Lecture Notes in Computer Science 3645

De-Shuang Huang Xiao-Ping Zhang
Guang-Bin Huang (Eds.)

Advances in Intelligent Computing

International Conference on Intelligent Computing, ICIC 2005
Hefei, China, August 23-26, 2005
Proceedings, Part II

Springer

Volume Editors

De-Shuang Huang
Chinese Academy of Sciences
Institute of Intelligent Machines
P.O. Box 1130, Hefei, Anhui, 230031, China
E-mail: dshuang@iim.ac.cn

Xiao-Ping Zhang
Ryerson University
Department of Electrical and Computer Engineering
350 Victoria Street, Toronto, Ontario, Canada, M5B 2K3
E-mail: xzhang@ee.ryerson.ca

Guang-Bin Huang
Nanyang Technological University
School of Electrical and Electronic Engineering
Nanyang Avenue, Singapore 639798
E-mail: egbhuang@ntu.edu.sg

Library of Congress Control Number: 2005930888

CR Subject Classification (1998): F.1, F.2, I.2, G.2, I.4, I.5, J.3, J.4, J.1

ISSN 0302-9743
ISBN-10 3-540-28227-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28227-3 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11538356 06/3142 5 4 3 2 1 0

Preface

The International Conference on Intelligent Computing (ICIC) was set up as an annual forum dedicated to emerging and challenging topics in the various aspects of advances in computational intelligence fields, such as artificial intelligence, machine learning, bioinformatics, and computational biology, etc. The goal of this conference was to bring together researchers from academia and industry as well as practitioners to share ideas, problems and solutions related to the multifaceted aspects of intelligent computing.

This book constitutes the proceedings of the International Conference on Intelligent Computing (ICIC 2005), held in Hefei, Anhui, China, during August 23–26, 2005. ICIC 2005 received over 2000 submissions from authors in 39 countries and regions. Based on rigorous peer reviews, the Program Committee selected 563 high-quality papers for presentation at ICIC 2005; of these, 215 papers were published in this book organized into 9 categories, and the other 348 papers were published in five international journals.

The organizers of ICIC 2005 made great efforts to ensure the success of this conference. We here thank the members of the ICIC 2005 Advisory Committee for their guidance and advice, the members of the Program Committee and the referees for reviewing the papers, and the members of the Publication Committee for checking and compiling the papers. We would also like to thank the publisher, Springer, for their support in publishing the proceedings in the Lecture Notes in Computer Science series. Particularly, we would like to thank all the authors for contributing their papers. Without their high-quality papers, the success of the conference would not have been possible. Finally, we are especially grateful to the IEEE Computational Intelligence Society and the IEEE Hong Kong Computational Intelligence Chapter as well as their National Science Foundation of China for their sponsorship.

15 June 2005

De-Shuang Huang
Institute of Intelligent Machines, Chinese Academy of Sciences, China

Xiao-Ping Zhang
Ryerson University, Canada

Guang-Bin Huang
Nanyang Technological University, Singapore

ICIC 2005 Organization

ICIC 2005 was organized and sponsored by the Institute of Intelligent Machines in cooperation with the University of Sciences & Technology, Hefei, and technically cosponsored by the IEEE Computational Intelligence Society and the Hong Kong Computational Intelligence Chapter.

Committees

General Chairs	De-Shuang Huang, China Jiming Liu, Hong Kong, China Seong-Whan Lee, Korea

International Advisory Committee

Songde Ma, China
Marios M. Polycarpou, USA
Shoujue Wang, China
Yuanyan Tang, Hong Kong, China
Guanrong Chen, Hong Kong, China
Guoliang Chen, China
DeLiang Wang, USA
Horace H.S. Ip, HK
Paul Werbos, USA
Qingshi Zhu, China
Yunyu Shi, China
Yue Wang, China
Youshou Wu, China
Nanning Zheng, China
Fuchu He, China
Okyay Knynak, Turkey
Roger H. Lang, USA
Zheng Bao, China
Yanda Li, China
Ruwei Dai, China
Deyi Li, China
Yixin Zhong, China
Erke Mao, China

Program Committee Chairs	Yiu-Ming Cheung, Hong Kong, China Xiao-Ping Zhang, Canada
Organizing Committee Chairs	Tao Mei, China Yunjian Ge, China
Publication Chair	Hujun Yin, UK
Publicity Chair	Anders I. Morch, Norway
Registration Chair	Haiyan Liu, China
International Liaison Chair	Prashan Premaratne, Australia
Exhibition Chair	Guang-Zheng Zhang, China
Finance Chair	Hongwei Liu, China

Program Committee

Yuanyan Tang (Hong Kong, China), Jonathan H. Manton (Australia), Yong Xu (UK), Kang Li (UK), Virginie F. Ruiz (UK), Jing Zhang (China), Roberto Tagliaferri (Italy), Stanislaw Osowski (Poland), John Q. Gan (UK), Guang-Bin Huang (Singapore), Donald C. Wunsch (USA), Seiichi Ozawa (Japan), Zheru Chi (Hong Kong, China), Yanwei Chen (Japan), Masaharu Adachi (Japan), Huaguang Zhang (China), Choh Man Teng (USA), Simon X. Yang (Canada), Jufu Feng (China), Chenggang Zhang (China), Ling Guan (Canada), Ping Guo (China), Xianhua Dai (China), Jun Hu (China), Jinde Cao (China), Hye-Ran Byun (Korea), Key-Sun Choi (Korea), Dewen Hu (China), Sulin Pang (China), Shitong Wang (China), Yunping Zhu (China), Xuegong Zhang (China), Jie Tian (China), Daqi Zhu (China), Nikola Kasabov (New Zealand), Hai Huyen Dam (Australia), S. Purushothaman (India), Zhu Liu (USA), Kazunori Sugahara (Japan), Chuanlin Zhang (China), Hailin Liu (China), Hwan-Gye Cho (Korea), Hang Joon Kim (Korea), Xiaoguang Zhao (China), Zhi-Cheng Ji (China), Zhigang Zeng (China), Yongjun Ma (China), Ping Ao (USA), Yaoqi Zhou (USA), Ivica Kopriva (USA) Derong Liu (USA), Clement Leung (Australia), Abdelmalek Zidouri (Saudi Arabia), Jiwu Huang (China), Jun Zhang (China), Gary Geunbae Lee (Korea), Jae-Ho Lee (Korea), Byoung-Tak Zhang (Korea), Wen-Bo Zhao (China), Dong Hwa Kim (Korea), Yi Shen (China), C.H. Zhang (Japan), Zhongsheng Wang (China), YuMin Zhang (China), HanLin He (China), QiHong Chen (China), Y. Shi (Japan), Zhen Liu (Japan), K. Uchimura (Japan), L. Yun (Japan), ChangSheng Xu (Singapore), Yong Dong Wu (Singapore), Bin Zhu (China), LiYan Zhang (China), Dianhui Wang (Australia), Kezhi Mao (Singapore), Saeed Hashemi (Canada), Weiqi Chen (China), Bjonar Tessem (Norway), Xiyuan Chen (China), Christian Ritz (Australia), Bin Tang (Canada), Mehdi Shafiei (Canada), Jiangtao Xi (Australia), Andrea Soltoggio (UK), Maximino Salazar-Lechuga (UK), Benjamin S. Aribisala (UK), Xiaoli Li (UK), Jin Li (UK), Jinping Li (China), Sancho Salcedo-Sanz (Spain), Hisashi Handa (Japan)

Organizing Committee

Xianda Wu, Jinhuai Liu, Zengfu Wang, Huanqing Feng, Zonghai Chen, Gang Wu, Shuang Cong, Bing-Yu Sun, Hai-Tao Fang, Xing-Ming Zhao, Zhan-Li Sun, Ji-Xiang Du, Hong-Qiang Wang, Fei Han, Chun-Hou Zheng, Li Shang, Zhong-Hua Quan, Bing Wang, Peng Chen, Jun Zhang, Zhong-Qiu Zhao, Wei Jia

Secretary

Hai-Mei Zhang, China

Table of Contents – Part II

Genomics and Proteomics

Adaptation and Decision Making

Applications and Hardware

Other Applications

Table of Contents – Part I

Perceptual and Pattern Recognition

Informatics Theories and Applications

Computational Neuroscience and Bioscience

Models and Methods

Learning Systems

Protein Secondary Structure Prediction Using Sequence Profile and Conserved Domain Profile

Seon-Kyung Woo[1], Chang-Beom Park[2], and Seong-Whan Lee[1,2]

[1] Department of Bioinformatics, Korea University,
Anam-dong, Seongbuk-ku, Seoul 136-713, Korea
skwoo@image.korea.ac.kr

[2] Department of Computer Science and Engineering, Korea University,
Anam-dong, Seongbuk-ku, Seoul 136-713, Korea
{cbpark, swlee}@image.korea.ac.kr

Abstract. In this paper, we proposed a novel method for protein secondary structure prediction using sequence profile and conserved domain profile. Sequence profile generated from PSI-BLAST (position specific iterated BLAST) has been widely used in protein secondary structure prediction, because PSI-BLAST shows good performance in finding remote homology. Conserved domains kept functional and structural information of related proteins; therefore we could draw remote homology information in conserved domains using RPS-BLAST (reverse position specific BLAST). We combined sequence profile and conserved domain profile to get more remote homology information, and propose a method which used the combined profile to predict the protein secondary structures. In order to verify the effectiveness of our proposed method, we implemented a protein secondary structure prediction system. Overall prediction accuracy reached 75.9% on the RS126 data set. The improvement by incorporating conserved domain information exceeded 3%, and this result showed that our proposed method could improve significantly the accuracy of protein secondary structure prediction.

1 Introduction

Protein is a large and complex molecule which plays important roles in life processes, therefore it is essential to know the specific functions of proteins to understand biological system. To know the function of an unknown protein, we should resolve the protein structure, because protein structure leads to protein function[1]. We can deduce the biological function of a protein from its structure, because protein can perform its own functions only when it has a unique three-dimensional structure.

Since Kendrew determined the structure of myoglobin in 1958, many protein structures were resolved using X-ray crystallography and NMR(nuclear magnetic resonance). These experimental methods provide very high-resolution structural information, but it takes too much time and money to determine a structure of a protein using these methods. Besides, these methods can not resolve the structures of all types of proteins, because of their experimental characteristics and difficulties. These methods

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 1 – 10, 2005.

are useful only when we can get a crystal structure of each protein, however some types of proteins can not be crystallized and it means that it is impossible to get the exact structural information of those proteins.

After the successful completion of the human genome project, a large amount of protein sequence data was generated using high throughput sequencing devices. Protein sequence data are meaningless without structural information or functional information, but it seems impossible to determine the structure of all proteins using biological experiments, because experimental determination of protein structure is very slow compared to sequencing speed. For instance, more than 40 million sequences are deposited in GenBank(http://www.ncbi.nlm.nih.gov/), but only about 30 thousand known structures were deposited in PDB(http://www.rcsb.org/pdb). The gap between protein sequences and structures are widening, therefore many researchers have tried to develop various computational methods for protein structure prediction which can predict the three-dimensional structure from the amino acid sequence to understand the biological function of proteins. In spite of intensive researches, there is no method which can predict three-dimensional structure of proteins accurately and confidently, because a protein with N amino acids can form approximately 10^N structures. Therefore many researchers predict the secondary structure of proteins to obtain the structural information from sequence, because protein secondary structures are fundamental elements of protein structure and provide useful knowledge on three-dimensional structures of proteins. Secondary structure prediction methods are not often used alone, but are instead often used to provide constraints for tertiary structure prediction methods or as part of fold recognition methods[2].

There are three fundamental secondary structures: α-helix(H), β-sheet(E) and coil(C), so the aim of protein secondary structure prediction is to predict the three secondary structures of proteins from amino acid sequence.

2 Related Works

Many secondary structure prediction methods were proposed through extensive studies over the past 30 years. Early works used single residue statistics to predict secondary structure of proteins. There were not many known structures in 1970s, so the prediction accuracy was very low. The number of known structures of proteins which was increased slowly by considerable efforts of biologists made possible to use evolutionary information and more sophisticated machine learning algorithms than single residue statistics. Rost and Sander proposed a method which used sequence profiles and neural networks[3]. Their proposed method showed 69.7% of prediction accuracy. This result was significantly better than the results of previous methods. The main improvement of this method comes from the use of PSSM (position-specific score matrix). PSSM is multiple alignment result of homologous sequences and it is very useful for protein secondary structure prediction, because multiple sequence alignments contain more information about the structure than a single sequence. A small window of amino acids from PSSM was used as input of the network for prediction, and the network predicted the secondary structure of the amino acid at the middle position of the input window. Rost and Sander showed a remarkable result but their method had following problems. Overall prediction accuracy was relatively high, but it was still below 70%. Their method was weak in β-sheet prediction and predicted structures were too short compared to real protein secondary structure. David T. Jones used PSSM which was generated from PSI-BLAST(position specific iterated BLAST). Because PSI-BLAST showed better performance in finding remote homology than previous alignment methods[2], PSSM from PSI-BLAST was adopted by almost all state-of-the-art protein secondary structure prediction systems. In 2001, Hua and Sun proposed a new method which was based on SVM (support vector machine)[4]. They assembled binary classifiers to predict three protein secondary structures and got a very good result even though their method did not use the sequence profile generated from PSI-BLAST. SVM showed good performance in protein structure prediction, therefore many researches based on SVM followed. Guo and his colleagues proposed a method using dual-layer SVM and PSI-BLAST profile. The first layer SVM classifier was for sequence-to-structure prediction and the second layer SVM classifier was for structure-to-structure prediction. The second layer SVM classifier was used to improve the low prediction accuracy of β-sheet structure. Nguyen and Rajapakse proposed a multi-class SVM for protein secondary structure prediction to overcome the shortcomings of binary classifiers[5]. Their experiments showed that multi-class SVM methods were more suitable for protein secondary structure prediction than the other methods, because multi-class SVM could solve an optimization problem in one step and additional assemble procedures are not required.

Recent improvements in secondary structure prediction were resulted from the growth of sequence databases, better database search methods, better prediction methods[6]. Therefore the prediction accuracy of protein secondary structure will be increased with larger structural database, more sophisticated machine learning algorithms and more efficient search methods to find remote homologies of related proteins.

3 Protein Secondary Structure Prediction Using Sequence Profile and Conserved Domain Profile

We propose a novel method for protein secondary structure prediction which used the combined profile of sequence profile and conserved domain profile. Unlike previous methods which used only sequence profiles from PSI-BLAST, our proposed method used a combined profile of PSI-BLAST and RPS-BLAST (reverse position specific BLAST), and we could improve the accuracy of protein secondary structure prediction because the conserved domain profile generated from RPS-BLAST had valuable information on remote homology. We implemented SVM classifiers to test the effectiveness of our proposed method, because SVM had shown good performance in solving various problems in bioinformatics fields. The whole procedure of the proposed protein secondary structure prediction method is illustrated in Figure 1.

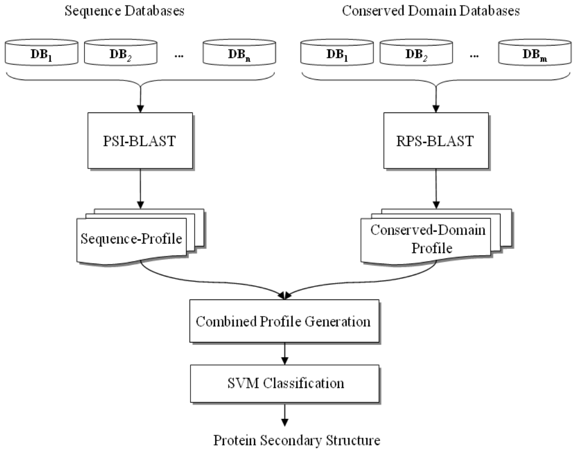


Fig. 1. Overview of the proposed protein secondary structure prediction method

3.1 Combined Profile of Sequence Profile and Conserved Domain Profile

Domains are functional and structural units of a protein, and can perform specific functions independently. Protein sequences are modified during evolution because of mutations, insertion and deletion in their residue. These modifications of sequences form protein families of related proteins. Although many modifications occurred in protein sequence, fundamental structure should be conserved to perform specific functions, and these conserved regions can be identified as conserved domain. Conserved domains are defined as recurring units in molecular evolution, whose extent can be determined by sequence and structure analysis[7]. Conserved domain contains information on common characteristics of related proteins and can be used to find remote homology. Conserved domain database is a collection of multiple sequence alignments for ancient domains and full-length proteins[8].

RPS-BLAST is a search tool used in conserved domain database. PSI-BLAST searches a database to find significantly similar sequence to the query sequence and build a PSSM for the query sequence[9]. PSI-BLAST iterates these procedures to refine the PSSM which contains information on remote homology. On the contrary, RPS-BLAST searches a database of pre-calculated PSSMs on conserved domains. PSSMs were already calculated and stored in conserved domain database; therefore RPS-BLAST can build the PSSM and show significant hits in a single pass. RPS-BLAST is useful to find the functional relationships of protein family. To compare the procedure of PSI-BLAST and RPS-BLAST, a diagram of PSI-BLAST and RPS-BLAST is depicted in figure 2.

PSI-BLAST is a powerful tool for extraction of remote-homologies among homologous proteins, because it refines the searching results iteratively. However, the iterative nature of the PSI-BLAST algorithm makes very sensitive profiles which are apt to be biased to outliers. Furthermore, PSI-BLAST generates sequence profile

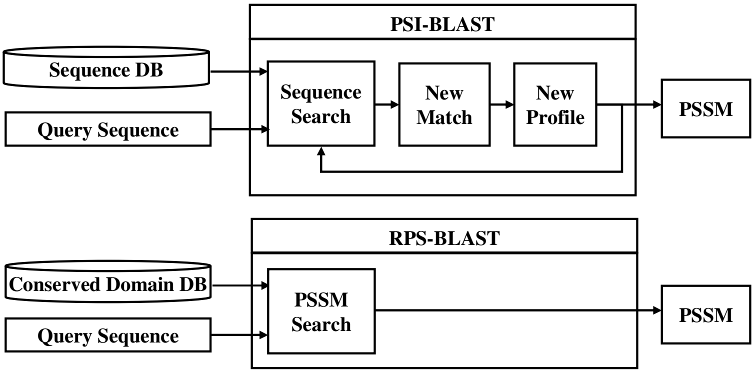


Fig. 2. Diagram of PSI-BLAST and RPS-BLAST

because PSI-BLAST uses sequence similarity to find remote homology. During evolution, sequences may change up to 85%, so it is hard to detect remote homologies only using sequence similarity. To improve this situation, we proposed a method which used contextual information of conserved domain, because conserved domains contained valuable information on functional and structural relationships among proteins.

To test the usefulness of conserved domain information, we performed prediction experiments on the RS126 data set using conserved domain profile generated from RPS-BLAST. Prediction accuracy was 1% or 2% lower compared to the accuracies of the previous methods which used sequence profile generated from PSI-BLAST. This result showed that conserved domain profile could be used for protein secondary structure prediction even though 20% of proteins in the RS126 data set were DUF(Domains of Unknown Function) or UPF(Uncharacterized Protein Families). The number of conserved domain in databases are growing rapidly, therefore the prediction accuracy will be increased as the number of available data will be increased.

The method which used only conserved domain profiles showed good prediction results, but we could not get better results because we could not get conserved domain profile of all proteins in query sequences. Therefore we proposed a method which used a combined profile of sequence profile and conserved domain profile. The PSSMs which were generated from PSI-BLAST and RPS-BLAST were basically homogeneous, because RPS-BLAST was a variant of the PSI-BLAST. We expected that we could improve the prediction accuracy, because our proposed method used sequence similarity and functional relationship together and because alignments which incorporate sequences with significant yet low sequence similarity to the target protein produce more accurate predictions that those which incorporate sequences which are very closely related to the target[2]. To test the effectiveness of the combined profile, we performed some experiments on the RS126 data set using PSSMs from PSI-BLAST profile, RPS-BLAST profile and combined profile. The results are

Table 1. Comparison with the results of different profiles

	Window Length(l)					
	$l = 9$	$l = 11$	$l = 13$	$l = 15$	$l = 17$	$l = 19$
PSI-BLAST Profile	71.5%	72.1%	72.3%	72.6%	72.7%	72.6%
RPS-BLAST Profile	70.5%	70.9%	71.1%	71.3%	71.4%	71.5%
Combined Profile	74.8%	75.2%	75.4%	75.4%	75.5%	75.9%

shown in Table 1. Among three profiles, combined profile showed the best results and it meant that combined profile had more information on remote homology of proteins and it could be used to improve the protein secondary structure prediction accuracy.

3.2 SVM for Protein Secondary Structure Prediction

SVM is a new type of binary pattern classifier based on a novel statistical learning technique that has been proposed by Vapnik. Previous prediction methods such as neural networks minimize the empirical training error, but SVM minimizes an upper bound of the generalization error by maximizing the margin between the separating hyperplane and the data. The Margin is the sum of the distances from the hyperplane to the closest data points of each class, and optimal separating hyperplane(OSH) has the maximum margin. When the two classes are not completely separable, SVM transforms the data using kernel from the input space into a high dimensional feature space by a nonlinear transformation where the two classes are linearly separable. SVM had shown good performance in solving various problems in bioinformatics fields, therefore we implemented a protein secondary structure prediction program based on SVM.

The basic SVM is a tool for two-class problem and there are three secondary structures of proteins. Therefore we made binary SVM classifiers and assembled binary classifiers to predict the secondary structures of proteins according to the method proposed by Hua and Sun[7]. We built three one-versus-all classifiers(H/~H, E/~E, C/~C) and three one-versus-one classifiers(H/E, E/C, C/H), and assembled these binary classifiers to make following five classifiers; SVM_TREE1, SVM_TREE2, SVM_TREE3, SVM_MAX_D and SVM_VOTE classifiers. SVM_TREE1, SVM_TREE2 and SVM_TREE3 classifiers were composed of one one-versus-all classifier and one one-versus-one classifier. One-versus-all classifiers were used first to predict the secondary structure of input residue. If the output result was positive, the secondary structure of the residue was predicted. If the output result was negative, then one-versus-one classifiers were used to predict the secondary structure. SVM_MAX_D classifier consisted of three one-versus-all classifiers and the structure of the input residue was determined by a classifier which had the largest positive distance to the optimal separating hyperplane. SVM_VOTE classifier combined all six binary classifiers and predicted the secondary structure of the input residue using simple voting method. Among five classifiers, SVM_MAX_D showed the best performance and was shown in figure 3.

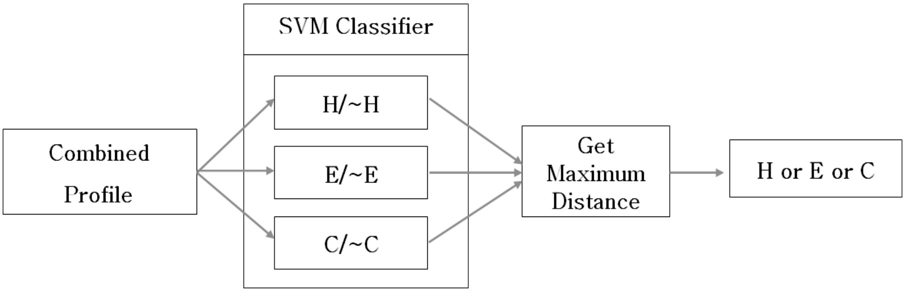


Fig. 3. Illustration of SVM_MAX_D Classifier

4 Experiments and Results

4.1 Dataset

The RS126 data set was used to measure the exact prediction accuracy of the proposed method and to compare with previous methods, because the RS126 data set had been widely used for protein secondary structure prediction. The RS126 data set was constructed by Rost and Sander[3] and contained 126 non-homologous proteins. Our proposed method was tested on these data sets using a seven-fold cross validation. We divided the data set into 7 subsets which had approximately similar size and structural information on three protein secondary structures to avoid biased subsets. Six subsets are used for training and one subset is used for testing.

4.2 Prediction Measurements

Three standard measurements were used to assess the prediction accuracy of the proposed method. The Q_3 accuracy indicates the percentage of correctly predicted residues of three secondary structures[10]. The Q_3 accuracy is the most obvious measurements for the overall accuracy of protein secondary structure prediction.

$$Q_3(\%) = \frac{\sum_{i \in \{H,E,C\}} \text{Number of residues correctly predicted in class i}}{\sum_{i \in \{H,E,C\}} \text{Number of residues observed in class i}} \times 100 \qquad (1)$$

4.3 Combined Profile Generation

We transformed RS126 data set into FASTA format to get evolutionary information using PSI-BLAST and RPS-BLAST. First, we tried to get PSSMs using web-based BLAST system which is provided by NCBI. However, web-based BLAST system was very slow, so it took so much time to get the profiles of 126 proteins. Therefore we had constructed a local BLAST sever in Linux environment, and got profiles from PSI-BALST and RPS-BLAST using this local BLAST server. We got sequence profile from PSI-BLAST after three iterations and conserved domain profile from RPS-BLAST for each protein in RS126 data set. About 20% of proteins in RS126 data set were DUF or UPF, but we could get conserved domain profile of those proteins using RPS-BLAST. After we got both sequence profile and conserved domain profile, we normalized and scaled the profiles to the [0, 1] range using the standard sigmoid function in equation 2

$$f(x) = \frac{1}{1 + \exp(-x)} \times 100 \tag{2}$$

where x is the raw profile matrix value.

Finally, we united sequence profile and conserved domain profile to construct a combined profile.

4.4 Protein Secondary Structure Prediction Using Combined Profile Based on Support Vector Machine

Each residue in profile had 20 columns and was coded to a 21-dimensional vector. 20 columns represent 20 amino acids and an additional element is used to indicate that the sliding window extends over the N- and the C- terminus. Therefore total length of feature vector is 21 x window length. We performed several experiments changing window length to know the influence of window length. Furthermore, we tested various kernel functions to find the most suitable kernel function for protein secondary structure prediction, and RBF(radial basis function) kernel was selected to train SVM. RBF kernel is calculated by:

$$K(x_i, x_j) = \exp(-r\left|x_i - x_j\right|^2) \tag{3}$$

where r is determined empirically for optimal performance. We used soft margin SVM, so the regularization parameter C is also needed to be determined. The regularization parameter C controls the trade-off between model complexity and misclassified training sequences by controlling the influence of a training sequence. We constructed SVM classifiers with the parameter $r = 0.001$ and $C = 1.0$.

SVM_TREE1, SVM_TREE2, SVM_TREE3, SVM_MAX_D and SVM_VOTE classifier were constructed to test the performance of our proposed method. Table 2 shows the Q_3 accuracies of the different classifiers. SVM_MAX_D showed the best performance among five classifiers. Our proposed method showed the best result when the window length was 19, and it meant that our proposed method required longer window length compared to previous methods, because previous methods showed the best results when the window length was 11 or 13. It seems that this characteristic is caused by the use of combined profile.

We compared our proposed method with previous methods using the RS126 data set. The results are shown in table 3. Results of PHD, DSC, Predator, NNSP, Multipred and Jpred were obtained from JPred webpage of the Barton group homepage (http://www.compbio.dundee.ac.uk/~www-jpred/accuracy.html). The results of Hua and Sun, Nguyen and Rajapakse were obtained from their papers[4][5]. Our proposed method showed better results compared to previous methods. This reflected that combined profile of sequence profile and conserved domain profile could improve the prediction accuracy.

Table 2. Q_3 accuracies of each classifier for RS126 data set

	Window Length(l)						
	$l = 9$	$l = 11$	$l = 13$	$l = 15$	$l = 17$	$l = 19$	$l = 21$
SVM_TREE1	73.1%	74.1%	74.1%	74.3%	73.9%	74.3%	74.5%
SVM_TREE2	73.0%	73.4%	73.9%	73.8%	73.7%	73.6%	73.6%
SVM_TREE3	74.0%	74.5%	74.2%	74.1%	74.5%	74.2%	74.4%
SVM_MAX_D	74.8%	75.2%	75.4%	75.4%	75.5%	75.9%	75.4%
SVM_VOTE	72.9%	73.7%	73.9%	73.8%	73.7%	74.1%	74.2%

Table 3. Comparison with other methods for RS126 data set

	Q_3(%)
PHD	73.5%
DSC	71.1%
Predator	70.3%
NNSP	72.7%
Multipred	67.2%
JPred	74.8%
Hua and Sun	71.1%
Nguyen and Rajapakse	72.8%
Proposed Method	75.9%

5 Conclusions and Further Research

Protein secondary structure prediction is one of the most promising research areas in bioinformatics because it provides key information to understand the structures of proteins. In this paper, we proposed a new method for protein secondary structure prediction which used sequence profile and conserved domain profile altogether. Conserved domain database is a collection of multiple sequence alignments for ancient domains and full-length proteins; therefore conserved domain profile which was generated from conserved domain database using RPS-BLAST has important information on remote homology and can be used to improve the accuracy of protein secondary structure prediction. Conserved domain profile can not be used alone for protein secondary structure prediction, because conserved domain database does not contain conserved domain information for all sequence queries. Therefore we proposed a new profile which combined sequence profile and conserved domain profile. To show the effectiveness of the proposed method, we performed some experiments using RS126 data set. On the RS126 data set, Q3 reached 75.9% which was nearly 3% higher than the methods which use sequence profile only. This result showed that conserved domain profile had useful information on remote homology and would be able to play an important role in protein secondary structure prediction.

The next step of our proposed method is to investigate the data similarity of profile data. The distributions of amino acid are different among protein families; therefore it will be useful information to increase the accuracy of protein secondary structure prediction. Moreover, current state-of-the-art protein secondary structure prediction methods use whole available structure data, therefore it is required to collect all structural information from PDB to increase the accuracy of the proposed method. Finally, more sophisticated combination method for sequence profile and conserved domain profile will increase the overall prediction accuracy.

References

1. Bourne, P.E. and Weissig, H.: Structural Bioinformatics, Wiley-Liss (2003)
2. Jones, D.T.: Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. Journal of Molecular Biology 292 (1999) 195-202
3. Rost, B. and Sander, C.: Prediction of Protein Secondary Structure at Better Than 70% Accuracy. Journal of Molecular Biology 232 (1993) 584-599
4. Hua, S. and Sun, Z.: A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach. Journal of Molecular Biology 308 (2001) 397-407
5. Nguyen, M.N. and Rajapakse, J.C.: Multi-Class Support Vector Machines for Protein Secondary Structure Prediction. Genome Informatics 14 (2003) 218 - 227
6. 6 Przybylski, D. and Rost, B.: Alignments Grow, Secondary Structure Prediction Improves. Proteins: Structure, Functions and Genetics 46 (2002) 197-205
7. http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.html
8. Bauer, A.M. et al.: CDD : a Conserved Domain Database for protein Classification. Nucleic Acids Research 33 (2005) 192-196
9. Altschul S. et al.: Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nuclei Acids Research 25 (1997) 3389-3402
10. Cuff, J.A. and Barton, G.J.: Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction. Proteins: Structure, Functions and Genetics (34) (1999) 508-519

Correlating Genes and Functions to Human Disease by Systematic Differential Analysis of Expression Profiles

Weiqiang Wang, Yanhong Zhou, and Ran Bi

Center for Bioinformatics, Huazhong University of Science and Technology,
Wuhan, Hubei 430074, China
yhzhou@hust.edu.cn

Abstract. Genome-wide differential expression studies of human diseases using microarray technology usually produce long lists of genes with altered expression, therefore, the genes causally involved in a disease cannot be effectively separated from innocent bystanders. Existing methods for differential analysis of gene expression profiles seem unable to solve this problem successfully. In this paper, we present a systematic strategy that combines gene-wise and function-wise differential analysis of gene expression profiles to interrelate genes and functions with human diseases. The gene-wise analysis adopts a modified T-test to analyze the expression alteration of each single gene, and the alteration is represented by quantitative significant *p*-value. The function-wise analysis uses a new combined S-test to identify coordinate alterations of genes within each functional category. A computational tool, MageKey, is developed based on this strategy, and its utility is demonstrated by the analysis results of gene expression dataset of human Amyotrophic Lateral Sclerosis disease. MageKey is freely available upon request to authors.

1 Introduction

Microarray technology, which can be used for measuring the expression levels of thousands of genes simultaneously in a single experiment and generating gene expression profiles, has become a powerful tool in the post-genome era [1-3]. One major application area of microarray technology is to discover disease relevant genes, which may lead to gaining pathogenetic insights into diseases and identifying potential therapeutic targets [3-6]. However, the great potential of microarray technology is currently limited by the lack of powerful computational tools to make sense of the generated gene expression profiles [7-9].

For the analysis of gene expression profiles, there is a widely held assumption that genes showing significant expression difference in different conditions are likely to be biologically important [3,5,10]. We define gene-wise differential analysis as the detection of expression difference of single gene. For this purpose, a simple approach is to do traditional two-sample T-test, and many sophisticated statistical methods have been proposed in consideration of relatively few samples with overmany genes and extremely low signal-to-noise ratio in expression profiles [11-14]. These methods usually

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 11 – 20, 2005.

produce long lists of differentially expressed genes, therefore, they cannot effectively separate the genes causally involved in a disease from innocent bystanders [3,5,10].

It seems reasonable to suppose that the presence of genetic alterations may frequently lead to changing the expression levels of many other genes functionally connected with disease genes [5]. One promising method is to analyze the alterations of gene expression at the functional level, such as biological pathways. Functional analysis of expression profiles may help disease gene discovery by revealing pathways or other functional information related to target genes [5,6,15]. We define function-wise differential analysis as the detection of coordinate alteration in the expression of a gene-set containing multiple genes with specific functional characteristics. Although some tools are available to provide such analysis, e.g. OntoExpress [16,17], these tools require a predefined threshold for gene selection and check for over-representation using only qualitative information. A recently developed tool, GOAL [18], uses the T-test statistic to aggregate an average as a combined measurement, but it can detect only those mis-regulations with consistent directions.

In this paper, we present a systematic strategy that combines gene-wise and function-wise differential analysis of gene expression profiles to interrelate genes and functions with human diseases. The gene-wise analysis adopts a modified T-test to analyze the expression alteration of each single gene, and the alteration is represented by quantitative significant *p*-value. The function-wise analysis uses a new combined S-test to identify coordinate alteration of genes within each functional category. A computational tool, MageKey, is developed based on this strategy, and its utility is demonstrated by the analysis results of gene expression dataset of human Amyotrophic Lateral Sclerosis (ALS) disease.

2 Method

The flowchart of MageKey is shown in Fig. 1. First, gene-wise differential analysis is used to evaluate the differential expression of each single gene. Then, genes are organized into gene-sets corresponding to different functional categories. Finally, function-wise differential analysis is used to evaluate coordinate expression change of each gene-set.

2.1 Gene-Wise Differential Analysis

Gene-wise differential analysis determines the expression difference of individual genes between two conditions. The simplest method is to use a two-sample T-test, which measures the difference in means divided by the sum of standard deviations under two conditions. When sample size per condition is small, however, the variance in T-test is difficult to estimate because of erratic fluctuations and heterogeneity [13,14]. One feasible way to improve T-test's performance is to obtain more stable variance estimate, which has already been taken into account in SAM [12] and Cyber-T [11]. In SAM, a small positive constant is added to the gene-specific variance estimates. With this modification, genes with small variance will not be considered as significant, however, genes with large

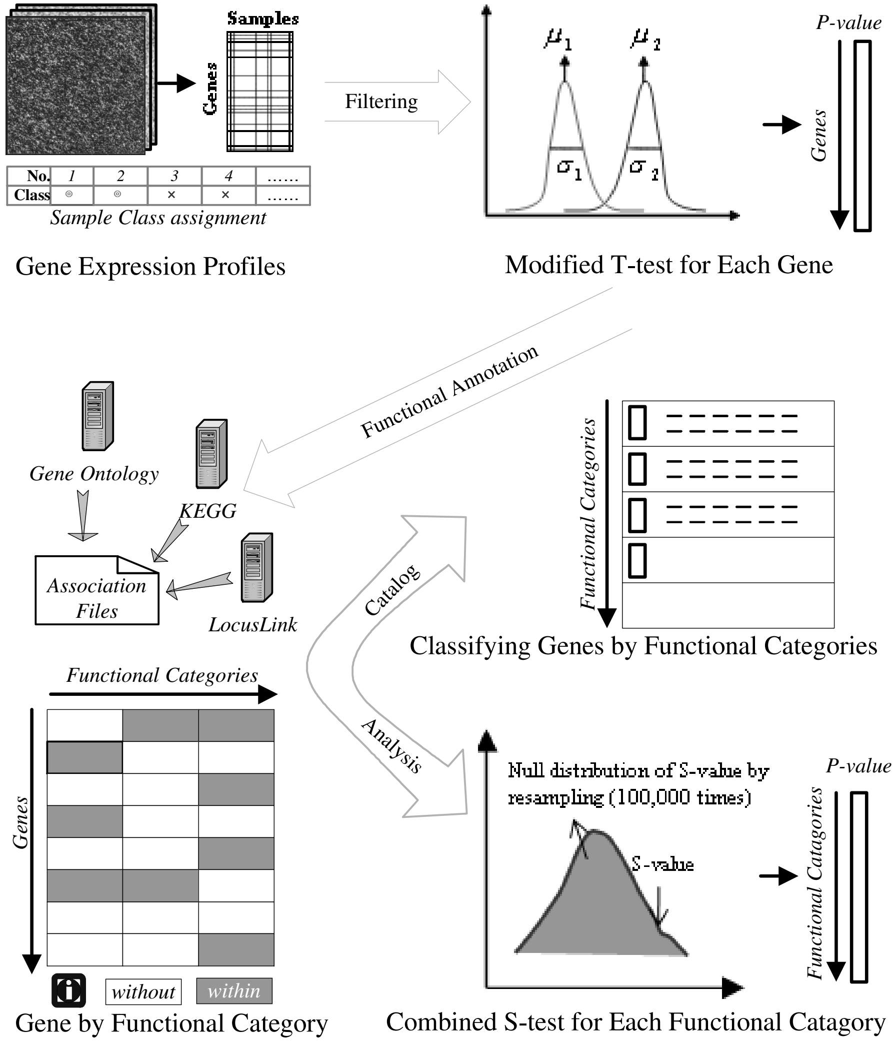


Fig. 1. Flowchart of MageKey. The goal of MageKey is to determine whether the expressions of multiple genes within a predefined functional category are consistently altered between two conditions. See text for details.

variance will be always considered as non-significant. In Cyber-T, it combines information from gene-specific and global average variance estimates and thus should be more stable. But it will generate bias when the assumption of normal distribution is violated.

We adopt a non-parametric T-test with modified variance estimates, which uses a tuning parameter (usually be 0.5) to determine the relative weights of gene-specific variance and global average variance, and reports its significance by permutation analysis.

Suppose that e_{ij} is the expression level of gene i in sample j $(i=1,2,\cdots,m; j=1,\cdots,n_1,n_1+1,\cdots,n_1+n_2)$, where the first n_1 and the last n_2 samples are obtained under two conditions, respectively. Then, for any gene i, the means of its expression under the two conditions can be determined by

$$\overline{e_{i(1)}}=\sum_{j=1}^{n_1}e_{ij}\Big/n_1 \ , \tag{1}$$

$$\overline{e_{i(2)}}=\sum_{j=n1+1}^{n_1+n_2}e_{ij}\Big/n_2 \ , \tag{2}$$

and the variances of its expressions can be determined by

$$s_{i(1)}^2=\sum_{j=1}^{n_1}(e_{ij}-\overline{e_{i(1)}})^2\Big/(n_1-1) \ , \tag{3}$$

$$s_{i(2)}^2=\sum_{j=n_1+1}^{n_1+n_2}(e_{ij}-\overline{e_{i(2)}})^2\Big/(n_2-1) \ . \tag{4}$$

Let the overall gene-specific variance of gene i and the overall global average variance be

$$s_i^2=\frac{(1/n_1+1/n_2)\cdot((n_1-1)\cdot s_{i(1)}^2+(n_2-1)\cdot s_{i(2)}^2)}{n_1+n_2-2} \ , \tag{5}$$

$$s^2=\sum_{i=1}^{m}s_i^2\Big/m \ . \tag{6}$$

We define the non-parametric T-test statistic with modified variance as

$$t_i=\frac{\overline{e_{i(1)}}-\overline{e_{i(2)}}}{\sqrt{\lambda\cdot s_i^2+(1-\lambda)\cdot s^2}} \qquad (\lambda\in[0,1]) \ . \tag{7}$$

In addition, permutation analysis by shuffling condition-labels is used to calculate the statistical significance values (p-values) for this modified test statistic under null hypothesis, and the p-value of gene i is denoted as p_i $(i=1,2,\cdots,m)$. A small p-value for a given gene indicates its significant expression difference.

2.2 Categorization of Gene-Sets

Genes belonging to same functional category are grouped into a gene-set according to their annotation information. For this purpose, KEGG [19] biochemical pathways and GO [20] functional terms are currently used in MageKey. The KEGG biochemical pathway annotation information is acquired from KEGG PATHWAY database (August 2004, http://www.genome.jp/kegg/pathway.html). The GO hierarchy and GO functional annotation information is acquired from Gene Ontology database (September 2004, http://www.geneontology.org). Other annotation information including

Locus ID, gene symbol, gene title and OMIM, which are acquired from LocusLink [21] database (September 2004, http://www.ncbi.nlm.nih.gov/LocusLink), have also been integrated in MageKey.

2.3 Function-Wise Differential Analysis

Function-wise differential analysis determines if the expressions of genes within a given gene-set are consistently altered between two conditions.

Suppose that the *k* genes numbered $I_j\,(j=1,2,\cdots,k;I_j\in[1,m]))$ belong to the same functional category *I*. Then the *p*-values of these *k* genes will be p_{I_j} $(j=1,2,\cdots,k;I_j\in[1,m])$. We define the average differential expression degree and the average non-differential expression degree of these *k* genes as

$$S_{diff}=\left(\sum_{j=1}^{k}-\log(p_{I_j})\right)\Big/k\ , \tag{8}$$

$$S_{non-diff}=\left(\sum_{j=1}^{k}-\log(1-p_{I_j})\right)\Big/k\ . \tag{9}$$

Then, we use log-ratio of these two measures as the combined S-test statistic to evaluate consistent expression alteration of the gene-set *I*.

$$S=\log\left(\frac{S_{diff}}{S_{non-diff}}\right)=\log\left(S_{diff}\right)-\log\left(S_{non-diff}\right)\ . \tag{10}$$

To calculate the statistical significance values (*p*-value) for this test statistic, permutation analysis by random resampling is used.

3 Result

To demonstrate the utility and validity of MageKey, here, we use it to analyze the publicly available human ALS disease expression profile dataset (GSE833 in GEO [22], http://www.ncbi.nlm.nih.gov/GEO).

This dataset measures expression levels of ~6,800 genes in postmortem spinal cord gray matter obtained from 7 individuals with ALS as well as 4 normal individuals, using high-density oligonucleotide microarray technology containing 7070 probe sets. This dataset has been preprocessed before we obtained it. To eliminate genes that had extremely low expression, we filter the 7070 probe sets as follows. First, only those probe sets for which there are at least a single measure greater than 100 are kept. Then, probe sets are mapped to Locus ID using NetAffx [23] (http://www.affymetrix.com/analysis/index.affx), and those probe sets without assigned "Locus" ID are discarded. Of the 7070 probe sets, 5858 probe sets meet this filtering criterion, representing 5164 genes. In gene-wise analysis, multiple probe sets for a particular gene are treated independently, that is, individual *p*-value is calculated for each probe set. In function-wise analysis, they are treated as a whole by transparently calculating the geometric mean *p*-value for the gene.

3.1 Gene-Wise Differential Analysis

We applied modified T-test with 1000 permutations to analyze differential expression of 5858 probe sets (5164 genes) between ALS and normal individuals, and then we obtained the approximate *p*-value for each gene. When comparing the distribution of actual *p*-values with that of random *p*-values generated by 10 separate random permutations of the disease-state labels in the actual dataset, there are overmany small *p*-values occurred in the actual dataset (data not shown), indicating the presence of some genes related with human ALS disease.

Table 1. A list of genes related with human ALS disease according to KEGG ALS pathway (05030). Also included are the t-value, the permutation test p-value, the rank of the probe set (by the absolute of t-value) and some other annotations.

probe set ID	t-value	*p*-value	rank*	Locus ID	gene symbol	OMIM
X15306_rna1_at	-3.832	0.042	49	4744	NEFH	162230
X02317_at	-3.745	0.021	54	6647	SOD1	147450
Y00067_rna1_at	-3.603	0.051	62	4741	NEF3	162250
Y00433_at	3.587	0.036	64	2876	GPX1	138320
X05608_at	-3.553	0.045	68	4747	NEFL	162280
U66879_at	-3.015	0.021	122	572	BAD	603167
L14778_s_at	-2.338	0.022	276	5530	PPP3CA	114105
U57341_at	-1.964	0.094	419	4747	NEFL	162280
Z23115_at	-1.712	0.031	552	598	BCL2L1	600039
Z69043_s_at	1.355	0.203	829	6748	SSR4	300090
M22898_at	1.112	0.006	1110	7157	TP53	191170
D31890_at	1.061	0.186	1191	3735	KARS	601421
M14745_at	-0.979	0.030	1318	596	BCL2	151430
X04085_rna1_at	-0.711	0.189	1921	847	CAT	115500
U57341_r_at	-0.390	0.833	3039	4747	NEFL	162280
U01824_at	-0.342	0.722	3279	6506	SLC1A2	600300
M13994_s_at	0.067	0.946	5295	596	BCL2	151430

Table 1 lists a set of 17 probe sets (14 genes) related with human ALS disease according to KEGG ALS pathway (05030). It can be seen that only 5 out of these 14 genes have been significantly mis-regulated (here, top 100 genes are considered to be significantly mis-regulated), but many other genes which are currently considered as unrelated with human ALS disease have shown significant expression alterations. This phenomenon, which is probably typical in most human diseases, implies that it is unrealistic to effectively identify disease relevant genes only by gene-wise differential analysis of gene expression profiles.

3.2 Function-Wise Differential Analysis by Using KEGG Biochemical Pathways

We then applied combined S-test with 100000 permutations to analyze 95 KEGG biochemical pathways (each pathway consists of 8~200 genes). Table 2 lists the top 5 most significantly mis-regulated pathways. If we consider the pathways with *p*-value less than 0.01 as significant, there will be only 2 significant pathways, one of which is the human ALS disease pathway.

After the identification of the most significantly mis-regulated ALS pathway, we reexamined the individual expression difference of genes within ALS pathway (see Table 1). Although none of these genes is extremely significantly mis-regulated, most of them show more or less expression alterations. Moreover, two known ALS susceptible genes, i.e. NEFH and SOD1, can be identified from the ALS pathway by virtue of their sufficient expression alterations relative to the others. A strategy combing gene-wise analysis and function-wise analysis will be more effective.

Table 2. Top 5 most significantly mis-regulated KEGG biochemical pathways identified from gene expression profiles of 7 ALS patients and 4 normal individuals. The number of profiled genes for each pathway is listed. The S-value for each pathway is calculated using the combined S-test statistics and the corresponding *p*-value is calculated using random resampling.

Function_ID	Function_Name	Gene_Count	S-value	*p*-value
KEGG Biochemical Pathway				
05030	Amyotrophic lateral sclerosis (ALS)	14	2.733	0.00046
00903	Limonene and pinene degradation	15	1.999	0.00835
00380	Tryptophan metabolism	54	1.265	0.01583
00350	Tyrosine metabolism	37	1.257	0.03153
00561	Glycerolipid metabolism	68	1.097	0.03319

3.3 Function-Wise Differential Analysis Using GO Functional Terms

We also applied combined S-test with 100000 permutations to analyze GO functional categories (containing 568 GO biological process terms, 127 GO cellular component terms, 418 GO molecular function terms, each term is covered by 8~200 genes). Table 3 lists the top 5 most significantly mis-regulated functional terms in all three aspects. If we consider functional terms with *p*-value less than 0.01 as significant, there will be 4 significant GO biological process terms, 3 significant GO cellular component terms, none significant GO molecular function terms. This suggests that consistently mis-regulated phenomenon often occurs in specific cellular components and specific biological processes. In addition, some of the identified GO biological processes, such as protein kinase C activation (GO:0007205) and antigen processing (GO:0030333), have been postulated to be associated with human ALS disease [24].

Table 3. Top 5 most significantly mis-regulated GO functional terms (for biological process, cellular component, and molecular function, respectively) identified from gene expression profiles of 7 ALS patients and 4 normal individuals. The number of profiled genes for each term is listed. The S-value for each term is calculated using the combined S-test statistics and the corresponding p-value is calculated using random resampling.

Function_ID	Function_Name	Gene_Count	S-value	*p*-value
GO Biological Process				
GO:0007194	negative regulation of adenylate cyclase activity	8	2.610	0.00614
GO:0030333	antigen processing	19	1.825	0.00663
GO:0045761	regulation of adenylate cyclase activity	23	1.661	0.00781
GO:0007205	protein kinase C activation	9	2.484	0.00833
GO:0006221	pyrimidine nucleotide biosynthesis	10	2.222	0.01154
GO Cellular Component				
GO:0005875	microtubule associated complex	32	1.837	0.00105
GO:0030529	ribonucleoprotein complex	124	1.179	0.00202
GO:0005840	ribosome	74	1.263	0.00519
GO:0005842	cytosolic large ribosomal subunit (sensu Eukarya)	25	1.519	0.01296
GO:0005830	cytosolic ribosome (sensu Eukarya)	44	1.275	0.02114
GO Molecular Function				
GO:0003887	DNA-directed DNA polymerase activity	10	2.338	0.01077
GO:0045012	MHC class II receptor activity	12	2.096	0.01185
GO:0008080	N-acetyltransferase activity	8	2.303	0.01275
GO:0016410	N-acyltransferase activity	9	2.328	0.01317
GO:0005003	ephrin receptor activity	12	1.900	0.01570

4 Conclusion

We proposed a systematic strategy that combines gene-wise and function-wise differential analysis of gene expression profiles to correlate genes and functions with human diseases, and a computational tool, MageKey, has been developed based on this strategy. When applied to analyze the gene expression dataset of human ALS disease, the results demonstrate that MakeKey could effectively associate biological functions with human ALS disease, both in KEGG biochemical pathways and GO functional category, and some of these biological functions have been postulated or known to be associated with human ALS disease. Furthermore, with the help of these identified biological functions, the genes responsible for the disease can be determined at a relative lower cost. In conclusion, the proposed method and the MageKey program is able to effectively interrelate genes and functions with human diseases.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grants No. 30370354 and 90203011).

References

1. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O.: Quantitative Monitoring of Gene Expression Patterns With a Complementary DNA Microarray. Science 270(5235) (1995) 467-470
2. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L.: Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays. Nat. Biotechnol. 14(13) (1996) 1675-1680
3. Schulze, A. and Downward, J.: Navigating Gene Expression Using Microarrays--a Technology Review. Nat. Cell Biol. 3(8) (2001) E190-E195
4. Young, R.A.: Biomedical Discovery With DNA Arrays. Cell 102(1) 2000 9-15
5. Meltzer, P.S.: Spotting the Target: Microarrays for Disease Gene Discovery. Curr. Opin. Genet. Dev. 11(3) (2001) 258-263
6. Lyons, P.A.: Gene-Expression Profiling and the Genetic Dissection of Complex Disease. Curr. Opin. Immunol. 14(5) (2002) 627-630
7. Brazma, A. and Vilo, J.: Gene Expression Data Analysis. FEBS Lett. 480(1) (2000) 17-24
8. Nadon, R. and Shoemaker, J.: Statistical Issues With Microarrays: Processing and Analysis. Trends Genet. 18(5) (2002) 265-271
9. Leung, Y.F. and Cavalieri, D.: Fundamentals of CDNA Microarray Data Analysis. Trends Genet. 19(11) (2003) 649-659
10. Miklos, G.L. and Maleszka, R.: Microarray Reality Checks in the Context of a Complex Disease. Nat. Biotechnol. 22(5) (2004) 615-621
11. Baldi, P. and Long, A.D.: A Bayesian Framework for the Analysis of Microarray Expression Eata: Regularized t-Test and Statistical Inferences of Gene Changes. Bioinformatics 17(6) (2001) 509-519
12. Tusher, V.G., Tibshirani, R., and Chu, G.: Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. Proc. Natl. Acad. Sci. U.S.A. 98(9) (2001) 5116-5121
13. Wu, T.D.: Analysing Gene Expression Data From DNA Microarrays to Identify Candidate Genes. J. Pathol. 195(1) (2001) 53-65
14. Cui, X. and Churchill, G.A.: Statistical Tests for Differential Expression in CDNA Microarray Experiments. Genome Biol. 4(4) (2003) 210
15. McCarthy, M.I., Smedley, D., and Hide, W.: New Methods for Finding Disease-Susceptibility Genes: Impact and Potential. Genome Biol. 4(10) (2003) 119
16. Khatri, P., Draghici, S., Ostermeier, G.C., and Krawetz, S.A.: Profiling Gene Expression Using Onto-Express. Genomics 79(2) (2002) 266-270
17. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., and Krawetz, S.A.: Global Functional Profiling of Gene Expression. Genomics 81(2) (2003) 98-104
18. Volinia, S., Evangelisti, R., Francioso, F., Arcelli, D., Carella, M., and Gasparini, P.: GOAL: Automated Gene Ontology Analysis of Expression Profiles. Nucleic Acids Res. 32(Web Server issue) (2004) W492-W499
19. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M.: The KEGG Resource for Deciphering the Genome. Nucleic Acids Res. 32(Database issue) (2004) D277-D280

20. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la, C.N., Tonellato, P., Jaiswal, P., Seigfried, T., and White, R.: The Gene Ontology (GO) Database and Informatics Resource. Nucleic Acids Res. 32(Database issue) (2004) D258-D261
21. Pruitt, K.D. and Maglott, D.R.: RefSeq and LocusLink: NCBI Gene-Centered Resources. Nucleic Acids Res. 29(1) (2001) 137-140
22. Edgar, R., Domrachev, M., and Lash, A.E.: Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. Nucleic Acids Res. 30(1) (2002) 207-210
23. Liu, G., Loraine, A.E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M.A.: NetAffx: Affymetrix Probesets and Annotations. Nucleic Acids Res. 31(1) (2003) 82-86
24. Malaspina, A. and de Belleroche, J.: Spinal Cord Molecular Profiling Provides a Better Understanding of Amyotrophic Lateral Sclerosis Pathogenesis. Brain Res. Rev. 45(3) (2004) 213-229

On the Evolvement of Extended Continuous Event Graphs

Duan Zhang[1], Huaping Dai[1], Youxian Sun[1], and Qingqing Kong[2]

[1] National Laboratory of Industrial Control Technology Zhejiang University,
Hangzhou, P.R. China 310027
{dzhang, hpdai, yxsun}@iipc.zju.edu.cn
[2] Department of Mathematics, University of Manchester,
Manchester M13 9PL, United Kingdom
kdragon80@163.com

Abstract. Extended Continuous Event Graphs (ECEG) are a special class of Continuous Petri Nets. As the limiting form of Extended Timed Event Graphs (ETEG), ECEGs can be used not only to describe the discrete events approximately, but also to describe the continuous processes. In this note, we obtain some of the global properties of ECEGs. In the end, a simple example is given to illustrate the feedback control of CEGs with input.

1 Introduction

In the research of Discrete Event Dynamic Systems (DEDS), the modeling and analysis are usually in three basic hierarchies: algebraic hierarchy, logic hierarchy and statistic performance hierarchy. In algebra hierarchy, the most important method is dioid algebra and its two special forms max-plus algebra and min-plus algebra. One of the key tools in logic hierarchy is Petri Net. Event Graphs are special Petri Nets, that is, upstream transition or downstream transition of every position is no more than one. Cohen is the first one who derived the relation between the max-plus algebra and Petri Nets is Cohen. He constructed the max-plus algebra linear model of TEGs [1]. The further research on analyses and syntheses of TEGs based on this model can be found i[1]n [2-4]. In another way, the analyses of TEGs can be treated as solving a set of dioid equations [5,6]. Extended Timed Event Graph (ETEG, or Timed Event MultiGraph) is the extended form of TEG. The dioid algebraic model of ETEG is construct in different ways by Cohen [7] and Dai [8] independently. The analyses and syntheses of ETEG based on dioid algebraic model had been studied in [9]. In [7], Cohen also considered the dioid algebraic representation for the event graph with continuous positions and discrete transitions. In [10], the min-plus algebraic model of a class of Continuous Event Graphs (CEGs) was gived. All these algebraic models can convert serious logic non-linearity into algebraic linear form.

[1] This work was supported by the National Natural Science Foundation of China (No. 20206027), the Natural Science Foundation of Zhejiang Province (No. 202046) and the National Basic Research Program (973) of China (No. 2002CB312200).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 21 – 30, 2005.

15 years ago, David and Alla proposed continuous Petri Nets, in which transition fire continuously [11,12]. However the problem of analyses and syntheses of continuous Petri Nets is far from solving. In this paper, we focus on ECEGs, which are the event graph forms of continuous Petri nets, and are also the limiting forms of ETEGs [11,12]. ECEG can be used not only to describe discrete systems approximately, but also to describe continuous processes. So the studies of ECEGs and the corresponding algebraic models are of great significance.

The essential of the algebraic method to study TEG and ETEG is that the algebraic linear equations will be derived by the local relation (within transition, its upstream positions and the upstream transitions of those positions) in TEG or ETEG, then the global properties can be elicited by using max-plus algebra and dioid algebra. But it becomes rather complex for ECEGs. In this note, we obtain some of the global properties of ECEGs. We also discuss briefly about the feedback control of CEGs with input.

2 Prelimitationary

Denote the positive integer set as $\mathbb{N}$, the real set as $\mathbb{R}$. Let $\omega = +\infty$, $\mathbb{R}_{\min} := \mathbb{R} \bigcup \{\omega\}$, $\mathbb{R}^{+} := [0, +\infty)$ and $\mathbb{R}^{+}_{\min} := \mathbb{R}^{+} \bigcup \{\omega\}$.

For any $a, b \in \mathbb{R}_{\min}$, we define $a \oplus b = \min(a, b)$ and $a \otimes b \;\; = a + b$,where $a \oplus \omega = \omega \oplus a = a$ and $a \otimes \omega = \;\; \omega \otimes a = \omega$. $\langle \mathbb{R}_{\min}, \oplus, \otimes \rangle$ denotes the min-plus algebra on $\mathbb{R}_{\min}$. The symbol "$\otimes$" will be omitted sometimes for convenience.

$\mathbb{R}^{m \times n}_{\min}$ is the set of $m \times n$ matrixes, and the entries of the matrixes belong to $\mathbb{R}_{\min}$. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}_{\min}$, then

$$(\boldsymbol{A} \oplus \boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij} \oplus (\boldsymbol{B})_{ij} \tag{1}$$

For any $\boldsymbol{A} \in \mathbb{R}^{m \times k}_{\min}$ and $\boldsymbol{B} \in \mathbb{R}^{k \times n}_{\min}$, then $\boldsymbol{AB} \in \mathbb{R}^{m \times n}_{\min}$ and

$$(\boldsymbol{AB})_{ij} = (\boldsymbol{A} \otimes \boldsymbol{B})_{ij} = \bigoplus_{h=1}^{k} [(\boldsymbol{A})_{ih} \otimes (\boldsymbol{B})_{hj}]. \tag{2}$$

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}_{\min}$. $\boldsymbol{A}^0 \in \mathbb{R}^{n \times n}_{\min}$, where $(\boldsymbol{A}^0)_{ij} = 0$ for $i = j$ and $(\boldsymbol{A}^0)_{ij} = \omega$ for $i \neq j$, $\boldsymbol{A}^1 = \boldsymbol{A}$, $\boldsymbol{A}^2 = \boldsymbol{AA}$, $\cdots$, and $\boldsymbol{A}^* = \bigoplus_{i=0}^{\infty} \boldsymbol{A}^i$. If for any $i \in \mathbb{N}$, there is no negative value in diagonal entries of $\boldsymbol{A}^i$, we say $\boldsymbol{A}$ has no negative weight circuit; if each diagonal entry is greater than zero, then say $\boldsymbol{A}$ only has positive weight circuit.

For any $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^{n \times 1}_{\min}$, we have:

Theorem 1 [2-4]*:* If $\boldsymbol{A} \in \mathbb{R}^{n \times n}_{\min}$ has no negative weight circuit, there exists a maximum solution to equation $\boldsymbol{x} = \boldsymbol{Ax} \oplus \boldsymbol{b}$ which is given by $\boldsymbol{x} = \boldsymbol{A}^* \boldsymbol{b}$; Moreover, if $\boldsymbol{A}$ only has positive weight circuit, this solution is unique.

Theorem 2 [2-4]*:* If $A \in \mathbb{R}_{\min}^{n\times n}$ has no negative weight circuit, then

$$A^* = A^0 \oplus A^1 \oplus A^2 \oplus \cdots \oplus A^{n-1} . \tag{3}$$

3 Definition of ECEG

Definition 1: A ECEG is a 6-tuplet $< P,T,R ,\bar{V},M_0,W >$. P is the set of positions; T , satisfying $P \bigcap T = \phi$, is the set of transitions; $R = R_{PT} \bigcup R_{TP}$, where $R_{PT} \subseteq P\times T$ and $R_{TP} \subseteq T\times P$; $\bar{V} : T \to (0,+\infty)$ is the maximum firing velocity of transitions; M_0 : $P \to \mathbb{R}^+$ is initial mark of places; $W : R \to \mathbb{R}^+$ is the weight of R . Since ECEGs are event graphs, for any $p \in P$, the cardinal number of $\{t \in T \mid < p,t >\in R_{PT}\}$ and $\{t \in T \mid < t,p >\in R_{TP}\}$ are both no more than one. If $W(R) = \{1\}$, a ECEG degenerates to a CEG.

ECEGs can be expressed by directed graphs with double circles representing positions and rectangle representing transitions (Figure 1).

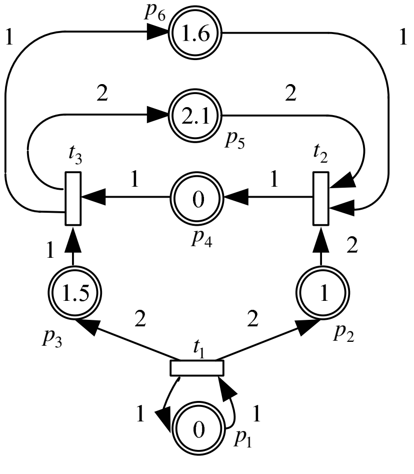


Fig. 1. A ECEG

Definition 2: For a transition t , ${}^\circ t = \{ p\in P \mid< p,t >\in R_{PT} \}$ and $t^\circ = \{ p\in P \mid< t, p >\in R_{TP} \}$ are called upstream and downstream of t respectively. For a position p , ${}^\circ p = \{t\in T \mid< t, p >\in R_{TP} \}$ and $p^\circ = \{t\in T \mid< p,t >\in R_{PT} \}$ are similarly called upstream and downstream of p respectively. Let *Nodes* be a set of transitions or positions, that is, ${}^\circ Nodes = \bigcup\limits_{n\in Nodes} {}^\circ n$ and $Nodes\,{}^\circ = \bigcup\limits_{n\in Nodes} n^\circ$.

Throughout this note, $mark_p(\tau)$ denotes mark of position p at time τ, and $V_t(\tau)$ denotes the firing velocity of transition t at time τ. Obviously, $mark_p(0) = M_0(p)$.

Definition 3: Let $T_{in} := \{t \in T \mid {}^\circ t = \phi\}$. The firing velocity of $t \in T_{in}$ can be arbitrary within $[0, \bar{V}(t)]$, so we may treat this transition as a input. For all $t \notin T_{in}$, the value of $V_t(\tau)$ can be obtained by the following algorithm:

step1. Let $T_S = \{t \in T \mid (\forall p \in {}^\circ t \;\; mark_p(\tau) > 0)$ or $(t \in T_{in})\}$. If $T_S = \phi$, we have $V_t(\tau) = 0$ for all $t \in T$, and stop; otherwise, we have $V_t(\tau) = \bar{V}(t)$ for any $t \in T_S - T_{in}$, and for any $t \in T_{in}$, $V_t(\tau)$ is known. Set $T_d = T_S$.

step2. Let T_u be the difference set $T - T_d$. If $T_u = \phi$, stop. Set $T_n = \{t \in T_u \mid \forall p \in {}^\circ t \;\; ((mark_p(\tau) > 0) \text{ or } ({}^\circ p \subseteq T_d))\}$. If $T_n = \phi$, we have $V_t(\tau) = 0$ for all $t \in T_u$ and stop; otherwise we have

$$V_t(\tau) = \min\left\{ \min_{\substack{t_i \in {}^\circ({}^\circ t) \bigcap T_d, \, {}^\circ p_i = \{t_i\} \\ mark_{p_i}(\tau) = 0}} \{V_{t_i}(\tau) \frac{W(t_i, p_i)}{W(p_i, t)}\}, \; \bar{V}(t) \right\} \tag{4}$$

for any $t \in T_n$, then adjust T_d as $T_d = T_d \bigcup T_n$ and turn to step2.

Definition 4: The mark consumed by position p within time interval $[\tau_1, \tau_2]$ is denoted by $\Delta mark_p(\tau_1, \tau_2)$. Supposing $\{t_1\} = {}^\circ p$ and $\{t_2\} = p^\circ$, we have:

$$\Delta mark_p(\tau_1, \tau_2) = W(t_1, p) \int_{\tau_1}^{\tau_2} V_{t_1}(\tau) du - W(p, t_2) \int_{\tau_1}^{\tau_2} V_{t_2}(\tau) du . \tag{5}$$

The definition indicates that $mark_p(\tau)$ is continuous.

4 Main Results

We only consider ECEGs without input, namely $T_{in} \neq \phi$, in this section, so "without input" is usually omitted for briefness. If a position satisfies $p^\circ = \{t\}$ and ${}^\circ p = \phi$, we can add a position p_0 with zero initial mark and a transition t_0 satisfying $p_0^\circ = {}^\circ p_0 = \{t_0\}$, ${}^\circ t_0 = \{p_0\}$, $t_0^\circ = \{p, p_0\}$ and $W(t_0, p_0) = W(t_0, p_0) =$ $W(t_0, p) = 1$ (Fig. 2.). Without modified the firing velocities of those original transitions, this treatment has no effect on the evolution of the ECEG. Now, for any $t \in T$ equation $|{}^\circ t| = |{}^\circ({}^\circ t)|$ holds ($|S|$ denotes the cardinal number of set S).

Definition 5: For any $t \in T$, $M_t : [0, +\infty) \to \mathbb{R}_{\min}$ is non-decrease continuous function and represents the total mark consumed by t in time interval $[0, \tau]$, i.e., $M_t(\tau) = \int_0^\tau V_t(\tau) du$.

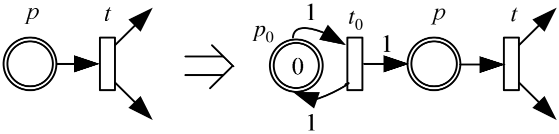


Fig. 2. Treatment for places without upstream

Lemma 1: For any $t \in T$ in a ECEG, the range of $V_t(\tau)$ is a finite set.

Proof: Using definition 3, it is clear that

$$\{V_t(\tau) \mid t \in T\} \subseteq \{0\} \cup \left\{ \overline{V}(t') \cdot \prod_{i=1}^{k} \frac{W^i}{W_i} \mid t' \in T, W^i \in W(R_{TP}), W_i \in W(R_{PT}), k = 1, \cdots, |P| \right\} \tag{6}$$

so the range is a finite set.

Definition 6: $\overline{P}(\tau) := \{p \in P \mid mark_p(\tau) > 0\}$.

According to the algorithm in definition 3, the following property of ECEGs without input can be obtained:

Lemma 2: If $\overline{P}(\tau_1) = \overline{P}(\tau_2)$, then $V_t(\tau_1) = V_t(\tau_2)$ for any $t \in T$; else if $\overline{P}(\tau_1) \subseteq \overline{P}(\tau_2)$, then $V_t(\tau_1) \le V_t(\tau_2)$.

Consider two transitions $t_1, t_2 \in T$, a elementary path from t_1 to t_2 is a path from t_1 to t_2 (including t_1 and t_2) without any circuit except itself, if it is. Here we exclude neither the case that $t_1 = t_2$ nor the elementary path that not includes any other transition. Elementary circuits in ECEGs are the circuits without any sub-circuit except itself.

Lemma 3: If $t \in T$ satisfies $V_t(\tau) = 0$, there exists $t' \in T$ satisfing $V_{t'}(\tau) = 0$, and there exist a elementary path Pa from t' to t and a elementary circuit Ci via t' . Moreover, every position $\overline{p}$ in Pa or Ci satisfies $mark_{\overline{p}}(\tau) = 0$ and every transition in Pa or Ci satisfies $V_{\overline{t}}(\tau) = 0$.

Proof: The equation $V_t(\tau) = 0$ indicates that there exists $p_1 \in {}^\circ t$ such that $mark_{p_1}(\tau) = 0$, $t_1 \in {}^\circ p_1$ and $V_{t_1}(\tau) = 0$. Similarly there exists $p_2 \in {}^\circ t_1$ such that $mark_{p_2}(\tau) = 0$ $t_2 \in {}^\circ p_2$ and $V_{t_2}(\tau) = 0$. This process can go ahead and we obtain p_i and t_i ($i = 1, 2, \cdots$). But the number of transitions and positions in ECEG is limited, so there exist i_1 and $i_2 > i_1$ such that $t_{i_1} = t_{i_2}$. Denote the minimum value of such i_1 as i. t_i is exactly the transition t' we want to find. □

Definition 7: Consider $t \in T$ and $V_t(\tau) = 0$ with $\tau \geq 0$, a transition $t' \in D_t(\tau)$ if it satisfies the following conditions (By Lemma 3, we ensure that these conditions are reasonable):

1) There exists an elementary path Pa from t' to t, and for every position $\overline{p}$ in Pa, $mark_{\overline{p}}(\tau) = 0$; for every transition $\overline{t}$ in Pa, $V_{\overline{t}}(\tau) = 0$.

2) There exists an elementary circuit Ci via t', and every position $\overline{p}$ on Ci satisfies $mark_{\overline{p}}(\tau) = 0$, every transition $\overline{t}$ satisfies $V_{\overline{t}}(\tau) = 0$.

In the case of $V_t(\tau) \neq 0$, a transition $t' \in D_t(\tau)$ if it satisfies the following two conditions:

1) For any $p \in {}^{\circ}t'$, $mark_p(\tau) > 0$ or the transition $t'' \in {}^{\circ}p$ satisfies $V_{t''}(\tau) > \dfrac{W(p,t')}{W(t'',p)} V_{t'}(\tau)$.

2) If the elementary path from t' to t is $Pa = (t' = t_0, p_1, t_1, \cdots p_{k-1}, t_{k-1}, p_k, t = t_k)$, then $mark_{p_i}(\tau) = 0$ and $V_{t_i}(\tau)\prod_{j=i}^{k-1} \dfrac{W(t_j, p_{j+1})}{W(p_{j+1}, t_{j+1})} = V_t(\tau) = \overline{V}(t)$.

In intuition, we may consider that $V_t(\tau)$ "depends on" $D_t(\tau)$. Note that $|D_t(\tau)| \geq 1$ for any $t \in T$. If $\overline{V}(t_1) = 1$, $\overline{V}(t_2) = 0.9$ and $\overline{V}(t_3) = 1$, we have $D_{t_1}(0) = \{t_1\}$, $D_{t_2}(0) = \{t_2\}$ and $D_{t_3}(0) = \{t_2\}$, see figure 1.

Lemma 4: Consider $t \in T$ with $t' \in D_t(\tau)$, we have $V_{t'}(\tau) \geq V_{t'}(\tau_1)$ for any $\tau_1 > \tau$.

Proof: If $V_{t'}(\tau) = 0$, according to definition 7, there exists an elementary circuit via t' with $mark_{\overline{p}}(\tau) = 0$ for all $\overline{p}$ in this circuit and $V_{\overline{t}}(\tau) = 0$ for all $\overline{t}$ in this circuit. Thus, after time τ, the firing velocity of any transition in the circuit will be identically vanishing. If $V_{t'}(\tau) > 0$, using definition 7, $V_{t'}(\tau)$ is exactly $\overline{V}(t')$. Therefore $V_{t'}(\tau_1) \leq V_{t'}(\tau) = \overline{V}(t')$ for any $\tau_1 > \tau$.

Lemma 5: $V_t(\tau)$ is a right continuous step function.

Proof: According to lemma 1, if $V_t(\tau)$ is a right continuous function, then it must be a step function. Now we only need to prove that $V_t(\tau)$ is a right continuous function. Consider $\overline{P}(\tau)$ given by definition 6:

1) If there exists $\Delta\tau > 0$ such that $\overline{P}(\tau) = \overline{P}(\tau_1)$ for any $\tau_1 \in [\tau, \tau + \Delta\tau]$, then, according to lemma 2, $V_t(\tau_1) = V_t(\tau)$ for any $t \in T$. Consequently, $V_t(\tau)$ is right continuous.

2) Otherwise, since $mark_p(\tau)$ of $p \in P$ is continuous, there exists $\Delta\tau > 0$ small enough so that $\overline{P}(\tau_1) = \overline{P}(\tau_2) \supset \overline{P}(\tau)$ for any $\tau_1, \tau_2 \in (\tau, \tau + \Delta\tau]$. Using lemma 2, it is clear that $V_t(\tau_1) = V_t(\tau_2) \geq V_t(\tau)$ holds for any transition t. Now we aim to prove $V_t(\tau_1) = V_t(\tau)$. Consider $t' \in D_t(\tau)$, we have $V_{t'}(\tau) \geq V_{t'}(\tau_1)$ by lemma 4, and therefore $V_{t'}(\tau) = V_{t'}(\tau_1)$. With definition 7, we ensure that there exists a elementary path Pa from t' to t, and $mark_p(\tau) = 0$ holds for any position p in Pa . Supposing $V_t(\tau_1) > V_t(\tau)$, there must be a position p_1 in $Pa = (t' = t_0, p_1, t_1, \cdots p_{k-1}, t_{k-1}, p_k,$ $t = t_k)$ satisfying $mark_{p_1}(\tau_1) > 0$, otherwise, by definition 4 and the invariability of $\overline{P}$ in the time interval $(\tau, \tau + \Delta\tau]$, we will come to the conclusion that $V_t(\tau_1) = V_{t'}(\tau_1)\prod_{j=0}^{k-1}\frac{W(t_j, p_{j+1})}{W(p_{j+1}, t_{j+1})} \leq \quad V_{t'}(\tau)\prod_{j=0}^{k-1}\frac{W(t_j, p_{j+1})}{W(p_{j+1}, t_{j+1})} = V_t(\tau)$, which is contradictable. Let $t_1 \in {}^{\circ}p_1$. Since the mark of p_1 varies and the firing velocity of t_1 keeps invariable in time interval $(\tau, \tau + \Delta\tau]$, we can predicate that $V_{t_1}(\tau_1) > V_{t_1}(\tau)$. Because of $V_{t'}(\tau) = V_{t'}(\tau_1)$, then $t_1 \neq t'$. Similarly, there must be a position p_2 satisfying $mark_{p_2}(\tau_1) > 0$ in Pa_1 which is from t' to t_1 as a sub-path of Pa , and a transtion $t_2 \in {}^{\circ}p_2$ such that $V_{t_2}(\tau_1) > V_{t_2}(\tau)$ and $t_2 \neq t'$. This process is endless, but the number of transition in Pa is finite, so $V_t(\tau_1) > V_t(\tau)$ is impossible. Hence, $V_t(\tau_1) = V_t(\tau)$ is true for any $t \in T$ and $\tau_1 \in (\tau, \tau + \Delta\tau]$. As a weaker property, the right continuity of $V_t(\tau)$ is also true.

Lemma 6: In an ECEG, $V_t(\tau_1) \geq V_t(\tau_2)$ for all $t \in T$ and $0 \leq \tau_1 \leq \tau_2$, that is, $V_t(\tau)$ is non-increasing.

Proof: We prove it by contradiction. Suppose that $V_t(\tau)$ is not non-increasing. Using Lemma 1 and 5, there exists a transition t and $\tau_0, \Delta\tau_1, \Delta\tau_2 > 0$ such that

$$V_t(\tau) = \begin{cases} V_1 & \tau \in (\tau_0 - \Delta\tau_1, \tau_0) \\ V_2 & \tau \in [\tau_0, \tau_0 + \Delta\tau_2) \end{cases}. \tag{7}$$

where V_1 and V_2 are two constants and $V_1 < V_2$, $\Delta\tau_1$ and $\Delta\tau_2$ are small enough so that the firing velocity of any transition within $(\tau_0 - \Delta\tau_1, \tau_0)$ and $(\tau_0, \tau_0 + \Delta\tau_2)$ is fixed. For $\tau_1 \in (\tau_0 - \Delta\tau_1, \tau_0)$, Let $t' \in D_t(\tau_1)$ and Pa be the elementary path from t' to t. Since the firing velocity of every transition is fixed in $[\tau_1, \tau_0)$ and the firing velocity of every transition in Pa is not greater than its maximum firing velocity, the mark of every position in Pa fixes at zero in $[\tau_1, \tau_0)$. And because of the continuity of the mark of positions, the mark of every position in Pa is still zero at the time τ_0. Using Lemma 4, the firing velocity of $D_t(\tau_1)$ after time τ_1 is less than or equal to that at the time τ_1. Hence, according to the algorithm in definition 3, the inequality $V_t(\tau_0) \le V_t(\tau_1)$ holds. It is contradictory.

Using Lemma 1 and Lemma 6, we have

Corollary 1: For any $t \in T$ in an ECEG, $V_t(\tau)$ can only change for finite times in the time interval $[0, +\infty)$.

Let us focus on the CEGs. In [10], we deduce the following theorem.

Theorem 3: Performances of a CEG can be represented by min-plus algebraic linear equations set. Let $\boldsymbol{x} = (M_{t_1}(\tau), M_{t_2}(\tau), \cdots, M_{t_n}(\tau))^T$ and $\boldsymbol{v} = (\bar{V}(t_1)\cdot\tau, \bar{V}(t_2)\cdot\tau, \cdots, \bar{V}(t_n)\cdot\tau)^T$, we can deduce the following min-plus linear algebraic equations

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} \oplus \boldsymbol{v} \tag{8}$$

where $\boldsymbol{A} \in \mathbb{R}_{\mathrm{mi}}^{n\times n}$ and

$$(\boldsymbol{A})_{ij} = \begin{cases} \omega & t_j \notin {}^\circ({}^\circ t_i) \\ \bigoplus\limits_p \{M_0(p)\} & t_j \in {}^\circ({}^\circ t_i) \text{ and } p \in t_j^\circ \cap {}^\circ t_i \end{cases}. \tag{9}$$

Corollary 2: The maximum solution to the equation (9) is $\boldsymbol{x} = \left(\bigoplus\limits_{i=0}^{n-1} \boldsymbol{A}^i\right)\boldsymbol{v}$. And it is unique if $\boldsymbol{A}$ has only positive weight circuits.

Proof: There is no negative value in (9) so that $\boldsymbol{A}$ has no negative weight circuit. Then using Theorem 1, 2 and 3 the corollary holds.

5 Example

A manufacturing system represented as a CEG is shown in figure 3. The real line part has input transition u and our aim is to maximize the mark consumed by output transition y at the time τ. To do so we add a position p_u to the real line part in figure 4 such that ${}^\circ p_u = p_u^\circ = \{u\}$ and $mark_{p_u}(0) = 1$.

If we introduce output feedback as imaginal line part in figure 4 with three new parameters $\bar{V}(f)$, m_1 and m_2. By Theorem 3, we have the equation

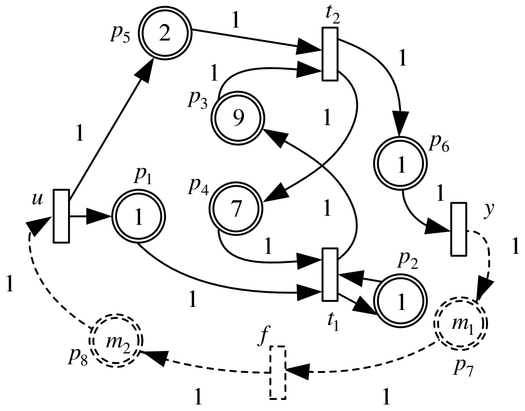


Fig. 3. Feedback Control of CEG

$$\begin{bmatrix} M_u(\tau) \\ M_{t_1}(\tau) \\ M_{t_2}(\tau) \\ M_y(\tau) \\ M_f(\tau) \end{bmatrix} = \begin{bmatrix} \omega & \omega & \omega & \omega & m_2 \\ 1 & 1 & 7 & \omega & \omega \\ 2 & 9 & \omega & \omega & \omega \\ \omega & \omega & 1 & \omega & \omega \\ \omega & \omega & \omega & m_1 & \omega \end{bmatrix} \begin{bmatrix} M_u(\tau) \\ M_{t_1}(\tau) \\ M_{t_2}(\tau) \\ M_y(\tau) \\ M_f(\tau) \end{bmatrix} \oplus \begin{bmatrix} \bar{V}(u)\cdot\tau \\ \bar{V}(t_1)\cdot\tau \\ \bar{V}(t_2)\cdot\tau \\ \bar{V}(y)\cdot\tau \\ \bar{V}(f)\cdot\tau \end{bmatrix} . \tag{10}$$

By Corollary 2, the unique solution of equation (10) is

$$\begin{bmatrix} M_u(\tau) \\ M_{t_1}(\tau) \\ M_{t_2}(\tau) \\ M_y(\tau) \\ M_f(\tau) \end{bmatrix} = \begin{bmatrix} 0 & m_1+m_2+10 & m_1+m_2+1 & m_1+m_2 & m_2 \\ 1 & 0 & (m_1+m_2+2)\oplus 7 & m_1+m_2+1 & m_2+1 \\ 2 & 9 & 0 & m_1+m_2+2 & m_2+2 \\ 3 & 10 & 1 & 0 & m_2+3 \\ m_1+3 & m_1+10 & m_1+1 & m_1 & 0 \end{bmatrix} \tag{11}$$

$$\otimes \begin{bmatrix} \bar{V}(u)\cdot\tau \\ \bar{V}(t_1)\cdot\tau \\ \bar{V}(t_2)\cdot\tau \\ \bar{V}(y)\cdot\tau \\ \bar{V}(f)\cdot\tau \end{bmatrix} .$$

which imply that mark of every position in figure 4 is finite.

Moreover, consider the minimum capacity at position

$$\begin{aligned}
&\max_{\tau}\{mark_{p_1}(\tau)\} = \max_{\tau}\{mark_{p_1}(0) + M_u(\tau) - M_{t_1}(\tau)\} \\
&= \max_{\tau}\{1 + (2.2\cdot\tau) \oplus (m_1 + m_2 + 10 + 2\cdot\tau) \oplus (m_1 + m_2 + 1 + 2.1\cdot\tau) \\
&\oplus (m_1 + m_2 + 2.15\cdot\tau) \oplus (m_2 + \bar{V}(f)\cdot\tau) - (2\cdot\tau) \oplus (m_2 + 1 + \bar{V}(f)\cdot\tau)\} \qquad (12) \\
&= \begin{cases} 1 & \bar{V}(f) < 2 \\ m_2 + 1 & \bar{V}(f) = 2 \\ m_1 + m_2 + 10 & \bar{V}(f) > 2 \end{cases}
\end{aligned}$$

when $\bar{V}(f) = 2$ the capacity required is minimum if $m_2 = 0$, and when $\bar{V}(f) > 2$ the capacity required is minimum if $m_1 = m_2 = 0$. Let $\bar{V}(f) = 2 + e$, where e is error, then the capacity required of p_1 is at least $m_1 + m_2 + 10$.

References

1. Cohen, G., Bubois, V., Quadrat, J. P., et al: A Linear System-theoretic View of Discrete-event Processed and It Use for Performance Evaluation in Manufacturing. IEEE Trans Automatic Control, Vol 30, No. 3, Mar. (1985) 210 – 220
2. Baccelli, F., Cohen, G., Olsder, G. J., Quadrat, J. P.: Synchronization and Linearity: An Algebra for Discrete Event Systems. Wiley, New York (1992)
3. Chen, W., Qi, X.: Discrete Event Dynamic Systems: A Max-Plus Algebraic Approach. Science Press, Beijing (1994)
4. Zheng, D., Zhao, Q.: Discrete Event Dynamic Systems, Tsinghua University Press, Beijing (2001)
5. Cofer, D. D., Garg, V. K.: Supervisory Control of Real-time Discrete Event Systems Using Lattice Theory. IEEE Trans Automatic Control, Vol 41, No. 2, Feb (1996) 199 – 209
6. Takai, S.: A Characterization of Realization Behavior in Supervisory Control of Timed Event Graphs, Automatic, Vol 33, No. 11, Nov (1997) 2077-1080
7. Cohen, G., Gaubert, S., Quadrat, J. P.: Timed-events Graphs with Multipliers and Homogeneous Min-plus Systems, IEEE Trans. Automat.Contr., Vol 43, No. 9, Sept (1998) 1296–1302
8. Dai, H., Sun, Y.: An Algebraic Model for Performance Evaluation of Timed Event Multigraph, IEEE Trans. Automat. Contr., Vol 48, No. 7, pp.1227–1230, Jul. 2003.
9. Chen, W.: Analyses of Extended Timed Event Graphs, Acta Mathematicae Applicatae Sinica, Vol 26, No. 3 (2003) 451-457
10. Zhang, D., Dai, H., Sun, Y.: An Algebraic Model for Performance Evaluation of a Class of Continuous Petri Nets. IEEE Conference on Systems, Man, and Cybemetrics, (2004) 4965-4970
11. David, R., Alla, H.: Petri Nets Grafcet Tools for Modeling Discrete Event Systems. Prentice-Hall, London (1992)
12. David, R., Alla, H.: Petri Nets for Modeling of Dynamic Systems, Automatic, Vol 30, No. 2 (1994) 175-202

Combined Literature Mining and Gene Expression Analysis for Modeling Neuro-endocrine-immune Interactions

Lijiang Wu and Shao Li*

MOE Key Laboratory of Bioinformatics, Department of Automation, Tsinghua University,
Beijing 100084, P. R. China
wulj03@mails.tsinghua.edu.cn
shaoli@tsinghua.edu.cn
http://www.au.tsinghua.edu.cn/szll/lishao/default.htm

Abstract. Here we develop a new approach of combined literature mining and gene expression analysis (CLMGE) to model the complex neuro-endocrine-immune (NEI) interactions. By using NEI related PubMed abstracts and the Human Genome Organisation gene glossary for subject oriented literature mining (SOLM), it is found that the NEI model serves as a scale-free network and the degree of nodes follows a power-law distribution. Then we evaluate and eliminate the redundant of SOLM-based NEI model by multivariate statistic analysis basing on selected gene expression data. Each involving expression data is tested by cross validation with Leave One Out strategy. The results suggest that the performance of CLMGE approach is much better than that of SOLM alone. The integrated strategy of CLMGE can not only eliminate false positive relations obtained by SOLM, but also form a suitable solution space for analyzing gene expression data. The reasonable biological meanings of the CLMGE-based NEI model are also evaluated and demonstrated by classifying its sub-functions according to DAVID and SwissProt databases.

1 Introduction

The neuro-endocrine-immune (NEI) system [1], the central component of the complex human biological stress response, controls host defenses and homeostasis at the molecular level. A great amount of evidence regarding to this theme has been accumulated rapidly. However, the complex actions of NEI are still mysterious which beyond the scope of any individual experiment. Now many integrated approaches in the field of bioinformatics are developed, such as literature mining (LM) [2] which retrieves information contained in valuable literatures and almost is free of the limitation of individual experimental conditions. Our previous work has constituted a preliminary approach of subject oriented literature mining (SOLM) [3] for constructing the network special for a given biological system. While on the gene expression level, microarray datasets can reflect the activities of thousands of genes on a large scale; gene expression information may provide a chance to explore

* Corresponding author.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 31 – 40, 2005.

cooperative relations of the whole system according to the hypothesis that genes co-expressed would be co-regulated [4].

Unfortunately, either LM-based or microarray data-based approach has its limits for the network construction. For example, the LM derived network is relatively crude and has many redundancies. The co-occurrence method is unable to represent various types of relations such as the direct-ratio or the inverse-ratio. On the other hand, for the microarray data, it is difficult to deal with thousands of variables directly to construct a complex network. In view of the SOLM model forms a global solution space for further analyzing microarray data, and the microarray data may appropriately deal with the problems of validating the confidence of network and differentiating many types of relations. With the goal of taking advantage of both methods whereas avoid their limits to promote the ability for modeling complex biological systems, here we develop a Combined Literature Mining and Gene Expression analysis (CLMGE) approach to construct and analyze the NEI interactions.

2 Methods

2.1 Subject Oriented Literature Mining

The Human Genome Organisation (HUGO) glossary (http://www.gene.ucl.ac.uk), which includes about 20,000 none redundant gene symbols, is prepared for SOLM. A total of 21229 PubMed abstracts (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi, 2005/03/01) are retrieved by using 11 keywords of NEI: “nerve-endocrine-immune”, “hypothalamic-pituitary”, “neuroendocrinology”, “neuro-endocrine-immune”, “neuroimmunology”, “neuroimmunomodulation”, “immune-neuroendocrine”, “neuroimmunoendocrinology”, “psychoneuroimmunology”, “nuclear factor-kappaB pathway” and “pineal-immune”. We defined the co-occurrence as a pair of symbols co-occurring in a sentence. Then took count of the co-occurrence of all symbol pairs to form the weighted matrix, in which a row or column is a gene and each entry is the number of co-occurrence of each pair genes. Note that the symbols are unified as the corresponding unique names when counting the co-occurrence. Given importance and frequency of the symbol, the co-occurrence should not share the same significance. So the weighted matrix is normalized as: $D = (d_{ij})_n$, where $d_{ij} = 2c \big/ (\sum_{k=1}^{n} c_{ik} + \sum_{k=1}^{n} c_{kj})$, and $(d_{ij}) \in [0,1]$; and the adjoint matrix, $R = (r_{ij})_n$, is defined as

$$r_{ij} = \sum_{k=1}^{n} \frac{2d_{ik}{}^2 d_{kj}{}^2}{(d_{ik})^2 + (d_{kj})^2} \tag{1}$$

2.2 Microarray Data and Missing Value Estimation

The NEI-related B-cell lymphoma microarray dataset, including about 14,000 genes and 133 gene expression experiments, is downloaded from the Stanford Microarray

Database (SMD, http://genome-www5.stanford.edu/), which stores raw and normalized data from microarray experiments, and provides data retrieval, analysis and visualization interfaces. The missing values in this dataset are estimated using the K-nearest neighbors method [5]. Briefly, for x_{ij}, a missing value of gene $\mathbf{x_i}$ in a observation *j*, we can find the other K genes which have the most similar expression profile by the Pearson correlation analysis. Then the missing value is recovered with the weighted mean of the corresponding observation of the K genes.

2.3 Multivariate Statistic Analysis for Gene Expression Data

We take the expression level of *n* genes as variable **X₁, X₂, …, Xₙ**, the dataset with *m* observations and the n variables, is denoted by

$$[\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{2}$$

The relationship between gene k and l can be formulated as $d_{kl} = d(\mathbf{x_k}, \mathbf{x_l})$ for the adjoint matrix. So far, the most commonly used definition of $d(\mathbf{x_k}, \mathbf{x_l})$ is the Pearson coefficient of the variable pair $(\mathbf{x_k}, \mathbf{x_l})$. The true relations between d_{kl_1} and d_{kl_2}, however, may not be independent. The interactions among multiple variables should be considered.

Suppose the network follows a linear model [6], we can build the model through multiple variables selection. If the prior knowledge about the network is unavailable, there are two problems need to deal with. First, it is not advisable to utilize the method directly for thousands of candidate variables. The next, for the limitation of the algorithm, the number of variables should not exceed the number of observations to avoid matrix singularity. Therefore, it is necessary to get some prior information from the NEI network relating literatures.

For the linear model, $x_k = \beta_{k0} + \beta_{k1}x_{k1} + \beta_{k2}x_{k2} + \cdots + \beta_{kl}x_{kl} + e$, where the $x_{k1}, x_{k2}, \cdots, x_{kl}$ are the neighbor nodes of x_k in the SOLM-based NEI model, and e is the random error. Because of the “Over Fitting”, many variables in the model are false positive and invalid. The significance of the model is verified using F-test, defined as

$$F_H = \frac{SSR/l}{SSE/(m-l-1)} \tag{3}$$

where SSE and SSR represent the residual of random error and the variables of the entire model; m, l are the number of observations and the number of the variables of entire model respectively. The P-value is defined as $p = P(F_{l,m-l-1} > F_H)$, in which F_H obeys the distribution of $F(l, m-l-1)$.

Following step-wise model selection, for each attributive variable, we define its neighbor nodes in the SOLM-based NEI model as the candidate variables of the complete model. Starting from null model, we add one variable at the beginning of each iteration, then try to delete the invalid variable influenced by the new variable, and finally calculate the P value of each updated model. Repeat above steps for all the candidate variables until no variable can be added or deleted. Accordingly, a credible model with a lower P-value (<0.01) can be obtained.

2.4 Cross Validation with Leave One Out (LOO) Method

Cross validation with LOO strategy is used to estimate the gene expression of both the SOLM-based and the CLMGE-based NEI models. At each iteration of LOO, we omit one observation j of the gene $\mathbf{x_i}$ and its neighbor ones $\mathbf{x_{i1}}, \mathbf{x_{i2}}, \cdots, \mathbf{x_{il}}$, then reconstruct the linear network based on the remaining observations $\mathbf{x_{-ji}}$ and $\mathbf{x_{-ji1}}, \mathbf{x_{-ji2}}, \cdots, \mathbf{x_{-jil}}$. Therefore, the omitted x_{ji} can be reevaluated by the corresponding observations of neighbor genes.

Next, we evaluate the square sum of errors for all gene expression values in both models. It can be formulated as $SSE_{\text{mod}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ji} - \hat{x}_{ji}\right)^2$, in which x_{ji} is the true expression and $\hat{x}_{ji}$ is the reevaluated expression. And we define the Z-score of mean square error, Z_{mod}, as

$$Z_{\text{mod}} = \frac{\sqrt{SSE}/mn - E\{x_{ji}\}}{STD\{x_{ji}\}} \tag{4}$$

where $E\{x_{ji}\}$ is the expectation and $STD\{x_{ji}\}$ is the standard deviation of the gene expression.

2.5 Describing Functions from DAVID and SwissProt Databases

Finally, we submit the 350 genes into the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://apps1.niaid.nih.gov/david/), a web provide integrated solutions for the annotation and analysis of genome-scale datasets derived from high-throughput technologies such as Microarray and proteomic platforms. Then we divide the sub-function of genes into three systems of immune, endocrine and nerve, as defined in the DAVID and the SwissProt databases (http://www.hgmp.mrc.ac.uk/Bioinformatics/Databases/swissprot-help.html).

3 Results

3.1 Network of SOLM-Based NEI Model

First, we constructed SOLM-based NEI model from 21229 PubMed abstracts with a threshold of co-occurrence as 2, which including 925 genes as nodes and 3675 relations as edges. The average degree is 7.9. The degree of nodes follows a power-law distribution, obviously appearing as a scale-free network. In the top 20 hub nodes whose degrees are more than 50, there are 16 nodes fall into the sub-function classifications, namely neuro-system (8 nodes), immune-system (7 nodes) and endocrine-system (10 nodes), according to DAVID.

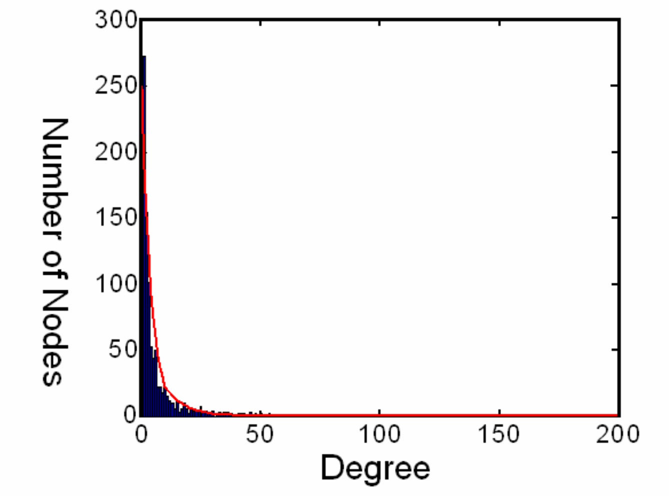


Fig. 1. The degree of nodes in SOLM-based NEI model follows a power-law distribution. The max degree is 197 and the corresponding node is tumor necrosis factor (TNF). Degree of heparanase (HPSE), corticotropin releasing hormone (CRH), and interleukin (IL)-6 is 123, 105, and 103 respectively.

3.2 Comparison of Both SOLM-Based and CLMGE-Based NEI Networks

Based on the SOLM-based NEI model, the overlapped 380 genes and 1196 edges are extracted from the prior model and the microarray dataset. A predominant sub-network (Fig. 2A) of 350 nodes and 1162 edges is selected for further analysis. For variables of each node along with its neighbors, we use the microarray data by statistic analysis to eliminate the redundant edges and form an updated model (Fig. 2B) with a threshold of P <0.01. At the same time, we calculated the P-value of each model and find that the CLMGE-based NEI model is much more believable than the previous one. By comparing both networks, we found that the average path length is almost uninfluenced though the average of degrees varies obviously, indicating that redundant edges are eliminated while the essentials are remained.

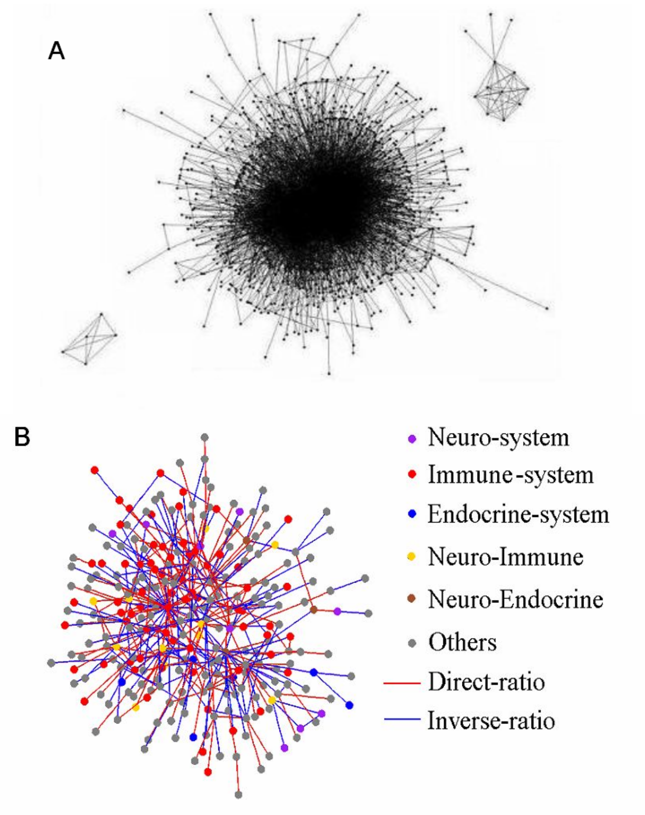


Fig. 2. Predominant sub-networks of SOLM-based (A) and CLMGE-based (B) NEI models. Sub-sets of neuro-system, immune system, endocrine-system, neuro-immune and neuro-endocrine are classified according to DAVID; "Others" include genes with functions such as signal, glycoprotein, and phosphorylation.

Table 1. Summary of the comparison between the SOLM-based NEI network and the CLMGE-based NEI Network

Networks	Number of Nodes	Number of Edges	Average Degrees	Average Shortlength
SOLM-based NEI network	861	3621	8.40	6.4432
CLMGE-based NEI network	276	426	3.09	8.4585

As shown in Table 1, both SOLM-based and CLMGE-based NEI networks have 861 versus 276 nodes, 3621 versus 426 edges, 8.4 versus 3 average degrees, and 6.4 versus 8.4 average path length (geodesic distance). The results suggest that the edges elimination do not impact seriously the average path length although affect the average degrees, comparing with the other figure. In other words, the eliminated edges are almost inessential edges, and the backbone of the network is retained.

3.3 Cross Validation for Both Networks

Next, cross validations for both SOLM-based and CLMGE-based NEI Networks are carried out by using the method of Leave One Out. It is found that the SSE_{mod} for SOLM-based or CLMGE-based NEI model is 1.25e+4 or 1.21e+4, and the Z value is 0.443 or 0.428 (equation (4)) respectively. The gene expression estimations of TNF, the hub node with the highest degrees in both networks, are illustrated in Figure 3. Comparison and estimation for hub nodes such as CRH, IL6, and IL2 as well as the node with lowest degree such as oncostatin M receptor (OSMR) are summarized in Table 2. The results indicate that the performance of CLMGE method is much better than that of SOLM.

Table 2. Comparison and estimation results for four nodes and their sub-models of CRH, IL6, IL2, and OSMR in both SOLM-based and CLMGE-based NEI Networks

Genes	Sub-functions	Degrees of the node (SOLM)	Degrees of the node (CLMGE)	P value (SOLM)	P value (CLMGE)
CRH	Neuro-Endocrine	105	13	1.11e-5	<1.00e-7
IL6	Neuro-Immune	103	6	1.38e-3	<1.00e-7
IL2	Neuro-Immune	54	7	7.47e-4	<1.00e-7
OSMR	Others*	3	2	<1.00e-7	<1.00e-7

* “Others” represents genes with functions such as signal, glycoprotein, and phosphorylation.

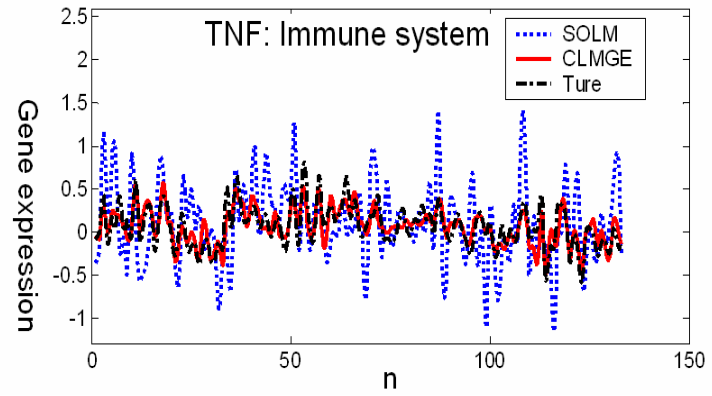


Fig. 3. The estimation result of expression of gene TNF (Id848), which falls into the immune system, with the degree of 197 in SOLM-based NEI model and the degree of 16 in CLMGE-based NEI model. The Y-axis represents the number of microarray experiments. And the P-value of these sub-models are P_{SOLM} =1.70e-3 and P_{CLMGE} <1.00e-7 by F-test.

4 Discussions

Building a network for complex biological systems has been the focus of extensive research for decades and is still a challenging problem. Up to now, there are many types of complex networks in biological systems including protein-protein interactions [7], genetic regulatory networks [8], signal transduction pathways [9] and metabolic networks [10]. For the complex NEI system, however, there is less prior information about the network topologies or other properties. It is almost impossible to directly search the entire parameter space to generate the NEI network. Thus, it is within this context that we preliminarily establish a new integrative and statistical approach, CLMGE, to constructing biological network maps for NEI interactions.

SOLM, the first step towards constructing NEI network in the present work, is based on co-citation, which shares the assumption with many existing literature mining systems [2,11,12]. This approach holds that when two genes are co-cited in the same text unit, there should be an underlying biological relationship. Next, CLMGE method developed in this work may be a way to model NEI interactions by applying SOLM to extracting all the gene interactions from literatures, using gene expression analysis for validation, and eliminating the statistically independent relations.

The SOLM method specifically designed to collect and extract PubMed abstracts-derived information about NEI includes the discoveries of many researchers. It could be provide a holistic frame of the NEI network, and forms a wide solution space. In the SOLM-based NEI network, it is found that hub nodes such as TNF, IL-6 and CRH are main players involving in the NEI functions and also taking important roles during the course of many pathological changes and complex diseases such as inflammation, infection, stress and autoimmune disorders [13,14,15].

On assuming that the SOLM method could address most of the relations from literatures, the approach of CLMGE will theoretically reach the global optimum solution by eliminating the false positive relations. Moreover, the approach of CLMGE can draw more detailed information such as relation types, intensities, and directions from the gene expression data. In this work, the microarray data analysis is testified for reconstructing the complex NEI network by comparing both SOLM-based and CLMGE-based models before and after the processing of the multivariate analysis. Although the selected microarray dataset is public and only parts of them are overlapped with NEI literature derived data, our results suggest that the integrated approach of CLMGE raise to a more desirable quality for NEI network when comparing with that of SOLM alone. On the basis of the present work, not only the false positive relations are eliminated remarkably but also the backbone of the whole network is uncovered. This remarkable improvement is mostly resulted from the elimination of redundant relations and avoiding "Over Fitting". The more candidate variables, the more precise the model would be, but the risk of extrapolation would be expanded accordingly. For complex biological system, the false positive relations may be associated with indirect interactions, which are misapprehended as direct interactions by the method of literature mining.

As the statistic analysis can only eliminate redundant relations but may not discover additional relations, the threshold of co-occurrences should be relative low. On the other hand, we can choose a proper threshold for the variable selection accordingly. The microarray dataset used in this work is public and not special for our purpose, so we can only utilize a part of this dataset that overlapped with the LM derived data. On the precondition of having more appropriate microarray datasets, we believe that this integrated strategy of combined the literature mining and gene expression analysis would be much more precise and promising for reconstructing complex biological networks.

5 Conclusions

The integrated strategy of CLMGE can not only eliminate the false positive relations obtained by literature mining, but also form a suitable solution space for analyzing gene expression data. The approach of CLMGE may be helpful for further understanding the topologies and biological meanings of complex NEI interactions.

Acknowledgements

We would like to express our appreciation to Dr. Zhongqi Zhang, Chaolin Zhang, Tao Peng and Ying Huang for their helpful discussions and comments. This work is supported by the National Natural Scinece Foundation of China (Nos. 30200365 and 90209002), the Foundation for the Authors of National Excellent Doctoral Dissertations of China (No. 200366), and the National Basic Research (973) Program (No. 2003CB517106).

References

1. Besedovsky, H.O., Sorkin, E.: Network of Immune-neuroendocrine Interactions. Clin. Exp. Immunol 27 (1977) 1-12
2. Jenssen, T.K., Laegreid, A., Komorowski, J., Hovig, E.: A Literature Network of Human Genes for High-throughput Analysis of Gene Expression. Nat. Genet. 28 (2001) 21-28
3. Zhang, C., Li, S.: Modeling of Neuro-endocrine-immune Network Via Subject Oriented Literature Mining. Proceedings of BGRS 2 (2004) 167-170
4. D'haeseleer, P., Liang, S., Somogyi, R.: Genetic Network Inference: from Co-expression Clustering to Reverse Engineering. Bioinformatics 16 (2000) 707-726
5. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing Value Estimation Methods for DNA microarrays. Bioinformatics 17 (2001) 520-525
6. D'Haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear Modeling of mRNA Expression Levels during CNS Development and Injury. Pac. Symp. Biocomput. (1999) 41-52
7. Ramani, A.K., Bunescu, R.C., Mooney, R.J., Marcotte, E.M.: Consolidating the set of known Human Protein-protein Interactions in Preparation for Large-scale Mapping of the Human Interactome. Genome Biol. 6 (2005) R40
8. Halfon, M.S., Michelson, A.M.: Exploring Genetic Regulatory Networks in Metazoan Development: Methods and Models. Physiol. Genomics. 10 (2002) 131-43
9. Lappe, M., Holm, L.: Algorithms for Protein Iinteraction Networks. Biochem. Soc. Trans. 33 (2005) 530-534

10. Sun J., Zeng, A.P.: IdentiCS - Identification of Coding Sequence and in Silico Reconstruction of the Metabolic Network directly from Unannotated Low-coverage Bacterial Genome Sequence, BMC Bioinformatics 5 (2004) 112
11. Stapley, B.J., Benoit G.: Information Retrieval and Visualization from Co-occurrences of Gene Names in Medline Abstracts. Pac. Symp. Biocomput. (2000) 529-540
12. Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., Mostafa, J.: Detecting Gene Relations from Medline Abstracts. Pac. Symp. Biocomput. (2001) 483-495
13. Li, S., Lu, A.P., Li, B., Wang, Y.Y.: Circadian Rhythms on Hypothalamic-pituitary-Adrenal Axis Hormones and Cytokines of Collagen Induced Arthritis in rats. J. Autoimmun.22 (2004) 277-285
14. Wilder, R.L.: Neuroimmunoendocrinology of the Rheumatic Diseases, past, present, and future. Ann. N. Y. Acad. Sci. 966 (2002) 13-19
15. Russell, S.H., Small, C.J., Stanley, S.A., Franks, S., Ghatei, M.A., Bloom, S.R.: The In vitro Role of Tumour Necrosis Factor-alpha and Interleukin-6 in the hypothalamic-pituitary gonadal axis. J. Neuroendocrinol 13 (2001) 296-301

Mean Shift and Morphology Based Segmentation Scheme for DNA Microarray Images

Shuanhu Wu[1], Chuangcun Wang[1], and HongYan[2,3]

[1] School of Computer Science & Technology, Yantai University,
30 QingQuan Rd., Yantai 264005, China
wushuanhu@163.com

[2] Department of Computer Engineering and Information Technology,
City University of HongKong, Kowloon, Hong Kong
h.yan@cityu.edu.hk

[3] School of Electrical and Information Engineering,
University of Sydney, NSW 2006, Australia

Abstract. Image segmentation is supposed to be the most important step in microarray image analysis. In this work, we proposed a new template-based segmentation method for DNA microarray images. Different from the local-based segmentation techniques adopted by all the available analysis softwares, our algorithm segments images from global view of point. Based on mean shift filtering technique, we first segmented image into some different homogenous regions in which all the spots appeared as different local maximum regions. Then an initial spot segmentation template was extracted by morphological H-reconstruction. Finally, a refined spot segmentation template was obtained by histogram analysis. Experimental results showed that our algorithm is robust and can obtain accurate spot segmentation results. Especially, compared to all the available algorithms, our template-based spot segmentation scheme not only can facilitate downstream intensity extraction step but also can be very helpful to improve the accuracy of intensity extraction.

1 Introduction

DNA microarrays allow the monitoring of expressions from tens of thousands of genes simultaneously. This technology makes use of the so-called hybridization reaction in which two segments of single-strand DNA bind together or hybridize if the bases on one strand are complementary to the bases on the other strand [1]. Applications of microarray technology range from the study of gene expression in yeast under different environmental conditions to the comparison of gene expression profiles of tumors from cancer patients. In addition to the enormous scientific potential in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research.

In a microarray experiment, two samples of DNA, which are reversed transcribed from mRNA , are labeled with different fluorescent dyes (usually Cy3 and Cy5) to constitute the cDNA probes. The two cDNA probes are then hybridized onto a DNA microarray, which holds hundreds or thousands of spots, each of which contains a

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 41 – 50, 2005.

known different DNA sequence. If a probe contains a cDNA whose sequence is complementary to the DNA on a given spot, that cDNA will hybridize to the spot, where it will be detectable by its fluorescence. Spots with more bound probes will have more fluorescent dyes and will therefore fluoresce more intensely. The ratio of the two fluorescence intensities at each spot indicates the relative abundance of the corresponding DNA sequence in the two cDNA samples that are hybridized to the DNA sequence on the spot. Fig.1(a) is an example of a combined gray microarray image representation, where the red channel sets as the Cy5 image, the green channel sets as the Cy3 image, and the blue channel is set to zero. In this image, one can see the differential expression of genes in the two cDNA samples. Intense red fluorescence at a spot indicates a high level of expression of the gene labeled using the red-fluorescent Cy5 dye, with little expression of the gene labeled using the green fluorescent Cy3 dye. Conversely, intense green fluorescence at a spot indicates a high level of expression of the gene labeled using the Cy3 dye, with little expression of the gene labeled using the Cy5 dye. When the genes in both samples express at similar levels, the observed spot is yellow.

Image analysis is a very important procedure for in a microarray experiment. In general, it includes three tasks: gridding, segmentation, and information extraction in which segmentation is a crucial step for accurate information extraction. A number of microarray image processing techniques have been proposed and implemented in several commercial software and freeware packages in the last several years, such as ScanAlyze [2], GenePix [3], Spot [4], QuantArray [5] and the newly developed processing techniques [6], [7]. Existing segmentation methods for microarray images can be grouped into three categories according to the geometry of the spots they produce fixed circle segmentation, adaptive circle segmentation and adaptive shape segmentation. Fixed circle segmentation fits a circle with a constant diameter to all the spots in the image. The advantages of this method are that it is easy to implement and that it works correctly when all the spots are circular and of the same size [2]. The adaptive circle segmentation is adopted by GenePix[3] in its early version, where the circle's diameter is estimated separately for each spot. Although latest GenePix Pro. 6.0 [3] has no restrains on spot shape, it still need to give appropriate spot's diameter before processing. Except for those features, the above algorithms are all based on local segmentation methods after image gridding. In local segmentation methods, it may be difficult to determine whether a spot is present or absent if the spot is bigger than the corresponding area in the grid or does not exist at all.

In this paper, we propose a new template-based spot segmentation algorithm by mean shift filtering technology and mathematical morphological operator. Different from the local-based segmentation algorithm adopted by all the available algorithms, our algorithm segments images from global point of view. Firstly, based on mean shift filtering, we segmented image into some homogeneous regions in which all the spots appeared as local maximum regions. Then an initial segmentation template was extracted by morphological H-reconstruction. Finally, a refined segmentation template was obtained by histogram statistical analysis. Comparison with available algorithm, our algorithm does not place any restriction on the spot shapes and intensity distribution. Especially, the template-based spot segmentation algorithm can facilitate the downstream information extraction process and also can be very helpful to improve the accuracy of intensity extraction.

2 Segmentation Method Based on Mean Shift Filtering

2.1 The Mean Shift Procedure

The mean shift procedure is based on kernel density estimation and was proposed in 1975 by Fukunaga and Hostetler [8] for mode detection and clustering and recently rekindled in Cheng's paper [9]. Given *n* data points $x_{i,}$ i=1~n in the d-dimensional space R^d, the multivariate kernel density estimator with kernel *K(x)* with bandwidth *h*, computed in the point x is given by

$$\hat{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K(\frac{x-x_i}{h}) \tag{1}$$

The d-variate kernel $K(x)$ is a bounded function with compact support satisfying:

$$\begin{array}{ll} \int_{R^d} K(x)dx = 1 & \lim\limits_{\|x\|\to\infty} \|x\|^d \ K(x) = 0 \\ \int_{R^d} xK(x)dx = 0 & \int_{R^d} xx^T K(x)dx = c_K I \end{array} \tag{2}$$

where c_K is a constant. The multivariate kernel may be generated from different ways, but here we only interested in a special class of radially symmetric kernels satisfying

$$K(x) = a_{k,d} k(\|x\|^2) \tag{3}$$

in which it suffices to define the function $k(x)$ called the profile of the kernel, only for x>0, and $a_{k,d}$ is a strictly positive constant which makes $K(x)$ integrate to one.

The quality of a kernel density estimator is measured by the mean of the square error between the density and its estimate, integrated over the domain of definition. In practice, however, only an asymptotic approximation of this measure (denoted as AMISE) can be computed. For kernel function in (3), the AMISE measure is minimized by the Epanechnikov kernel [10] having the profile

$$k(x) = \begin{cases} 1-x & 0 \le x \le 1 \\ 0 & x > 1 \end{cases} \tag{4}$$

Employing the above profile, the density estimator (4) can be rewritten as

$$\hat{f}_{h,K}(x) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k(\| \frac{x-x_i}{h} \|^2) \tag{5}$$

The first step in image segmentation with the underlying density $f(x)$ is to find the modes of each pixel. The modes are located among the zeros of the gradient $\nabla f(x)$=0 and the mean shift procedure is an elegant way that will converge to these zeros without estimating the density along following mean shift vector[11]:

$$m_{h,G}(x) = \frac{\sum_{i=1}^{n} x_i g(\| \frac{x - x_i}{h} \|^2)}{\sum_{i=1}^{n} g(\| \frac{x - x_i}{h} \|^2)} - x \tag{6}$$

where $g(x)=-k^{'}(x)$ and $\sum_{i=1}^{n} g(\| (x-x_i)/h \|^2)$ is a positive number.

Mean shift procedure above is also named as mean shift filtering. By this procedure, each pixel can find its own mode (or cluster) and then image can be segmented into some different homogeneous regions. Especially, mean shift filtering is a procedure of discontinuity preserving smoothing that is a very crucial for accurate image segmentation.

2.2 Feature Space and Kernel

The feature space for image can be represented as a two-dimensional lattice of p-dimensional vectors (pixels), where p=1 in the gray-level case, p = 3 for color image. The space of lattice is known as the spatial domain, while the gray level, color is represented in the range domain. When the location and range vectors are concatenated in the joint spatial-range domain of dimension d= p+2, their different nature has to be compensated by proper normalization. Thus the multivariate kernel is defined as the product of two radially symmetric kernels and the Euclidean metric allows a single bandwidth parameter for each domain

$$K_{h_s,h_r}(x) = \frac{C}{h_s^2 h_r^p} k(\| \frac{x^s}{h_s} \|^2) k(\| \frac{x^r}{h_r} \|^2) \tag{7}$$

where x^s is the spatial part, x^r is the range part of a feature vector, k(x) the common profile used in both two domains, h_s and h_r the employed kernel bandwidths, and C the corresponding normalization constant. In practice, we adopted Epanechnikov kernel because of its fast speed and excellent performance, what the user has to do is to set the bandwidth parameter $h = (h_s , h_r)$.

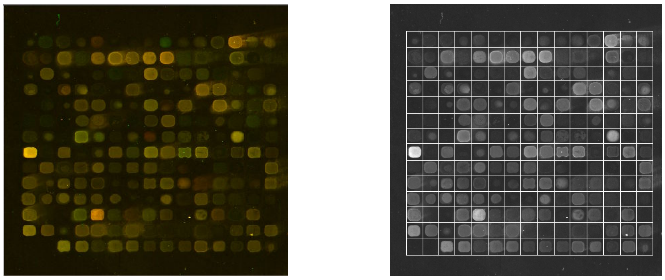
2.3 Microarray Image Preprocessing and Bandwidth Computation

The microarray images consist of a pair of 16-bit images in TIFF format scanned by laser with different wavelength. We denote the two input images as “R” and “G” for “red” and “green”, with R corresponds to the red-fluorescent dye Cy5 and G corresponds to the green-fluorescent dye Cy3, respectively. For improving computational efficiency, we transform them into a single 8-bit combined gray image *RG*, given by

$$RG(i,j) = max(R^{'}(i,j), G^{'}(i,j)) \tag{8}$$

where $R^{'}$ and $G^{'}$ are obtained by a square-root transformation of the original inputs, *R* and *G*, respectively.

Another crucial issue is determining the bandwidth parameters. Here we proposed an automatic method for determining the bandwidth parameter $h = (h_s , h_r)$. Firstly, we gridding combined 8-bit gray image RG by mathematic morphology technique proposed by Hirata[12](see Fig.1(b)). Then we calculate the variance of pixels in a central small regions (small than spots) in each grid and take the median as the bandwidth parameter h_r. Obviously, bandwidth parameter, h_s, should be smaller than the size of objective (diameter of spot). In practice, we have found that taking h_s=2, i.e. its detecting neighborhood is 5*5, is very robust for segmenting a variety of microarray images, even for noisy images.



(a) (b)

Fig. 1. DNA microarray image and gridding. (a) an example of microarray images; (b) an example of microarray image gridding.

2.4 Microarray Image Segmentation by Mean Shift Filtering

Let x_i and z_i , i = 1,2,…,n, be the d-dimensional input and filtered image in the joint spatial-range domain, respectively. In our case, the input is a combined 8-bit gray image, so the feature space is a two-dimensional lattice of 1-dimensional vectors (pixels). Based on the kernel (7), the segmentation process by mean shift filtering can be implemented as follows:

1. Initialize j = 1 and $y_{i,1} = x_i$;
2. Run the mean shift filtering procedure by kernel (7) and compute $y_{i,j+1} = y_{i,j} + m_{h,G}$ according to (6) until it converges to a stationary point $y = y_{i,c}$;
3. Assign $z_i = (x_i^{s}, y^{r}_{i,c})$, where the superscripts s and r denote the spatial and range components of a vector, respectively; if $i = n$, goto step 4, otherwise goto step 2.
4. Grouping together all z_i which are closer than h_s in the spatial domain and h_r in the range domain;
5. Eliminate spatial regions containing less than M pixels (used to filter out noise).

By above filtering procedure, an image can be segmented into some homogeneous regions where each spot appeared as local maximum region in the dark background (see Fig.3(b)).

3 Spot Segmentation Template Extraction

The final aim of microarray images analysis is extracting the gene expressions, so we must accurately know the position of each spot. In this section, we introduce our spot segmentation extraction method that can accurately locate the positions of spot and its local background based on mathematical morphology and histogram analysis.

3.1 Initial Template Extracting by Morphological H-Reconstruction

By above mean shift filtering technique, an image can be segmented into some homogeneous regions and all the spots appear as some local maximum regions. These local maximum regions can be accurately detected by morphological H-reconstruction technique [13] or contrast opening of *f* of size *H*. H-reconstruction or construct opening defined as

$$\gamma_H^d(f) = \gamma^{rec}(f, f - H) \tag{9}$$

Where γ^{rec} is the morphological reconstruction operator by dilation and is defined by

$$\gamma^{rec}(g, f) = \delta_g^i(f) \tag{10}$$

such that for some *i*, $\delta_g^i(f) = \delta_g^{i+1}$ that is implemented using the geodesic dilation operator based on restricting the iterative dilation of a function marker *f* by structuring element *B* to a function mask *g* by

$$\delta_g^i(f) = \delta_g^1 \delta_g^{i-1} \tag{11}$$

where $\delta_g^1(f) = \inf(\delta_B(f), g)$. Based on (9), the local maxima in an image can be extracted by setting $H = 1$ by

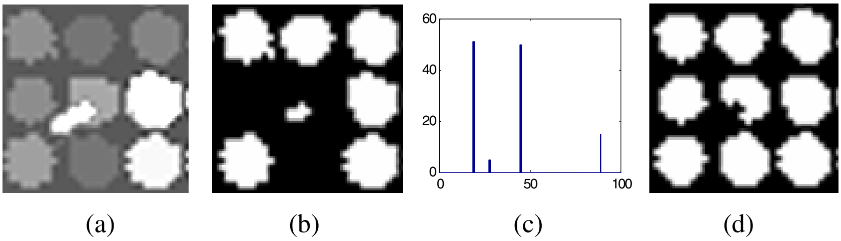
$$Maxima(f) = f(x) - \gamma_1^d(f) \tag{12}$$

According to (12) and taking the segmentation results obtained by mean shift filtering as input, an initial binary spot segmentation template can be generated (similar to Fig.3(c)) where the regions with the value of 1 represent spots and rest of regions with the value of 0 represents background. An important property is that the filters by reconstruction involve the notion of connectivity and can preserve the 'edges' of the structures that is very desired for segmenting images accurately.

3.2 Template Refining by Histogram Analysis

For most of spots, above segmentation results are very satisfactory, but there still exist a very few spots that can't be correctly detected due to the effect of the noise. In Fig.2(b), we showed an example. It is obvious that three spot don't be correctly detected: the central spot, the spot at the left and the one under the center spot. The reason resulting this phenomenon may be that the adjoining spots are very close to each other so that the distance between them is less than the size of minimum morphological operator or the neighboring spots is connected each other. We solve this problem by local histogram analysis for each spot. Based on the initial segmentation result (see example in Fig.2 (a)) obtained by mean shift algorithm and initial template obtained above (see example in Fig.2 (b)), we give our refining scheme as follows:

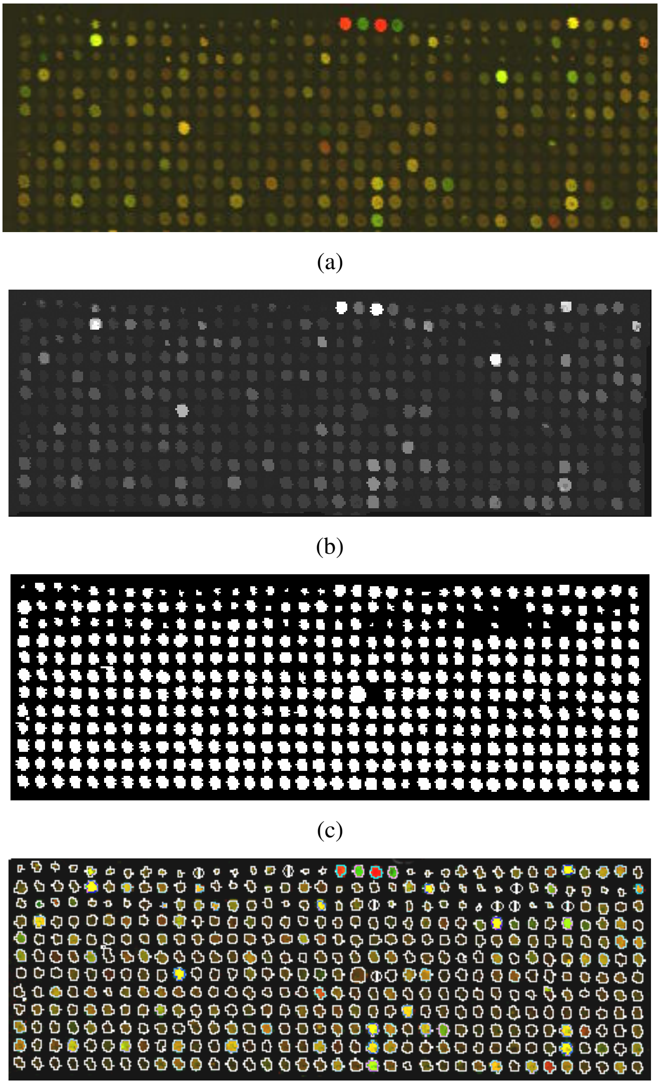
1. Locate the first spot in the initial segmented image, which consists of some homogeneous regions, by the gridding results for original image;
2. Calculate the maximum value of the pixels in the gridded spot region and its histogram; Especially, for judging if a spot is present, we calculate the minimum value of the pixels in a larger area that encloses previous gridded spot region for containing all the spots with different size; note that there are only a few position in the histogram not to equal to 0 since the gridded spot image consist of a few homogeneous regions;
3. Select a small disk region located in the center of gridded spot image and calculate its minimum pixel value. If this minimum value equals to the minimum value calculated in step 2, we can ascertain that this region has no spot available, then locate next spot and go to step 2;
4. Calculate the optimal threshold by histogram analysis (refer to Fig.2(c)): if the abscissa value of the maximum value in histogram is equals to the minimum value calculated in step 2, we can ascertain it correspond the background region and then select the abscissa value of maximum value among the rest of discrete values in the histogram as the optimal threshold; otherwise the optimal threshold, that correspond to the abscissa value of maximum peaks, can be selected among all no-zero points in the histogram; it is obvious that the pixels with the optimal threshold correspond to a largest foreground region (spot);
5. If the optimal threshold is equal to the maximum value calculated in step 2 and the spot has already been well detected and available, then we locate next spot and go to step 2; otherwise we find the maximum connected region by using this optimal threshold as the final spot segmentation result.



(a) (b) (c) (d)

Fig. 2. Illustration of image segmentation and template extraction (a) The part of segmentation image by mean shift algorithm; (b) The initial template extracted by morphological operation for Fig.2(a); (c) The histogram of the center spot image in (a); (d) The final template refined by histogram analysis.

Fig.2 is an illustration of the procedure of spot segmentation template refining using our scheme. Fig.2 (a) is the part of segmentation image by mean shift algorithm; Fig. (b) is the initial template extracted by morphological operation for Fig.2(a); Fig.2 (c) is the histogram of the center spot image in Fig.2(a); Fig.2(d) is the final template refined by histogram analysis. From the Fig. 2(d) and Compared to Fig.2(a), it showed that our scheme is effective and accurate.



(a)

(b)

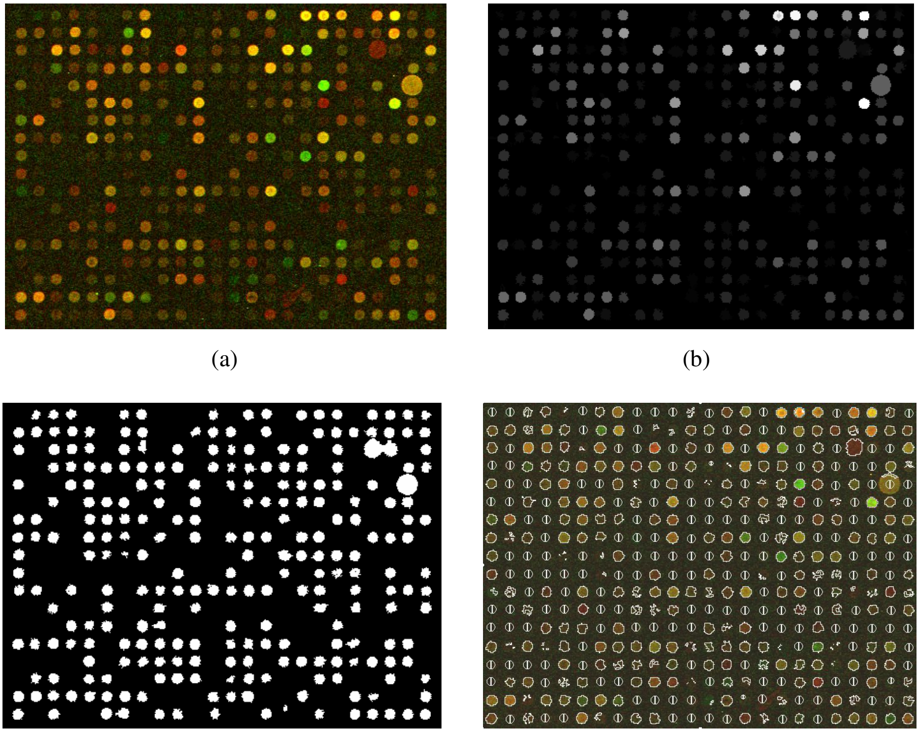
(c)

(d)

Fig. 3. Comparison of segmentation results: (a) Original microarray image. (b) Image segmentation results by the mean shift based technique. (c) Spot segmentation template obtained using our algorithm. (d) Spot segmentation obtained using Genepix Pro 6.0.

4 Experiment Results

We tested our segmentation scheme for many real microarray images and the results show that our algorithm is robust and efficient. We give comparisons of segmentation results obtained by our algorithm and popularly used commercial software Genepix Pro.6.0 in Fig. 3 and Fig.4 that include the initial spot block and the results of segmentation. Note that GenePix uses a special marking (|) for indicating the absent spots. As you can see in Fig.3, the results obtained by using GenePix and our algorithm are similar enough and can accurately extract the spots with arbitrary shape, The comparison in Fig.4 is very interesting. Fig.4 (a) is a low density microarray image with some noise. The results obtained by our algorithm are very good, while GenePix introduces more absent spots. This showed that our method is also robust for noise microarray images. Especially, for the very big Spot, for example, the spot in row 5 and column 23 that GenePix can't detect it, while our methods can detected and segment it perfectly. This may be caused by local based segmentation scheme.



(a) (b)

(c) (d)

Fig. 4. Comparison of segmentation results: (a) Original microarray image. (b) Image segmentation results by the mean shift based technique. (c) Spot segmentation template obtained using our algorithm. (d) Spot segmentation obtained using Genepix Pro. 6.0.

5 Conclusions

Image segmentation is supposed to be the most important step in the processing of microarray image analysis and should be helpful to the following extraction of gene expression information. In this paper, we proposed a new spot segmentation scheme based on mean shift filtering and morphological H-reconstruction and histogram analysis. The difference between our and traditional methods is that our method is based on the technique of global segmentation and template extracting but the traditional is based on local segmentation technique. The advantage of our template-based segmentation scheme, compared to the traditional technique, is that it can definitely indicate where the background is and where the spot is and therefore can greatly facilitate the following analysis of background and foreground (spot) and further can be very helpful to improve the accuracy of information extraction. Experimental results show our scheme is efficient and accurate.

References

1. Schena, M., Heller, R. A., Theriault, T. P., Konrad, K., Lachenmeier, E. and Davis, R. W.: Biotechnology's Discovery Platform For Functional Genomics. Trends in Biotechnology, 16 (1998) 301-306
2. Eisen, M. B.: ScanAlyze. http://rana.stanford.edu/software for software and documentation (2002)
3. Axon Instruments, Inc. http://www.axon.com/GN_GenePixSoftware.html for GenePix Pro 6.0 User's Guide and software (2004)
4. Buckly, M. J.: The Spot user's guide. CSIRO Mathematical and Information Sciences. http: // www. cmis. csiro. au/IAP/Spot/spotmanual.htm (2000)
5. GSI Lumonics: QuantArray Analysis Software, Operator's Manual (1999)
6. Angulo, J., and Serra, J.: Automatic Analysis of DNA Microarray Images Using Mathmatcal Morphology. Bioinformatics, 19 (2003) 553-562
7. Wu, S., and Yan, H.: DNA Microarray Image Processing Based On Minimum Error Segmentation and Histogram Analysis. Proceedings of SPIE, Color Imaging X: Processing, Hardcopy, and Applications. San Jose, CA, USA, 5667 (2005) 562-568
8. Fukunaga, K. and Hostetler, L. D.: The Estimation of The Gradient of A Density Function, with Application In Pattern Recognition. IEEE Trans. Information Theory, 21 (1975) 32-40
9. Cheng, Y.: Mean Shift, Mode Seeking, and Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17 (1995) 790-799
10. Scott, D. W.: Multivariate Intensity Estination. Wiley (1992)
11. Comaniciu, D. and Meer, P.: Mean Shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002) 603-619
12. Hirata, J. R., Barrera, J., Hashimoto, R. F., and Dantas, D. O.: Microarray gridding by mathematical morphology. Computer Graphics and Image Processing, 2001 Proceedings of XIV Brazilian Symposium on 2001, (2001) 112 –119
13. Grimaud, M.: A New Measure of Contrast: Dynamics. Proceedings of SPIE, Image Algebra and Morphological Image Processing III San Diego, 1769 (1992) 292-305

The Cluster Distribution of Regulatory Motifs of Transcription in Yeast Introns*

Jun Hu[1,2] and Jing Zhang[1,**]

[1] Department of Statistics, The Center for Applied Statistics,
Yunnan University, 650091, Kunming, China
zhangjing@ynu.edu.cn
Tel.:+86-871-6541419
[2] School of Science and Information Engineering,
Yunnan Agricultural University, 650201, Kunming, China

Abstract. A comparative analysis of olignucleotide frequencies in two sets of introns of genes highly-transcribed and lowly-transcribed respectively has suggested that the existence of potential positive regulatory motifs of transcription in yeast introns. To further reveal the distribution feature of these motifs, we detected significant clusters of these motifs (mainly pentanucleotides) in introns by r-scan analysis. The results showed that there are more clusters of regulatory motifs in the introns of ribosomal protein genes (highly-transcribed genes) than in lowly-transcribed introns. Experimental studies show that the transcription factors function cooperatively in transcriptional activation and the corresponding binding sites for factors generally cluster in DNA. Accordingly, we speculated that the cluster distribution of regulatory motifs of transcription in yeast introns could be correlated with the cooperative action of transcription factors and the transcriptional rates of genes could be improved by the cooperativity.

1 Introduction

A great number of experimental studies have indicated that many introns of eukaryotic genes could function either as enhancers or as promoters or both in transcriptional regulation [1- 4]; however, the sequence feature of these introns and its regulatory mechanisms have not been thoroughly and systematically understood. In a previous work, we have detected some potential positive regulatory motifs of transcription in yeast introns by comparative analysis of olignucleotide frequencies in two sets of introns with higher and lower transcription rates respectively[5,6]. The occurrence frequencies of these motifs in highly-transcribed introns are significantly higher than those in lowly-transcribed introns, and most of these motifs form many wider overlapping clusters in highly-transcribed introns. So, we speculated that the sequence structures of highly-transcribed introns could be favorable to transcription of genes, and the highly-transcribed introns could also provide enough binding sites

[*] This work was supported by grants from The National Natural Science foundation of China (no.30360027).

[**] Corresponding author.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 51 – 60, 2005.

for transcription factors to perform the cooperative transcription [6]. Cooperativity is a frequent phenomenon in eukaryotic transcriptional control. Many transcription factors show cooperativity in transcriptional activation, which manifests itself in clustering binding sites [7, 8]. These meaningful results lead us to focus our research on two questions: Whether do these regulatory motifs cluster in intron sequence? Are these regulatory motifs and their distribution feature in intron responsible for the high transcription rates?

In this work, we used *r*-scan analysis [9-11] to detect possible clusters of regulatory motifs and then compared the cluster distribution of regulatory motifs in the introns of ribosomal protein genes (highly-transcribed genes) with that in lowly-transcribed introns. The results showed that there are more clusters of regulatory motifs in the introns of ribosomal protein genes than in lowly-transcribed introns. It is recognized that cooperativity is reflected in the occurrence of multiple clumped binding sites in DNA sequence[7,8] and cooperativity contributes to the improvement of transcription rates of genes [12]. Thus, the cluster distribution of regulatory motifs in introns could be correlated with the cooperative action of transcription factors.

2 Samples and Methods

2.1 Samples

In previous work, we intuitively chose 30 and 10 mRNAs/h as cutoff for high and low transcription rates respectively. In fact, almost of all highly-transcribed genes encode ribosomal proteins (RPs) except ybr084w and yfl039c. Thus, from highly-transcribed introns we used the 75 RP genes as sample sequences and analyzed the distribution of regulatory motifs in the introns of RP genes. At the same time, 80 introns of lowly-transcribed genes were chosen. The relevant highly-transcribed and lowly-transcribed genes can be found in Table 1 and Table 2 of reference [5].

Transcription factor ABF1, RAP1 and TAF are involved in the transcriptional regulation of RP genes [13-15]. According to the binding sites for transcription factor ABF1, RAP1 and TAF [16,17], we divided the regulatory motifs (mainly pentanucleotides) into three groups. These three groups pentanucleotides are called ABF1 motif, RAP1 motif and TAF motif respectively (see Table1).

2.2 Methods

Step 1 Definition of *r*-scan Statistic

Firstly, we put the introns of 75 RP genes together into a long sequence of length *L*. This long sequence is called intron Ⅰ hereafter for brevity. The positions of motifs of each group in intron Ⅰ are indicated by $Y_1^* \le Y_2^* \le ... \le Y_n^*$. The distance between two adjacent motifs are defined by $X_s = Y_{s+1}^* - Y_s^*$ $(1 \le s \le n-1)$. Then

Table 1. The potential regulatory motifs for transcription factors ABF1, RAP1 and TAF. The motifs in parentheses denote the reverse complements. Grouping is done according to TRANSFAC[16,17].

ABF1 motif		RAP1 motif		TAF motif	
AAAAT	(ATTTT)	AAATA	(TATTT)	AATAT	(ATATT)
AAATA	(TATTT)	AACTG	(CAGTT)	ATGTC	(GACAT)
AAATT	(AATTT)	AATAT	(ATATT)	ATTAA	(TTAAT)
AATAT	(ATATT)	ACCAC	(GTGGT)	GAATA	(TATTC)
AATGA	(TCATT)	ATGTC	(GACAT)	TGAAA	(TTTCA)
ACTAT	(ATAGT)	ATTCA	(TGAAT)		
AGAAT	(ATTCT)	ATTTA	(TAAAT)		
AGCAT	(ATGCT)	ATTTG	(CAAAT)		
ATAAT	(ATTAT)	CAATC	(GATTG)		
ATGTC	(GACAT)	CACCA	(TGGTG)		
ATTAA	(TTAAT)				
ATTGA	(TCAAT)				
ATTTA	(TAAAT)				
ATTTC	(GAAAT)				
ATTTG	(CAAAT)				
CACCA	(TGGTG)				
CGATA	(TATCG)				
GATAA	(TTATC)				
GGATA	(TATCC)				
TAATA	(TATTA)				
TAGCA	(TGCTA)				

$$R_i = \sum_{j=i}^{i+r-1} X_j \quad , i=1,\ldots,n\text{-}r\,(r \geq 1)\,. \tag{1}$$

is the length of r+1 consecutive motifs. It will be referred to as a r-scan statistic [9-11]. Let $R_1^* \leq R_2^* \leq \ldots\ldots \leq R_{n-r}^*$ is the corresponding order statistics. Therefore, the kth (k=1, 2, 3) minimum of the r-scan [9-11] is $m_k^{(r)} = R_k^*$.

The same process was performed for 80 lowly-transcribed intorns, and the joined long sequence is called intron II. Similarly, r-scan statistic is defined by substituting intron II for intron I.

Step 2 Goodness of Fit Test for Uniform Distribution

The Kolmogorov test for goodness of fit[18]is used to assess whether $Y_1^*, Y_2^*, \ldots\ldots Y_n^*$ are drawn from a uniform on (0, L). Table2 summarizes the results of test for ABF1 motifs, RAP1 motifs and TAF motifs respectively. Because the value of P < 0.05 for RAP1 motifs in intron II, we suggest rejection of the hypothesis of uniform distribution at a 0.05 level of significance. This means that the positions of RAP1 motifs in intron II show a deviation from the uniform distribution. Thus, the cluster distribution of RAP1 motifs in intron Ⅱ will not be taken into further analysis.

Table 2. The Kolmogorov test for uniform distribution in intron Ⅰ and intron Ⅱ. n is the occurrence number of motifs in the same group in intron Ⅰ or intron Ⅱ. P is asymptotic significance level.

Motif	Intron Ⅰ		Intron Ⅱ	
	n	p	*n*	p
ABF1 motif	3337	0.848	719	0.367
RAP1 motif	1346	0.608	280	0.043
TAF motif	874	0.104	178	0.337

Step 3 R-scan Analysis of Distribution of Motifs Cluster [9-11]

If $Y_1^*, Y_2^*, \ldots\ldots Y_n^*$ has a uniform distribution on (0, L) then

$$\lim_{n\to\infty} P\left\{\frac{m_k^{(r)}}{L} < \frac{x}{n^{1+1/r}}\right\} = 1 - \exp\left\{-\frac{x^r}{r!}\right\}\sum_{i=0}^{k-1}\left(\frac{x^r}{r!}\right)^i \frac{1}{i!}. \tag{2}$$

The equation (2) here has been proposed by Dembo A et al [9]. With x chosen so as to the right side of equation (2) is equal 0.01[10,11], the r-scan is referred to a significant cluster if the $m_k^{(r)}$ is less than $L\frac{x}{n^{1+1/r}}$.

3 Results

According to the three steps mentioned above, we have detected significant clusters of motifs for ABF1, RAP1 and TAF in intron Ⅰ and intron Ⅱ respectively.

Table 3. The results of *r*-scan analysis in Intron Ⅰ

Motif	Intron Ⅰ						
	r	$m_1^{(r)}$	Position	$m_2^{(r)}$	Position	$m_3^{(r)}$	Position
ABF1	8			9	18612/18613 /26093	11	18608/18609 /18619/18620
	9	9	18615/18616 18617	10	18612/18613 /26093	12	18608/18609 /18618/18619
	10	10	18615/18616	11	18612/18613	13	18608/18609 /18617/18618
RAP1	4			5	10419/25761	6	8652/17070 /21371
	5	5	2461/18608/ 25761/25762			11	2462/ 15950
	6			12	2461	15	18608/18609
TAF	5	7	18617	8	18618	13	18620/18621
	6	9	18617	14	18620	15	18618
	7	16	18617				
	8	17	18617				

Table 4. The results of *r*-scan analysis in Intron Ⅱ

Motif	Intron Ⅱ						
	r	$m_1^{(r)}$	Position	$m_2^{(r)}$	Position	$m_3^{(r)}$	Position
ABF1	5			6	1642/1643	8	6163
	7					16	2190/9405
	8			18	2190		
TAF	3			7	9584		
	4			19	9584		

3.1 The Cluster Distribution of ABF1 Motifs

In intron Ⅰ, 25 significant overlapping clusters of ABF1 motifs were detected by 8-scans , 9-scans,and 10-scans. Because of the overlapping of motifs, these clusters form two wider DNA sequences. One of them consists of motifs AAAAT ,AAATA , AAATT AATAT , ATAAT , ATATT , ATTAA , TAAAT , TAATA , TATTA , TTAAT, extending for 28 bp in ylr287ca intron (see Table3 and Figure1), and the other consists of motifs AATAT , ATATT , ATTAA , ATTAT , TAATA , TATTA , TTAAT , TTATC , extending for 15 bp in yor182c intron.

In intron II, we found 3 significant overlapping clusters of ABF1 motifs in the case of r=5, one of which is in yjl041w intron and the rest two are in ydl029w intron (see Table4 and Figure2). Furthermore, we detected 3 significant clusters in the case of r=7 and 8. Two of these clusters form a stretch of DNA, extending for 26 bp in ydl189w intron; the remaining cluster, a non-overlapping cluster, occurs in yor318c intron.

3.2 The Cluster Distribution of RAP1 Motifs

In intoneⅡ, with r=4 we found 5 overlapping clusters of RAP1 motifs. The first cluster consists of TAAAT, AAATA, AATAT, ATATT, ATTCA, occurring in ygl189c intron. The second cluster consists of CAAAT, AAATA, AATAT, ATATT, ATTCA, occurring in yor096w intron. The third cluster consists of AAATA, AATAT, ATATT, TATTT, ATTTG, occurring in yer131w intron. The fourth cluster consists of CACCA, ACCAC, occurring in ykr057w intron. The fifth cluster consists of AAATA, TAAAT, AATAT, ATATT, occurring in ymr142c intron. With r=5 we found 2 significant overlapping clusters in introns of ybr189w and ylr287ca respectively. At the same time, we detected 4 non-overlapping clusters in intron of yor096w, ybr189w and yjr145c respectively. With r=6, 1 non-overlapping cluster was detected in ybr189w intron. 2 non-overlapping cluster were observed in ylr287ca intron, which consist of the same motifs TAAAT, AAATA, AATAT, ATATT. With r=7 and 9 we found 3 non-significant clusters, but they span two introns (the introns of ylr185w and ylr287ca) due to some of RAP1 motifs are close to the 5' end of the top strand of the ylr287ca intron.

Because motifs that are not follow a uniform distribution in the intron can not be analyzed with the method introduced above, the clusters of RAP1motifs in lowly-transcribed intorns were not considered.

3.3 The Cluster Distribution of TAF1 Motifs

In intron Ⅰ, 3 significant overlapping clusters are detected by 5-scans and 6-scans. 6 non-overlapping clusters are detected by 5-scans, 6-scans, 7-scans and 8-scans. All clusters consist of the same motifs AATAT, ATATT, ATTAA, TTAAT, extending for 22 bp in ylr287ca intron.

In intron II, we found an overlapping cluster in yor318c intron by 3-scans and a non-overlapping cluster by 4 -scans.

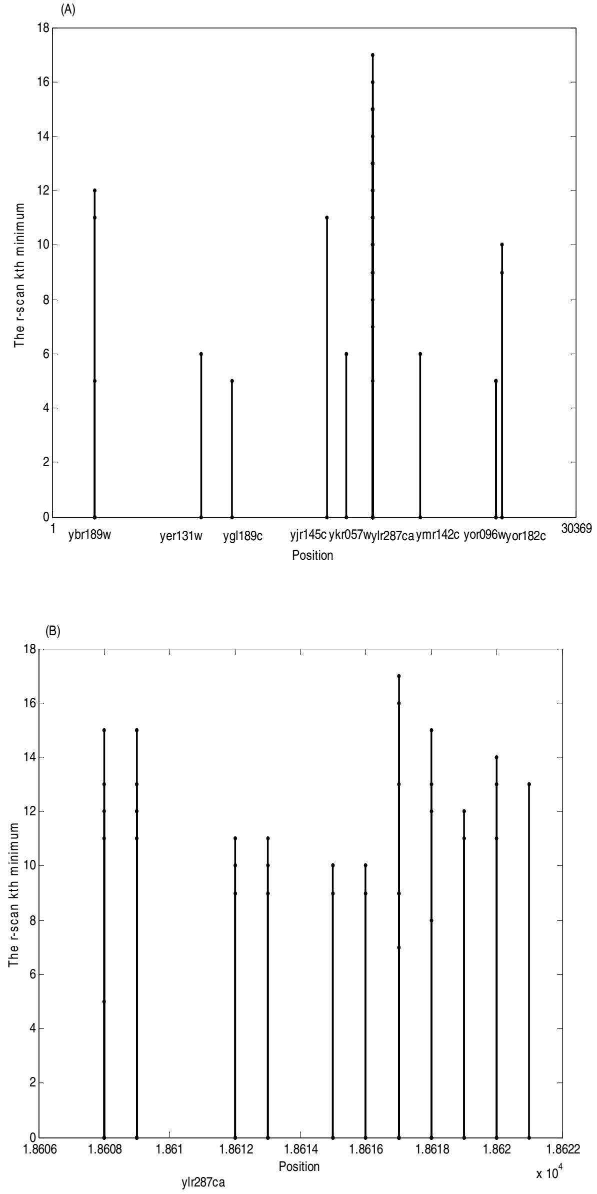


Fig. 1. The lengths of significant clusters and their start positions in intron I. (A) Each point in the plane corresponds to a significant cluster, and abscissa of each point indicates the position in intron I; ordinate of each point indicates the r-scan kth minimum. The bottom of each vertical line indicates the corresponding introns of the ribosomal protein genes in which significant clusters occur. (B) The detail of the significant clusters which locate at between position 18608 and 18635.

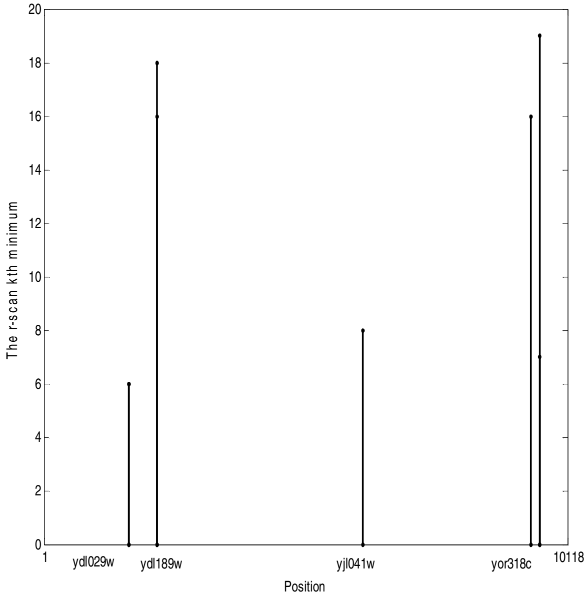


Fig. 2. The lengths of significant clusters and their start positions in intron II. Each point in the plane corresponds to a significant cluster, and abscissa of each point indicates the position in intron II; ordinate of each point indicates the r-scan kth minimum.

4 Discussion

In introns of RP genes and introns of lowly-transcribed genes we have detected significant clusters of motifs for ABF1, RAP1 and TAF by *r*-scan analysis. However, the cluster number in introns of RP genes is larger than that in introns of lowly-transcribed genes. We also noticed that most of clusters contain copies of motifs. For example, a non-overlapping cluster of TAF motifs contains two AATAT. It is known that most transcription factors function cooperatively; this is characterized by the clustering occurrence of binding sites for transcription factors [7,8,19]. Thus clustering binding sites of ABF1, RAP1 and TAF could demonstrate they bind to multiple cooperative sites respectively, and these motif cluster regions may form regulatory modules.

It is noteworthy that of the 75 introns of RP genes, only two introns (introns of yor182c and ylr287ca) have the clusters of ABF1 motifs and one intron (intron of ylr287ca) has the clusters of TAF motifs; in contrast , we found the clusters of RAP1 motifs in the 8 introns. RAP1 is a common activator of RP gene of yeast [14]. Our results further showed that the sequence structures of introns of RP genes in yeast are favorable for RAP1 binding.

The results of goodness-of-fit test showed that the positions of RAP1 motifs are uniformly distributed in intron I, but not in intron II. Almost all of known biological function genes in lowly-transcribed genes do not code ribosomal protein so far. Therefore, the differences in distribution of RAP1 motifs between introns of RP genes and those of lowly-transcribed genes suggest that clustering of RAP1 motifs may be a cause that is responsible for higher transcription rates of RP genes.

References

1. Bhattacharyya, N., Banerjee, D.: Transcriptional Regulatory Sequences within the First Intron of the Chicken Apolipoprotein AI(apo AI) Gene. Gene, 1999, 234 (2): 371~380
2. Brinster, R. L., Allen, J.M., Behringer, R.R., Gelinas, R.E., Palmiter, R.D.: Introns Increase Transcriptional Efficiency in Transgenic Mice. Proc Natl Acad Sci U S A, 1988, 85 (3): 836~840
3. Chen, J., Hayes, P., Roy, K., Sirotnak, F..M.: Two Promoters Regulate Transcription of the Mouse Folypolyglutamate Synthetase Gene: Three Tightly Clustered Sp1 Sites within the First Intron Markedly Enhance Activity of Promoter B. Gene, 2000, 242(1-2): 257~264
4. Surinya, K.H., Cox, T. C., May, B. K.: Identification and Characterization of a Conserved Erythroid-specific Enhancer Located in Intron 8 of the Human 5-Aminolevulinate Synthase 2 Gene. J Biol Chem, 1998, 273(27): 16798~16809
5. Zhang J, Shi X F.: Statistical Analysis of Sequence Features of Introns with Positive Transcriptional Regulation in Yeast Gene. Prog Biochem Biophys, 2003, 30(2) : 231~238
6. Zhang, J., Hu ,J., Shi X F., Cao, H., et al.: Detection of Potential Positive Regulatory Motifs of Transcription in Yeast Introns by Comparative Analysis of Oligonucleotide Frequencies. Comput Biol Chem, 2003, 27(4~5): 497~506
7. Wagner, A.: A Computational Genomics Approach to the Identification of Gene Networks. Nucleic Acids Research, 1997, 25 (18): 3594~3604
8. Wagner, A.: Genes Regulated Cooperatively by One or More Transcription Factors and Their Identification in Whole Eukaryotic Genomes. Bioinformatics, 1999, 15 (10): 776~784
9. Dembo, A., Karlin, S.: Poisson Approximations for R-scan Process. Annals of Applied Probability, 1992, 2(2): 329~357
10. Karlin, S., Brendel, V.: Chance and Statistical Significance in Protein and DNA Sequence Analysis. Science, 1992, 257 (3): 39~49
11. Karlin, S., Macken, C.: Some Statitcal Problems in the Assessment of Inhomogeneities of DNA Sequence Data. Journal of American Statistical Association□1991, 86 (413): 27~35
12. Xue, W., Wang, J., Huang, Q.L., Zheng, W.J., Hua, Z.C.: Synergistic Activation of Eukaryotic Gene Transcription by Multiple Uptream Sites. Prog.biochem.biophys,2002(4), 29,510–513
13. Della, Seta F., Ciafre, S.A., Marck, C., Santoro, B., Presutti, C., Sentenac, A., Bozzoni, I.: The ABF1 Factor Is the Transcriptional Activator of the L2 Ribosomal Protein Genes in Saccharomyces Cerevisiae. Mol Cell Biol, 1990, (5): 2437~2441
14. Lieb J, D., Liu, X., Botstein, D., Brown, P.O.: Promoter-specific Binding of RAP1 Revealed by Genome-wide Maps of Protein-DNA Association. Nat Genet, 2001, 28(4): 327~334
15. Dorsman, J. C., Doorenbosch, M. M., Maurer, C. T., Winde, J.H., Mager, W.H., Planta, R.J., Grivell, L.A.: An ARS/silencer Binding Factor Also Activates Two Ribosomal Protein Genes in Yeast. Nucleic Acids Res, , 1989, 17 (13): 4917–4923
16. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., Urbach, S.: The Transfac System on Gene Expression Regulation. Nucl Acids Res, 2001, 29(1): 281- 283

17. Matys,V.E., Fricke, R., Geffers, E., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., et al.: TRANSFAC®: Transcriptional Regulation, from Patterns to Profiles. Nucl. Acids Res, 2003, 31(1): 374- 378
18. Bickel, P.J., Doksum, K.A.: Mathematical Statistics: Basic Ideas and Selected Topics. Englewood Cliffs, NJ: Prentice Hall.(1977) 378–388
19. Carey, M., Smale, S.T.: Transcriptional Regulation in Eukaryotes Concepts Strategies and Techniques. New York: CSHL Press(2002) 8–42

E-Coli Promoter Recognition Using Neural Networks with Feature Selection

Paul C. Conilione and Dianhui Wang

Department of Computer Science and Computer Engineering,
La Trobe University, Melbourne, VIC 3086, Australia
csdhwang@ieee.org

Abstract. This paper investigates the effects on neural classification performance of biological data by features selection. Where the Relief-F and Symmetrical Tau feature selection algorithms were employed on a set of high level features of DNA and structural profiles. It was observed that even with a small percentage of the features used in neural classifiers, the recognition rate of E.coli promoters was not degraded significantly.

1 Introduction

Data is composed of examples, each example is an observation of a system with a discrete number of features or properties that describe the system. The number of features depends on what is being observed, from a few features for a mechanical system, to several thousand for biological sequences [1]. As the number of features increases, the volume of the feature space grows exponentially, this is called the *Curse of Dimensionality* [2].

Consequently, significant effort has been put into the area of feature selection algorithms (FSA's), which aim to reduce the number of features that are needed to describe the system, but still maintain or even improve the learners performance. In general the goal is to remove irrelevant and redundant features from the data. By reducing the feature space, it can increase learning speed, increase learner performance (e.g. classification accuracy), can make the learners model more easily understood, and reduce the learners storage requirements [3]. Research into the application of FSA's to biological sequences has been growing, for example they have been applied to DNA sequence classification [4], splice site prediction [5] and gene expression profiles [6].

One biological problem is the identification of a promoter region within a DNA sequence. A promoter is a region of DNA, recognised by and a binding target for RNA (ribonucleic acid) polymerase, which then starts transcription of the coding region at the Transcription Start Site (TSS). Using biochemical or genetic means to identify the promoter regions and pinpoint the binding site(s) at which the RNA polymerase comes into contact with the DNA is difficult. For this reason, previous techniques for identification of the promoter regions are based on statistical and alignment techniques. Research by [7], [8] and [9] compiled increasingly larger number of promoter regions of E.coli. Using statistical

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 61–70, 2005.

methods, they identified two major consensus sequences, which consist of two hexamers (6 base pairs (bps)) long. The first consensus sequence is TATAAT and is approximately 35 bps upstream from the TSS, (labelled -35 hexamer). The second consensus sequence is TTGACA and is approximately 10 bps upstream from the TSS (labelled -10 hexamer).

Previous researchers have applied ANN's to the problem of promoter recognition, [10], [11], and [12], achieving promoter recognition in the 90% range and false positive rates of around 5-10%. However, there is not a significant amount of work on the application of FSA's to the problem of promoter recognition in E.coli.

In this paper we analyse promoter DNA data using filter type FSA's, and measure the effects of varying the number of the *best* features selected by an FSA on the classification of E.coli promoters using neural networks.

2 Methods

2.1 Data

We used a pool of 872 E.coli (K12 strain) promoter sequences [13]. The promoter sequences were taken from 61 bases upstream of the TSS, to 20 bases downstream of the TSS. Three different types of non-target DNA sequences were used. The first type was randomly generated DNA sequences with the same base frequency as the target DNA (random-prom). The second type was taken from the gene coding regions of the E.coli K12 strain [13], with 872 sequences selected from the pool of approximately 4400 known genes, starting 100 bps downstream of the TSS. The third type used was randomly generated sequences, but using the same base frequencies of occurrences as the 872 gene DNA sequences, (random-gene). Table 1 summarises the different data-sets used in this paper.

Table 1. Details of the data sets used in this paper

D	Region	N(bps)	Size	Non-prom	Size	Total
A	-60 to +21	81	872	random-prom	872	1774
B	-60 to +21	81	872	gene	872	1774
C	-60 to +21	81	872	random-gene	872	1774

2.2 Feature Extraction

High Level Features. The following are definitions for the *high level* features of a DNA sequence as outlined in [14], and formally defined in [15];

Features 1 to 12 - Helical Parameters. Table 2 lists the 12 different patterns as defined in [16], where R and Y denotes purine (A and G) and pyrimidine (C and T) respectively and each feature takes on the number of times a non-overlapping pattern occurs.

Table 2. Features 1 to 12

Feature	Label	Pattern	Feature	Label	Pattern
1	twist1a	RRRRY	7	roll5a	RRYYY
2	twist1b	YRYYY	8	roll5b	YRRRY
3	twist3a	RRRYR	9	twist7a	RYRRR
4	twist3b	RYYYY	10	twist7b	YYRYR
5	roll4a	RRRYY	11	twist8a	YRYRR
6	roll4b	RYYYR	12	twist8b	YYYRY

Features 13 and 14 - Site Specific Information. Feature 13 is the number of times the gtg_motif occurs in a DNA sequence S, where it does not overlap. Feature 14 is the number of times the gtg_pair motif occurs, where the *spacer* is a multiple of 10 ± 1 bases from the beginning of each motif.

Table 3. Features 13 to 14

Feature	Label	Pattern
13	gtg_motif	GTG or CAC
14	gtg_pair	gtg_motif *spacer* gtg_motif

Features 15 to 16 - Local Secondary Structure. The *local secondary structures* are characterised by the presences of *tandem* and *invert* repeats. Let S be a sequence of N bases, drawn from an alphabet of $\{A, T, C, G\}$. $S = s_1, s_2, ..., s_m$, where s_i is a base at position i in S. The reverse of S is denoted S^{-1}. The complement of a base is the nucleotide that binds to it on the opposite strand of the DNA sequence and is denoted as $\overline{s_i}$, e.g. if $s_i = A$, then $\overline{s_i} = T$. The complement of a sequence is denoted $\overline{S}$.

Feature 15 - Tandem repeats. A tandem repeat is a sequences of nucleotides that occurs twice on the same DNA strand. We define an imperfect tandem repeat with no gaps between repeating sequences as $T = UV$, where the subsequences U and V are expressed as $U = u_1, u_2, ..., u_m$ and $V = v_1, v_2, ..., v_m$. The period p of T is the minimum integer such that $u_i = v_{i+p}$ for some i [17]. The mismatch between subsequences U and V is given by the hamming distance, $d(U, V) = c$, where c is the number of mismatches. The no-gap condition is met iff $u_1 = v_1$ and $u_m = v_m$.

Feature 16 - Inverted repeats. An inverted repeat is a sequence of nucleotides that is found to be repeated in the reverse order on the opposite strands of the DNA double helix. We define an imperfect inverted repeat as $I = UV$, where $U = u_1, u_2, ..., u_m$, $V = v_1, v_2, ..., v_m$, and the number of mismatches is given by the hamming distance $d(U, \overline{V}^{-1}) = c$.

Given a sequence S, all repeats of the same type are found and the size of the repeat n and number of matches $b = n - c$ are recorded. Then the probability of one or more of the repeats being found is calculated using the process detailed in [15], and the smallest probability is used for the feature value.

Features 17 to 19 - DNA compositions. The AT content, AG/TC ratio and AC/TG ratio are given in (1), (2) and (3) respectively.

$$AT\_content = \frac{A+T}{N} \tag{1}$$

$$AG\_TC\_ratio = \begin{cases} \frac{A+G}{T+C} & T+C \neq 0 \\ 0 & T+C = 0 \end{cases} \tag{2}$$

$$AC\_TG\_ratio = \begin{cases} \frac{A+C}{T+G} & T+G \neq 0 \\ 0 & T+G = 0 \end{cases} \tag{3}$$

where A, C, G and T are the number of adenines, cytosines, guanines and thymines respectively, and N is the total number of nucleotides in the sequence window.

DNA Structural Profiles. For a sequence $S = s_1, \ldots, s_N$, with N bases, its profile $P(S)$ is given by $\{p(s_i, .., s_{i+k})\}$ where $p(.)$ is the DNA property value for a given set of bases, $1 \leq i \leq N - k + 1$ and k is the number of nucleotides used to calculate its value. So for properties based on dinucleotides, $k = 2$ and for trinucleotide properties, $k = 3$. We used the DNA structural profiles GC trinucleotide frequency count and Stacking energy [18], please see [19] for further information.

2.3 Feature Selection Algorithms

There is a vast number of FSA's that have been developed, see [20] for a review. There are several types of FSA's, the filter model processes the data before it is feed to the learning algorithm. The wrapper model generates subsets of features and uses a specific classifier to measure whether the subset provides better performance compared to the full feature set. In this paper we have selected filter type FSAs, namely Relief-F (RF) as it is commonly used in the literature, and Symmetrical Tau (ST).

Relief-F. The original Relief algorithm was proposed by [21] and is given in Alg. 1., where $d(.)$ is normalised to the interval $[0, 1]$, which ensures the weights are in the interval $[-1, 1]$. The q highest ranking features according to W were used as the optimal feature set.

Symmetrical Tau. The aim of statistical methods, such as chi-square test, is to determine if a variable B is correlated with variable A. To begin, a contingency table, Tab. 4, is used to relate the two variables, where variable A has α categories, B has β categories, and A_i and B_j.

The problem with the most commonly used statistical, and information theory based feature selection methods, such as Chi-square criterion, Asymmetrical Tau, Information Gain and Gini indexing criterion, is that they tend to favour

Algorithm 1. The Relief-F Algorithm

```
procedure RELIEF-F($D$, $p$)
    $\boldsymbol{W} \leftarrow 0$
    for $i \leftarrow 1$ to $p|D|$ do
        $S \leftarrow$ random sample from $D$
        $H \leftarrow$ near Hit
        $M \leftarrow$ near Miss
        for $j \leftarrow 1$ to $|X|$ do
            $\boldsymbol{W}_j = \boldsymbol{W}_j + d(S_i, M_i) - d(S_i, H_i)$        ▷ $d(.)$ distance function
        end for
    end for
    Return $\boldsymbol{W}$
end procedure
```

Table 4. Contingency table

A	B				
	B_1	B_2	$\ldots$	B_β	Total
A_1	c_{11}	c_{12}		$c_{1\beta}$	c_{1+}
A_2	c_{21}	c_{22}		$c_{2\beta}$	c_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_α	$c_{\alpha 1}$	$c_{\alpha 2}$		$c_{\alpha\beta}$	$c_{\alpha +}$
Total	c_{+1}	c_{+2}		$c_{+\beta}$	c

features with more values. To overcome this problem, [22] proposed the Symmetrical Tau. In the case of a multinomial sampling model, the maximum likelihood estimator of τ is given in (4).

$$T = \frac{c\left[\sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(c_{ij})^2}{c_{+j}} + \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(c_{ij})^2}{c_{i+}}\right] - \sum_{i=1}^{I}(c_{i+})^2 - \sum_{j=1}^{J}(c_{+j})^2}{2c^2 - \sum_{i=1}^{I}(c_{i+})^2 - \sum_{j=1}^{J}(c_{+j})^2} \tag{4}$$

When there is perfect association between variables A and B, $T = 1$. Whilst if $T = 0$ there is no association.

To use ST to determine the worth of each feature, the class of the DNA sequence was represented by variable B and variable A was used to represent the feature under examination. Variable A will have the number of discrete values the feature can take. Using only the training data, the ST was calculated for each feature, and the q highest ranking features according to their T were selected as the optimal subset, Alg. 2..

2.4 Neural Network

The ANN architecture used for classification of the promoters was a fully connected three layer feed forward network with a single neuron to represent the two

Algorithm 2. The Symmetrical Tau Algorithm

```
procedure SymmetricalTau($D$)
    for $i \leftarrow 1$ to $|X|$ do
        create $\boldsymbol{C}_i$ from $D(X_i)$          ▷ Where $C_i$ is the Contingency table
        $\boldsymbol{W}_i \leftarrow T(C_i)$             ▷ Where $T(.)$ is from (4)
    end for
    Return $\boldsymbol{W}$
end procedure
```

classes. The activation function of all neurons was the logarithmic sigmoid function. All weights and biases were randomly initialised in the range of $[-0.01, 0.01]$ The training algorithm used depended on the size of the network being trained.

For the networks trained on DNA encoded using the high level features, the network size is comparatively small and so were trained using the Levenberg-Marquardt (LM) algorithm. Whilst the DNA structural profiles have 79 to 159 features, so the ANNs were trained using the Scaled Conjugate Gradient (SCG) algorithm, which is fast compared to other training algorithms for large networks.

The performance function used was the *mean square error*. When training the ANN, the target value was taken as 1 for a promoter and 0 for a non-promoter. All data were normalised to the range of $[-1, 1]$.

2.5 Performance Evaluation

The performance of the ANN classifier was measured using a confusion matrix and derived F-measure (5), and accuracy (6) metrics. A promoter that is correctly classified is called a *true positive* (TP), whilst a promoter classified as a non-promoter is called a *false positive* (FP). A non-promoter that is correctly classified is called a *true negative* (TN) and an incorrectly classified non-promoter is called a *false negative* (FN). The accuracy of classifying each class is given by the promoter and non-promoter sensitivity, (7) and (8).

$$F = \frac{TP}{TP + FP + FN} \tag{5}$$

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$S_P = \frac{TP}{TP + FP} \tag{7}$$

$$S_{NP} = \frac{TN}{TN + FN} \tag{8}$$

2.6 Training and Testing Process

To train and test the ANN-based promoter classifiers, K-fold cross validation was used. K-fold cross validation segments a data set D into K folds $F_1, .., F_K$ of approximately the same size, and all folds contained an equal number of instances from each class. For each step i, we trained an ANN on folds $F_j, 1 \leq j \leq K, j \neq i$, and tested using F_i.

To find the best generalisation of a data set D with q features at step i. The FSA is applied to the training data and the resulting q best features were used for both the training data and test data. The network was trained over a number of epochs and at regular periods, training was paused and the classification performance of the training data and test data were recorded in separate confusion matrices. The F-measure of the test data was calculated and if it was the best so far, then the confusion matrices of the train and test data were recorded for step i. If the networks test data set F-measure did not improve after a set number of epochs, training was halted. Once cross validation was complete, the training and testing confusion matrices were summed to get the overall training and test results. If the network output was above 0.5, the instance was classified as a promoter, whilst if the output was below 0.5, then the instances was classified as a non-promoter.

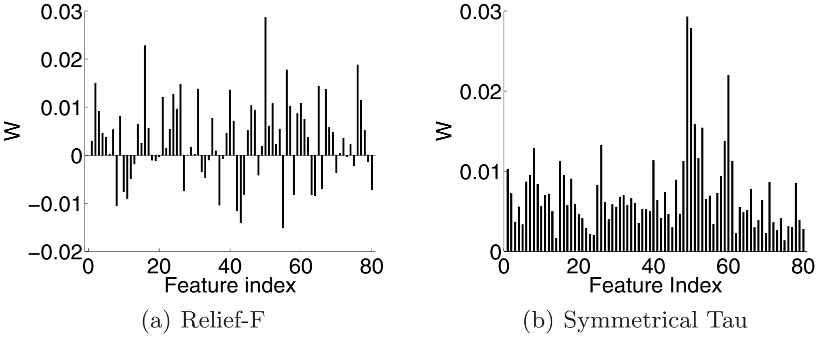
Five-fold cross validation was used to explored the effect of varying the number of features selected by the FSA's, where the number of neurons in the hidden layer was set as $\frac{1}{3}$ of the number of features.

The three data sets were converted to the high level features, stacking energy and GC-trinucleotide profiles. In addition, the stacking energy and GC-trinucleotide profiles of each data set were combined, to examine if classification accuracy would improve.

3 Results and Discussion

Each FSA was applied to each of the data-sets, for all extracted feature types. Given space limitations, only a few results are presented here. The analysis by RF and ST of the data-set A using the stacking energy profile is shown in Fig. 1(a) and Fig. 1(b).

Figure 2 is an example of how the training and test classification accuracy changes as the number of q best features are selected by the FSA's and Tab. 5 summarises the best classification results of the ANN's.



(a) Relief-F (b) Symmetrical Tau

Fig. 1. Data-set A encoded using the Stacking Energy profile

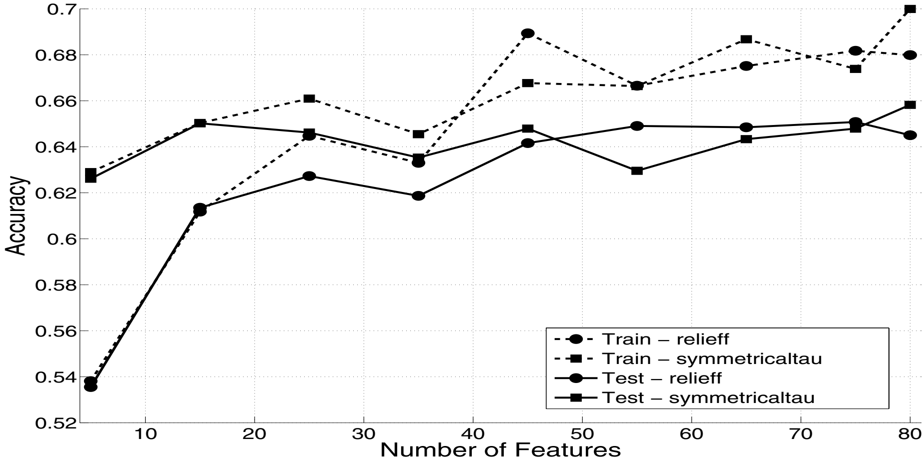


Fig. 2. Training and test data classification accuracy for data set A using the Stacking Energy profile

Table 5. Best results of the application of Relief-F (RF) and Symmetrical Tau (ST) FSA's to E.coli promoter classification using neural networks

D	**Perf**	**Feat**	**None**	$\|X\|$	**RF**	$\|X^*\|$	**ST**	$\|X^*\|$	**Feat**	**None**	$\|X\|$	**RF**	$\|X^*\|$	**ST**	$\|X^*\|$
A	acc	High Level	0.6339	19	0.6502	17	0.6491	15	Stacking	0.6517	80	0.6508	75	0.6583	80
	S_p		0.6491		0.6697		0.6502			0.6256		0.6261		0.6353	
	S_{np}		0.6187		0.6307		0.6479			0.6778		0.6755		0.6812	
B	acc		0.7942	19	0.8022	14	0.8010	12		0.8154	80	0.8194	75	0.8194	75
	S_p		0.7844		0.8039		0.7982			0.8251		0.8360		0.8303	
	S_{np}		0.8039		0.8005		0.8039			0.8056		0.8028		0.8085	
C	acc		0.8217	19	0.8308	16	0.8320	17		0.7996	80	0.8016	80	0.7976	80
	S_p		0.8119		0.8108		0.8108			0.7947		0.8108		0.7787	
	S_{np}		0.8314		0.8509		0.8532			0.8045		0.7924		0.8165	
A	acc	GC-Trinucl	0.6540	79	0.6600	55	0.6514	35	Stacking + GC	0.6637	159	0.6818	130	0.6829	150
	S_p		0.6497		0.6560		0.6594			0.6875		0.6617		0.6697	
	S_{np}		0.6583		0.6640		0.6433			0.6399		0.7018		0.6961	
B	acc		0.8102	79	0.8108	65	0.8119	79		0.8177	159	0.8268	150	0.8245	150
	S_p		0.8131		0.8177		0.8062			0.8268		0.8383		0.8406	
	S_{np}		0.8073		0.8039		0.8177			0.8085		0.8154		0.8085	
C	acc		0.8208	79	0.8349	65	0.8320	65		0.8308	159	0.8337	150	0.8343	150
	S_p		0.8360		0.8452		0.8062			0.8297		0.8314		0.8028	
	S_{np}		0.8056		0.8245		0.8578			0.8320		0.8360		0.8658	

For the high level encoding, both RF and ST showed that the the most important features were features 15, 16 and 17, which are tandem repeats, inverted repeats and AT content respectively. This indicates that the repeat structures are more likely to indicate the presence of a promoter than in a random or gene DNA sequence.

Whilst for all of the structural profiles, both FSA's were generally able to identify the -10 regions, and to a lesser extent the -35 regions, as being more strongly correlated with the class, than other regions of the DNA sequence, as illustrated in Fig's 1(a) and 1(b).

For the high level encoding, classification accuracy did not drop substantially until there were only 2 to 3 features used, namely feature 15, 16 and 17, showing their importance to classification.

After examining the classification results for all structure profile types, in general the classification accuracy did not suffer greatly as the number of features was reduced quite substantially. Even though the classification accuracy did not improve as the number of features were reduced, the classification accuracy was maintained, hence allowing for faster training and smaller network sizes.

From Tab. 1, it can be seen that the best classification accuracy was achieved with most or all of the features from the data-set. This indicates that even though most of the features are not strongly related to the class, they do provide some additional information for the ANN to make a classification.

By examining the results from both RF and ST the differences between the two algorithms are negligible. Indicate both are useful for the feature selection of large biological sequences. However, RF was approximately ten times slower in calculating the features weight compared to ST.

Ultimately due to the amount of data and time constraints, we were unable to determine a mean and standard deviation of our results.

4 Conclusion

The Relief-F and Symmetrical Tau FSA's were able to identify the -10 region of E.coli promoter as being more correlated to the promoter than other regions. The classification accuracy only degraded slightly as the number features used was reduced. We found that there was no significant difference between Relief-F and Symmetrical Tau in terms of determining each feature correlation with the instance class, or resulting classification accuracy by the ANN.

References

1. Xing, E.P., Jordan, M.I., Karp, R.M.: Feature Selection for High-Dimensional Genomic Microarray Data. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001) 601–608
2. Bellman, R.E.: Adaptive Control Processes: A Guided Tour. Princeton University Press (1961)
3. Molina, L.C., Belanche, L., Nebot, A.: Feature Selection Algorithms: A Survey and Experimental Evaluation. In: In Proc. of the International Conference on Data Mining (ICDM'02), Maebashi City, Japan, IEEE Computer Society (2002) 306–313 ISBN 0-7695-1754-4
4. Chuzhanova, N., Jones, A., Margetts, S.: Feature Selection for Genetic Sequence Classification. Bioinformatics **14** (1998) 139–143
5. Saeys, Y., Degroeve, S., Aeyels, D., Van de Peer, Y., Rouze, P.: Fast Feature Selection Using a Simple Estimation of Distribution Algorithm: A Case Study on Splice Site Prediction. Bioinformatics **19** (2003) 179–188

6. Park, C., Cho, S.B.: Genetic Search for Optimal Ensemble of Feature-Classifier Pairs in DNA Gene Expression Profiles. Proceedings of the International Joint Conference on Neural Networks, Volume 3 (2003) 1702–1707
7. Hawley, D.K., McClure, W.R.: Compilation and Analysis of Escherichia Coli Promoter DNA Sequences. Nucl. Acids. Res. **11** (1983) 2237–2255
8. Harley, C.B., Reynolds, R.P.: Analysis of E. coli Promoter Sequences. Nucl. Acids. Res. **15** (1987) 2334–2361
9. Lisser, S., Margalit, H.: Compilation of E.coli mRNA Promoter Sequences. Nucl. Acids. Res. **21** (1993) 1507–1516
10. Mahadevan, I., Ghosh, I.: Analysis of E.coli Promoter Structures using Neural Networks. Nucl. Acids. Res. **22** (1994) 2158–2165
11. Ma, Q., Wang, J.T.L., Shasha, D., Wu, C.H.: DNA Sequence Classification via an Expectation Maximization Algorithm and Neural Networks: A Case Study. IEEE Transactions on Systems, Man and Cybernetics, part c **31** (2001) 468–475
12. Ma, Q., Wang, J.T.L., Gattiker, J.R.: 30. In: Mining Biomolecular Data Using Background Knowledge and Artificial Neural Networks in Handbook of Massive Data Sets. Kluwer Academic Publishers (2002) 1141–1168
13. Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millán-Zárate, D., Díaz-Peredo, E., Sánchez-Solano, F., Pérez-Rueda, E., Bonavides-Martínez, C., Collado-Vides, J.: RegulonDB (version 3.2): Transcriptional Regulation and Operon Organization in Escherichia coli K-12. Nucl. Acids. Res. **29** (2001) 72–74
14. Hirsh, H., Noordewier, M.: Using Background Knowledge to Improve Inductive Learning of DNA Sequences. In Proceedings of the Tenth Conference on Artificial Intelligence for Applications (1994) 351–357
15. Conilione, P.C., Wang, D.: Effect of Non-Target Examples on E.coli Promoters Recognition Using Neural Networks. In Proceedings of International Joint Conference on Neural Networks IJCNN 2005, IEEE (2005) To be published.
16. Lennon, G.G., Nussinov, R.: Homonyms, Synonyms and Mutations of The Sequence/Structure Vocabulary. J Mol Biol. **175** (1984) 425–430
17. Kolpakov, R.M., Kucherov, G.: Finding Approximate Repetitions under Hamming Distance. In auf der Heide, F.M., ed.: ESA. Volume 2161 of Lecture Notes in Computer Science., Springer (2001) 170–181
18. Ornstein, R.L., Rein, R., Breen, D.L., Macelroy, R.D.: An Optimized Potential Function for The Calculation of Nucleic Acid Interaction Energies I. Base stacking. Biopolymers **17** (1978) 2341–2360
19. Conilione, P.C., Wang, D.: Neural Classification of E.coli Promoters Using Selected DNA Profiles. In: The Fourth IEEE Internation Workshop on Soft Computing and Transdisciplinary Science and Technology, Springer (2005) To be published.
20. Liu, H., Yu, L.: Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Trans. on Knowledge and Data Engineering (2004)
21. Kira, K., Rendell, L.A.: The Feature Selection Problem: Traditional Methods and a New Algorithm. In: AAAI. (1992) 129–134
22. Zhou, X.J., Dillion, T.S.: A Heuristic - Statistical Feature Selection Criterion For Inductive Machine Learning In The Real World. In: Systems, Man, and Cybernetics, 1988. Proceedings of the 1988 IEEE International Conference on. Volume 1. (1988) 548–552

Improve Capability of DNA Automaton: DNA Automaton with Three Internal States and Tape Head Move in Two Directions

Xiaolong Shi[1], Xin Li[2], Zheng Zhang[1], and Jin Xu[1]

[1] Department of Control Science and Engineering,
Huazhong University of Science and Technology,
430074 Wuhan, P.R.China
shixiaolong@mail.hust.edu.cn
[2] The First Affiliate Hospital of Wuhan University,
Wuhan University,
430060 Wuhan, P.R.China
Xinxin_star@hotmail.com

Abstract. DNA automaton is a simple molecular-scale automaton, in which the converting of information deploys in molecule-scale by DNA and DNA-manipulating enzymes autonomously. Finite automaton with two internal states has been applied to medical diagnosis. This paper analyses the computation ability of DNA automaton with different enzymes and the possibility of DNA finite automaton with three internal states which is more powerful than the two internal states finite automaton. Finally, we describe a DNA finite automaton with three internal states and proposal a scheme of DNA automaton model in which the tape head can move forward and backward, and symbols can be read from and write into the tape, thus extend the computation ability of DNA automaton and its application fields.

1 Introduction

Autonomous can convert information of one form into another according to a definite procedure, such as Turing and finite state automaton. DNA automaton is a simple molecular-scale automaton that convert information through biological process. Typically a DNA automaton is consists with three molecular units: hardware unit, a double-stranded DNA contains the alphabet taps to be read; software unit, double-stranded DNA encodes input and output information with programming amounts; regulation unit, which regulate software molecule concentrations, and hence make automaton transit according to a definite procedure. The key point of a DNA automaton is Restriction Enzyme that makes automaton transition possible, while the cutting site and recognizing site of Restriction Enzyme determined the computation ability of a DNA automaton.

A molecular automaton has been proposed by Yaakov benenson [1] to realize the basic features and processes of a finite automaton with two internal states using enzyme *FokI* with DNA computing. This automaton has an alphabet comprising two input symbols therefore can have eight possible transition rules, 255 possible transi-

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 71 – 79, 2005.

tion-rule selections and 3 possible selections of accepting states, resulting in 765 syntactically distinct programs.

Since DNA automatons can read and "logically" judge information contained in DNA according to some definite procedures automatically, and confidently give output DNA as compute result, it can surly be applied in gene diagnose and gene analysis. In 2004, Yaakov benenson [2] applied the finite automaton with two internal states described above in medical diagnosis, which can identify and analyze mRNA of disease-related genes associated with simplified models of small-cell lung cancer and prostate cancer.

The motivation of this paper is to study the computation ability of DNA automaton with different enzymes and improve the power of DNA automaton. The DNA finite automaton with two internal states can resolve some gene diagnose problems, but its ability is restricted by two internal states. The disease models for two states DNA finite automaton are highly simplified, and the logical relation of each gene expression of diagnostic rule must be "and", but generally the expression of genes is regulated by each other. Considering the condition that expressions of two genes have relationship with each other, diagnostic rules can not be expressed by the DNA finite automaton with two internal states.

2 Restriction Enzyme: Trigger of DNA Automaton

Restriction Enzyme is a critical component in DNA automaton, which makes possible status switching of DNA automaton. It is an endonuclease which recognizes a specific sequence of bases in a DNA molecule. Type I restriction enzymes bind to the recognition site and then cut randomly somewhere along the length of the molecule. Type II restriction enzymes bind at a recognition site and then cleave the molecule by clipping the DNA backbones somewhere within this sequence of bases. It is type II restriction enzymes that have been used extensively in DNA computing, and the restriction enzyme we discusses here are type II.

The scheme of DNA automaton aims at the representation of Turing computations through DNA computing. Therefore, various models of DNA computing of the Turing-universality have been proposed [3]. The pioneer conceiving of DNA automaton is the DNA Turing model proposed by C. Bennett [4] in 1973, but the feasible model of DNA automaton is constructed by Rothemund [5] using restriction enzyme in 1995. In this DNA computing model, the Turing tape is encoded by DNA sequence, the restriction enzyme recognizes the specific sequence of DNA sequence which acts as the tape head of automaton, cuts correspondence sites to make the tape head move forward and therefore the transition of automaton status achieved.

This model is quite perfect in theory, and makes a great contribution to the research of DNA automatons. In fact, the succeed models of DNA automatons are based on this model in which DNA sequence encoded tape symbols and restriction enzyme trigger the automaton. The drawback of this model is that it is too idealized to realize by experiment. Therefore, Smith and Scheweitzer [6] make effort to use DNA and standard laboratory technique to realize the DNA Turing model. Whereafter, Beaver [7] proposed a DNA Turing machine to operate single molecule. In 1995, winfree [8] proposed a cellular automata model of DNA computing, and realize a programmable self-assembly DNA model in 1998 by experiment [9].

In order to explore the feasibility of autonomous molecular computing, kensaku sakamoto [10] proposed a DNA computing model use the hairpin formation by single-strandedDNA molecules to solve a famous NP problem, the so-called satisfiability problem (SAT) in 2000. The problem is elaborately encoded with DNA hairpin structure, and the selected DNA hairpin structure is well-designed to contain a recognition site of given restriction enzyme to cut off them. This model approves the feasibility of DNA automatons, but it is so well-designed to extend.

Till 2001, Yaakov benenson [1] proposed the programmable molecule automaton model based on DNA computing, and the restriction enzyme in this DNA automaton is *FokI*. The main merit of this DNA automaton model is that it is programmable, in another words, it has the universal property and can be extend to other problems. In this model the distribution between states and symbols depends on the length of the spacer between the recognition site and the restriction site of the particular restriction enzyme employed.

Form the previous works above, we can see that the features of restriction enzyme determine the computation ability of DNA automaton. Thus we try to analyze the features of different restriction enzymes and establish the relationship between features of restriction enzyme and encoding scheme of DNA automaton.

Restriction enzymes can be divided into three kinds in the proper sense of DNA computing: the cut site is apart from the only recognition site, the cut site is inside the only recognition site and the cut site is between two recognition sites.

2.1 Restriction Enzyme with One Recognition Site and a Cut Site Inside

The restriction enzyme with one recognition site and a cut site inside is the most widely used restriction enzyme in biochemical engineering. The action of this kind of restriction enzyme can be demonstrated as follows *(Aat II in this example)*:

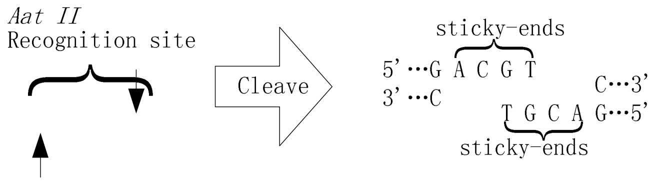


Fig. 1. Restriction enzyme with one recognition site and a cut site inside has no spacer between recognition site and restriction site, thus leave no space for the encoding of states and symbols of DNA automaton. It can only provide a single state transition mode, accordingly with one state and one symbol. For this reason, it has not been applied in DNA automaton model. Whereas, since it can provide a state transition mode, it may be used in DNA automaton with further discussion.

2.2 Restriction Enzyme with One Recognition Site and a Cut Site Beside

Restriction enzymes of this kind have some wonderful features, and have been successfully used in DAN computing. These enzymes act on DNA sequence as follows *(FokI in this example)*:

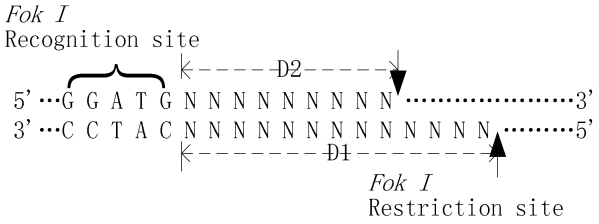

Fok I
Recognition site
D2
5' ···G G A T G N N N N N N N N N ···················3'
3' ···C C T A C N N N N N N N N N N N N N ·········5'
D1
Fok I
Restriction site

Fig. 2. N denotes any base that can be encoded. The distance between recognition site and restriction site of the restriction enzyme provides the encoding space for states and symbols. D1 and D2 determine the state transition mode the restriction enzyme can provide for DNA automaton. The restriction enzyme used in programmable DNA automaton model is of this kind [1].

Given a restriction enzyme with one recognition site and a cut site beside, the distances between recognition site and restriction site are D1 and D2 (Fig. 2.), and D1 > D2. We can make the following conclusion between features and the state transition mode of the restriction enzyme:

The length of encoded states S: The length of encoded states S is determined by the length of sticky-ends cleaved by the restriction enzyme,

$$S = (D1 - D2), D1 > D2 \tag{1}$$

The length of encoded symbols H: The encoded symbol must contain the states, for an n states automaton, each state must separate by at least one BP of nucleotide, on the other hand, the symbol and n states should be about to be cut off by the restriction enzyme. Therefore, the length of encoded symbols H must fit:

$$D2 - n + 1 \geq H \geq S + n - 1 \tag{2}$$

Probable maximum states T: From formula 2, we can conclude that the probable maximum states T a restriction enzyme can provide:

$$T \leq D2 - H + 1 \tag{3}$$

When H reaches the minimal value H = S + T – 1, T reaches its maximum value:

$$T \leq (D2 - S + 2)/2 = D2 - D1/2 + 1 \tag{4}$$

Here, we give out the formula to calculate probable maximum states directly from the features of restriction enzyme with one recognition site and a cut site beside. And from formula 4, we can calculate that the probable maximum states *FokI* can provide is T≤9-13/2+1=3.5, thus a DNA automaton using *FokI* can contain at most 3 internal states.

2.3 Restriction Enzyme with Two Recognition Sites and a Cut Site Inside

This kind of restriction enzyme have two recognition sites, thus gives more restriction to encoding process of DNA automaton, the action of these restriction enzyme can be demonstrate as *(BstXI in this example)*:

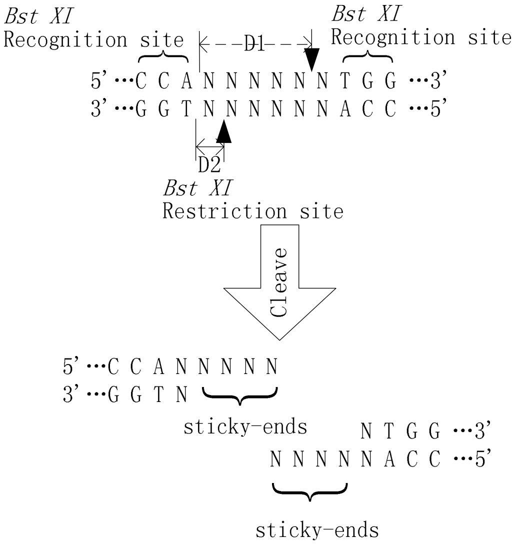


Fig. 3. *Bst XI* has two recognition site, D1 and D2 provide the encoding space for DNA automaton. Since D1 > D2, the separated pieces have single stranded "sticky-ends," which allow the complementary pieces to combine. Therefore, this kind of restriction enzyme can also provide several state transition modes due to D1 + D2 is not zero. In fact, *Bst XI* has been successfully applied in DNA automaton of hairpin model to solve SAT problem [10].

A restriction enzyme with two recognition sites and a cut site between them may provide some new state transition scheme for DNA automaton. With this feature, we can make tape head moves not only forward but also backward, which will be discussed in the follow sections.

3 Scheme of Complex DNA Automaton

With the features of restriction enzymes studied above, we can give some wonderful schemes of more complex DNA automaton than the 2 internal states automaton. Here we will give out the schemes for three states DNA automaton and free tape head direction DNA automaton.

3.1 Scheme for Three States DNA Automaton

The design of our three states DNA automaton incorporates ideas from designs for two states DNA automaton. According to formula 4, the Probable maximum state of *FokI* is three, and we will use *FokI* as the hardware of three states DNA automaton. The state transition diagram of three states DNA automaton is:

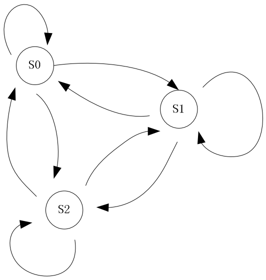


Fig. 4. State transition diagram of finite automatons with three internal states, for each acceptable symbol, there are nine possible state transitions

The nine transition molecules are encoded as Fig.5.

```
Given input symbol N encoding as:         5'GGCTCT3'
The correspondence three states of N is:
                                          N: 5'GGCTCT3'
                                          S0: 5'GGCT3'
                                          S1: 5'GCTC3'
                                          S2: 5'CTCT3'
ALL Transition molecules for symbol N:

Fok I
Recognition site

GGATGTAC              GGATGTACA              GGATGACGAC
CCTACATGCCGA          CCTACATGTGCTC          CCTACTGCTGCTCT

      S0→S0                  S1→S0                  S2→S0

GGATGTA               GGATGTAC               GGATGACGA
CCTACATCCGA           CCTACATGGCTC           CCTACTGCTCTCT

      S0→S1                  S1→S1                  S2→S1

GGATGT                GGATGTA                GGATGACG
CCTACACCGA            CCTACATGCTC            CCTACTGCCTCT

      S0→S2                  S1→S2                  S2→S2
```

Fig. 5. For each acceptable symbol of automaton, give 9 transition molecules as software molecules. Each transition molecule is comprised with *FokI* recognition site, transition regulating part (black alphabet in transition molecule) and complementary of encoded state of input symbol.

The process of the 3 states DNA automaton can be demonstrated as follows:

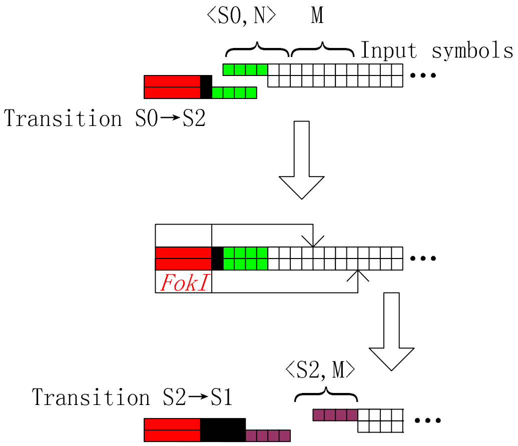


Fig. 6. Scheme and Process of the three states DNA automaton, which starts when the hardware, software and input are all mixed together and runs autonomously, if possible till termination

The automaton processes the input as shown in Fig. 6.The first input symbol is comprised with a four-nucleotide sticky end that encodes the initial state. The computation proceeds via a cascade of transition cycles. In each cycle the sticky end of an applicable transition molecule ligates to the sticky end of the input molecule, detecting the current state and the current symbol. The product is cleaved by *FokI* inside the next symbol encoding, exposing a new four-nucleotide sticky end.

The design of the transition molecules ensures that the 6 BP encodings of the input symbols are cleaved by *FokI* at three different status, the leftmost encoding the state S0 and the middle encoding S1 the rightmost encoding S2 (Fig. 5).

The exact next restriction site and thus the next internal state are determined by the current state and the size of the spacers (Fig. 2a, green) in an applicable transition molecule. The computation proceeds until no transition molecule matches the exposed sticky end of the input or until the special terminator symbol is cleaved, forming an output molecule that has a sticky end encoding the final state. In a step extraneous to the computation and analogous to a `print' instruction of a conventional computer, this sticky end ligates to one of two output detectors and the resultant output reporter is identified by gel electrophoresis.

3.2 Scheme for Free Tape Head Direction DNA Automaton

As the three states DNA automaton described above, the tape head can only move “forward”, in order to make the tape head move “backward”, we must employ some new state transition modes in the scheme of DNA automaton.

Considering features of the restriction enzyme with two recognition sites and a cut site between them in chapter 2, *Bst XI* can form a 4 BP long sticky end the same as

Fok *I* and provide a new state transition mode in the processing of DNA automaton. Therefore, we establish a free tape head DNA automaton model which incorporates *Bst XI* and *Fok I* together.

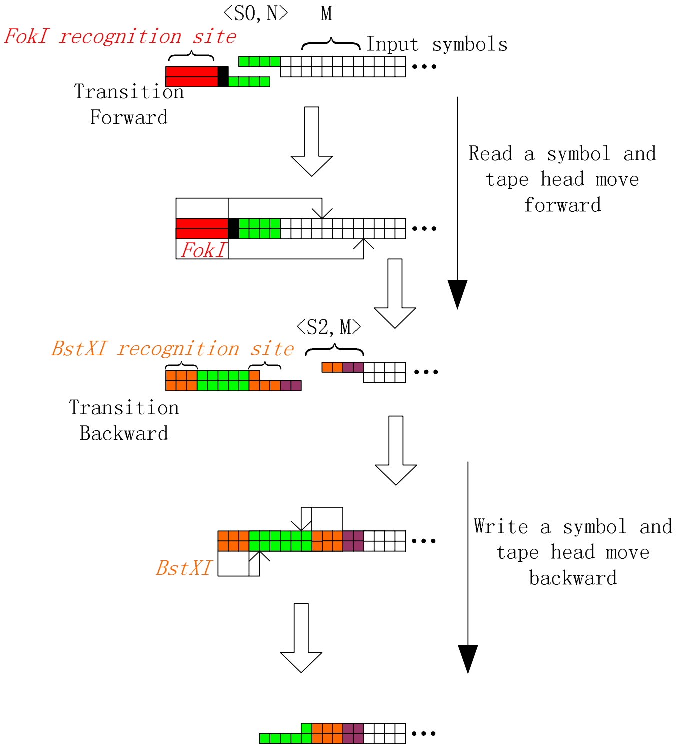


Fig. 7. Scheme and Process of free tape head direction DNA automaton, the length of DNA sequence encoded input symbols increased as *Bst XI* cleaves it, thus makes the tape head move backward and write a new symbol on the tape. Since the sticky end is upside down when tape head change its direction, the state transition molecule must change accordingly.

4 Conclusion

In order to improve the computation capability of DNA automaton, we studied the features of different restriction enzymes, which act as triggers of DNA automaton, and give out the formula to calculate probable maximum states from features of specific restriction enzyme. In succession, we establish several scheme of complex DNA automaton with different restriction enzymes and combination of them. The 3 states finite automaton model of DNA computing we proposed improve the capability of 2 states DNA automaton, while the free tape head DNA automaton model may lead to the realization of universal Turing machine based on DNA computing.

Acknowledgment

This project is supported by CNSF (Grant number: 60274026, 60403002, 60403001), Postdoctoral Foundation of HuaZhong University of Science & Technology.

References

1. Yaakov Benenson, Paz-Elizur T., Adar R., Keinan E., Livneh Z., Shapiro E.: Programmable and autonomous computing machine made of biomolecules. Nature, 22 (2001) 430–434
2. Yaakov Benenson, Binyamin Gil, Uri Ben-Dor1, Adar R., Shapiro E.: An autonomous molecular computer for logical control of gene expression. Nature, 27 (2004) 423~429
3. Yurke B., Andrew J. Turbereld, Allen P. Mills Jr, Friedrich C. Simmel & Jennifer L. Neumann: A DNA-fuelled molecular machine made of DNA. Nature, 10 (2000) 605~608
4. Bennett C. H.: On Constructing a Molecular Computer. IBM Journal of Research and Development, 17 (1973) 525~532
5. Paul Wilhelm, Karl Rothemund: A DNA and restriction enzyme implementation of Turing machine. http://www.ugcs.caltech.edu.~pwkr /oett.html
6. Smith W., Scheweitzer A.: DNA computer in Vitro and Vivo. DIMACS workshop on DNA based computing. Princeton, 1995
7. Beaver D.: Computing with DNA. Journal of computation biology, 3 (1996) 254~ 257
8. Erik Winfree: On the computational power of DNA annealing and ligation, Technical Report, California Institute of Technology, USA, 1995
9. Erik Winfree, et al.: Design and self-assembly of two-dimensional DNA crystals, Nature, 394 (1998) 539-544
10. Kensaku Sakamoto, et al.: Molecular computation by DNA hairpin formation, Science, 19 (2000) 1223-1226

A DNA Based Evolutionary Algorithm for the Minimal Set Cover Problem

Wenbin Liu[1], Xiangou Zhu[1], Guandong Xu[1], Qiang Zhang[2], and Lin Gao[3]

[1] School of Computer Science and Engineering,
Wenzhou Normal College, Wenzhou City 325027, China
wbliu69@sohu.com
[2] University Key Lab of Information Science & Engineering,
Dalian University, Dalian 116622, China
zhangq26@126.com
[3] School of Computer, Xidian University, Xi' an city 710071, China
lgao66@hotmail.com

Abstract. With the birth of DNA computing, Paun et al. proposed an elegant algorithm to this problem based on the sticky model proposed by Roweis. However, the drawback of this algorithm is that the "exponential curse" is hard to overcome, and therefore its application to large instance is limited. In this s paper, we present a DNA based evolutionary algorithm to solve this problem, which takes advantage of both the massive parallelism and the evolution strategy by traditional EAs. The fitness of individuals is defined as the negative value of their length. Both the crossover and mutation can be implemented in a reshuffle process respectively. We also present a short discussion about population size, mutation probability, crossover probability, and genetic operations over multiple points. In the end, we also present some problems needed to be further considered in the future.

1 Introduction

DNA computing is a new vista of computation that bridges between computer science and biochemistry. Because of its potential of massive parallelism, high density of information storage and energy efficiency, DNA computing has become an attractive research field since Adleman's seminal paper in 1994 [1]. Although great progress has been achieved both in theoretical and experimental aspects, its application to realistic problems still suffers from two major limitations. First, the generating and filtering approach employed by Adleman and other researchers couldn't overcome the "exponential curse" and therefore its application to large instance is limited. As estimated in [2], the quantity of DNA molecules needed for a Travel Salesman Problem (TSP) with 200 cities would larger than the weight of earth. This leads to the major criticism on DNA computing except for other factors. Another concern is the reliability of the biochemical protocols employed in DNA computing. Currently, great improvement has been witnessed both in reliability and speed of biological techniques over the past ten years.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 80 – 89, 2005.

At present, there exist two potential directions to tackle the first barrier. One is to convert known heuristic algorithms in traditional computers to DNA algorithms. Ogihara first proposed this idea for the 3-SAT and the results of computer simulation showed that the space complexity of 3-SAT might be reduced from 2^n to $2^{0.5n}$ (where n represents the number of the variables appeared in a formula) [3]. Another direction is to hybrid between DNA computing and evolutionary computing, which combines both the massive parallelism inherent in DNA computing and the directed search capability of evolutionary computing [4][5][6][7][8].

In this paper, we present a DNA based evolutionary algorithm for the *Minimal Set Cover* problem. The rest of this contribution is organized as follows. In section 2, we present a simple description of the Minimal Set Cover problem, a review of evolutionary algorithms and some potential advantages of DNA based evolutionary algorithms. In section 3, we then introduce how the DNA based evolutionary algorithm is implemented. We then give a short analysis of the computing process in section 4. Finally, we conclude in section 5 by present some problems to be further considered in the future.

2 Backgrounds

2.1 The Minimal Set Cover Problem

The Minimal Set Cover problem can be formulated as follows: given a finite set $S = \{1, 2, \cdots, p\}$ and a finite collection $C = \{C_1, C_2, \cdots, C_q\}$ of subset of S, find the smallest subset I of $\{1, 2, \cdots, q\}$ such that

$$\bigcup_{i \in I} C_i = S \tag{1}$$

Obviously, this problem can be solved through an exhaustive search of all the 2^q subsets of I. In reference [9], the authors presented an exhaustive method to tackle this problem based on the sticker model proposed by Roweis *et al.* [16]. Although the sticker model is an ingenious theoretical model, its implementation still remains suspicious by current bio-techniques except for the exponential barrier. Therefore, it is necessary to explore new method to solve it.

2.2 Evolutionary Algorithms

Conceptually, evolutionary algorithms (EAs) mainly get inspirations from Darwin's principle of natural selection — the fittest survives the best. They consist currently of three, more or less different, sub fields called evolutionary programming (EP), evolution strategies (ES), and genetic algorithms (GA). Because of their strong simplifications, EAs have been applied to solve practical problems, with a remarkable success, in a variety of application fields, such as global optimization, machine learning and automatic design.

Essentially, EAs could be regarded as population-based stochastic generate-and-test algorithms. First, an initial population is chosen randomly. Then the survival capacity of individuals is evaluated. The selection process is implemented through a probabilistic function based on the survival capacities of individuals. Individuals with higher survival capacities have higher chance of survival. Thirdly, crossover is introduced to implement information exchange between individuals, and then mutation induces variation in individuals. The basic process of a simple EA can be formulated as [10]:

```
Generate the initial population $P(0)$ at random, and
set $i=0$;
  Repeat
    (1) Evaluate the fitness of each individual in $P(i)$;
    (2) Select parents from $P(i)$ based on their fitness
        in $P(i)$;
    (3) Apply crossover and mutation to the parents and
        get generation $P(i+1)$;
  Until the population converges.
```

Through successive generations, the survival capacities of individuals are improved. As potential solutions are obtained through the evolution of an initial population rather than filtered from the whole solution space, the hybrid of the EAs with DNA computing may offer a promise to tackle the "exponential curse".

2.3 Some Advantages of the DNA Based Evolutionary Algorithms

From the birth of DNA computing, there have been calls [to consider carrying out evolutionary computations using genetic materials in vitro [11][12][13]. This motivation comes from the following facts [17][18][19]:

1. The massive parallelism inherent in biochemical reaction allows the processing of populations billions of times larger than that for conventional computers. The large population is expected to be able to sustain large range of genetic variation, and thus high quality individuals can be generated in fewer generations.
2. Current biotechnology of in vitro evolution can be easily adopted to implement mutation and crossover in one point or multiple points.
3. In the deterministic algorithms of DNA computing, the imperfectness of biological operations is undesirable. While it is tolerable in executing DNA based EAs. To some extent, errors may be regarded as contribution to some kind of variation.

In addition, the massive information storage of DNA molecules also provides a potential power to the computation process. At present, one challenge of this method is how to design a feasible fitness criterion that can physically separate DNA strands according to their fitness.

3 A DNA Based Evolutionary Algorithm for the MSCP

3.1 Encoding Method

From section 2.1, it is easy to see that any possible cover of set S can be represented by a binary string of collections $C_i\,(1 \le i \le q)$. In order to facilitate the operation of genetic crossover and mutation, the following data structure, proposed by Tom Head in the splicing systems [14], is used to represent a candidate solution by DNA sequences (the orientation is from $5'$ to $3'$):

$$E_0 h E_1 C_1^{t/f} E_2 \cdots E_q C_{q+1}^{t/f} E_{q+1} t E_{q+2}$$

where subsequences $E_0, E_1, \cdots, E_{n+2}$ denote sites at which restriction enzymes $RE_0, RE_1, \cdots, RE_{n+2}$ can cut them respectively, and we assume that the resulted sequences are all with particular sticky ends. Subsequences h (head) and t (tail) are mainly used to amplify those sequences representing candidate solutions by polymerase chain reaction (PCR). Subsequence C_i^t (or C_i^f) denote that collection C_i is (or not) contained in some candidate solution.

Assuming k is the maximal cardinality (the cardinality of a collection means the number of its elements) of these collections C_i $(1 \le i \le q)$, then all these collections can be represented through k blocks, where the first k' $(1 \le k' \le k)$ blocks denote the elements $s \in S$ included in C_i, and the rest is empty blocks. Figure 1 shows an example of some collection C_i in case of $k = 6$ and $k' = 4$. For each element $s \in S$, a distinct DNA subsequence with fixed length, say 20bp, is used to denote it. All the empty blocks can be encoded by one particular subsequence with the same length as s for all collections. Based on this, the length of each collection becomes $20k$ bp. Concerning C_i^f for all collections, it is enough to use just one particular sequence that is half-length of C_i to represent them. The length of those candidates thus ranges from $10kq$ bp to $20kq$ bp.

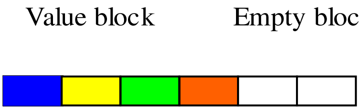


Fig. 1. An illustration of the representation for collection C_i with $k = 6$ and $k' = 4$

3.2 Initialization, Fitness Evaluation and Selection Process

Initialization of the Minimal Set Cover problem consists of the preparation of a combinatorial mixture of the initial population with size N, each individual essentially corresponds to a possible candidate solution. The mix and split combinatorial synthesis technique described in reference [15] can be used to synthesize the initial population, and we recommend readers interested in it to refer to that paper. Here we assume that the synthesized population are kept in tube T_0.

How to define the fitness function of individuals is very important in DNA based EAs is a key step as this will have a great influence on the way how to physically separate individuals efficiently. For the Minimal Set Cover problem, it is natural to define the fitness function of an individual a as

$$f(a) = -l_a \tag{2}$$

where l_a is the length of this individual. Thus, the fewer collections an individual covers the set S, the higher fitness it takes and the shorter its corresponding length is. As some binary representation may contain individuals that don't cover the set S at all, we should first separate them from tube T_0 proceeding the selection process. This can be implemented through p sequential extractions as described in Lipton's paper [15].

The selection process is implemented as the following strategy:

1. Separate individuals in tube T_0 by gel electrophoresis;
2. Regroup them in three distinct tubes, say T_1, T_2 and T_3 according to their length in an increasing order;
3. Amplify the content of the three tubes by PCR and dilute them to the previous concentration, then extract fractions V_l from tube T_l ($1 \le l \le 3$) into tube T_0 such that

$$\sum_{l=1}^{3} V_l = V_0 \tag{3}$$

$$V_1 > V_2 > V_3 \tag{4}$$

3.3 Genetic Crossover and Mutation

In this section, we presents a volume-controlled strategy to implement the crossover and mutation respectively:

3.3.1 Genetic Mutation

1. partition the content of tube T_0 into tubes T_c and $T_c^{'}$ with volumes $V_c = f_1 V_0$ and $V_c = (1 - f_1)V_0$ respectively ($0 \le f_1 \le 1$).
2. Assuming that the ith bit ($1 \le i \le q$) is intended to undergo mutation for some individuals, in other word, to flip their ith bit from C_i^t to C_i^f. Preparing some double substrands $E_i C_i^f E_{i+1}$ in a new tube T, digest them with enzyme RE_i completely, then with enzyme RE_{i+1}. Finally, this will result in double strands with particular sticky ends at both ends.
3. Digest the individuals in tube T_c with enzyme RE_i completely, then with enzyme RE_{i+1}. As this finished, the mutation site , C_i , of these individual will drop off.

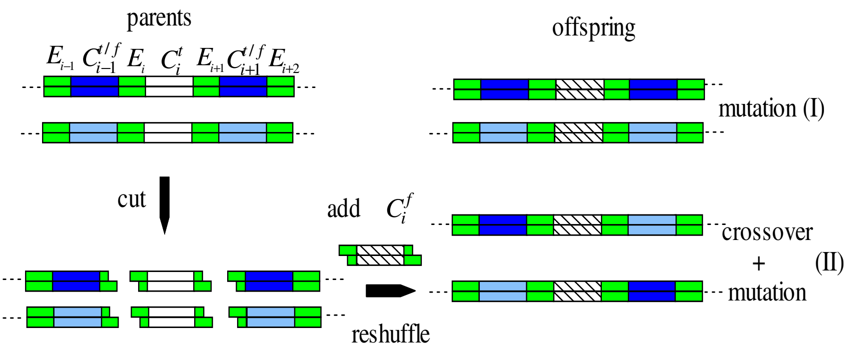

parents
E_{i-1} $C^{t/f}_{i-1}$ E_i C^t_i E_{i+1} $C^{t/f}_{i+1}$ E_{i+2}
offspring
mutation (I)
cut
add C^f_i
crossover
+ (II)
mutation
reshuffle

Fig. 2. A schematic diagram of the genetic mutation process between two individuals is presented. The short green boxes indicate those restriction sites, while other boxes indicate the value bits. This process is mainly implemented through the cut and reshuffle operations which result in two types of possible produced offspring: one pure mutation (I) and the other with crossover (II).

4. Separate the mutation site C_i through gel electrophoresis and clear them, then put the rest back to tube T_c .
5. Add the content in tube T and some ligase to tube T_c , then the ith bit will become C^f_i .

In figure 2, a very simple mutation process is presented between two individuals, where we assume that the ith bit of them is both as C^t_i . In this case, the final recombination process will result in two types of mutation: one pure mutation (I) and the other with both mutation and crossover (II).

3.3.2 Genetic Crossover

1. partition the content of tube T_0 into tubes T_c and $T_c^{'}$ with volumes $V_c = f_2 V_0$ and $V_c = (1 - f_2)V_0$ respectively ($0 \le f_2 \le 1$).
2. Assuming that the ith bit ($1 \le i \le q$) is intended to undergo crossover for some individuals, digest the individuals in tube T_c with enzyme RE_i completely, As this finished, all the individuals will be cut at the ith bit and resulted in segments with a sticky ends in one side.
3. Clearing the enzyme RE_i from tube T_c and then add some ligase to it, then the recombination process will lead to some crossover between those individuals

In figure 3, a very simple mutation process is presented between two individuals, where we assume that the ith bit of them is both as C^t_i . In this case, the final recombination process will lead to some crossover between the two individuals, and some individual may recover to its original situation.

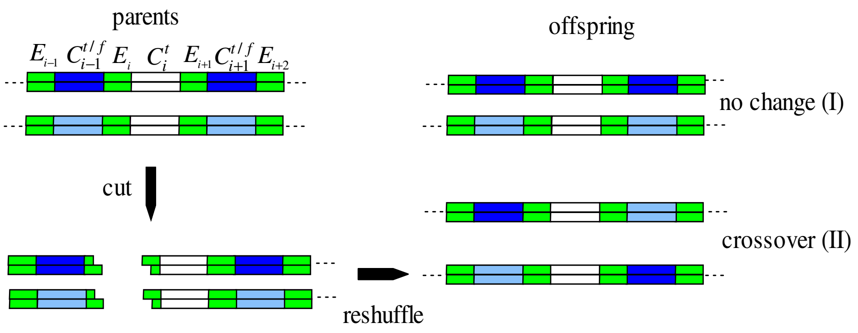

parents
E_{i-1} $C_{i-1}^{t/f}$ E_i C_i^t E_{i+1} $C_{i+1}^{t/f}$ E_{i+2}
offspring
no change (I)
cut
crossover (II)
reshuffle

Fig. 3. A schematic diagram of the genetic crossover process between two individuals is presented. The short green boxes indicate those restriction sites while other boxes indicate the value bits. This process is mainly implemented through the cut and reshuffle operations which leads to some crossover between the two individuals.

4 Discussions

In this section we give some discussion on the genetic parameter involved in the DNA based EAs, such as the population size N, the mutation probability p_m and the crossover probability p_c, and the genetic operations over multiple points.

4.1 Population Size

In traditional evolutionary algorithm, the population size N is usually kept in a constant during the evolution process. But in DNA based EAs, this doesn't hold anymore. This is mainly because of the following 2 reasons:

1. DNA based EAs, each individual may exist in many copies, so it is impossible for us to select all the copies of the individual selected during the genetic mutation operation and crossover operation.

2. It is hard for us to control the how many new individuals may be produced during the genetic mutation and crossover process as the complexity of real bio-chemical reaction.

4.2 Mutation Probability and Crossover Probability

Both the mutation probability p_m and crossover probability p_c are two important parameters in DNA based EAs. Like the population size N, these two parameters also vary in a complex way. From section 3, it is easy to see that:

1. The selection probability f_1 and f_2 have a great influence on the mutation probability p_m and crossover probability p_c. Therefore, we can control the mutation probability p_m and crossover probability p_c by them.

2. In figure 2 and figure 3, we just present a very simple situation of the mutation and crossover process between two individuals. In real computing process, there

involved thousands of individuals to be processed. So the final products of the genetic operation may contain pure mutation, pure crossover, and their combination. This makes the mutation probability p_m and crossover probability p_c even more complicate.

3. As the ligation process plays an important role both in the mutation and crossover process, its efficiency also takes contribution to the mutation probability p_m and crossover probability p_c.

4.3 Genetic Operations over Multiple Points

For some instances, researches have found that it is more beneficial to perform crossover and mutation in multiple points than in single point. Although in this paper we present an implementation of DNA based EAs by one point crossover and mutation, the method can be easily adapted to perform multiple points through the participation of more restriction enzymes in the reshuffle process.

5 Conclusions

Since Adleman's pioneering work, DNA computing has received more and more attention. However, the most urgent issue in DNA computing is how to tackle the "exponential curse" that hinders its application to practical instances. In this paper, we present a DNA based evolutionary algorithm for the Minimal Set Cover problem, which combines both the massive parallelism and the evolution strategy. As the potential solution could be reached through the evolution process, DNA based EAs therefore provides a promising alternative to overcome this disadvantage. Though DNA based EAs show an exciting prospect, there are still many problems to be studied further:

First, at present, the easiest way to evaluate individuals is to separate them according their length. But in practice, the fitness function may present in various form, and it is not always possible for us to transform them into the length of individuals. Therefore, it is desired to develop more sophisticate and flexible means to distinguish DNA molecules.

Second, as the complexity of bio-chemical reaction, it is difficult for us to precisely control the population size N, the crossover probability p_c and the mutation probability p_m during the evolution process. Therefore, how to control these parameters effectively so that the evolution process proceeds to the expected direction is very important for the reach of the potential solution.

Third, theoretical studies should also be begin as the implementation of DNA based EAs differs dramatically from that of the traditional EAs. And developing new genetic operations is also helpful.

The last but not the least, the instances could be solved by our algorithm relies too much on the number of the restriction enzymes. In order to overcome this disadvantage, a specifically designed protein nucleic acid (PNA) may be a promising technique, which can suppress restriction of particular restriction sites and it may be possible to use the same restriction enzyme for multiple stations in the future [20].

Acknowledgement

The work of this paper is partially supported by the National Scientific Foundation of China under grants 60403002, 60403001 and 30470414 and the China Postdoctor Scientific Foundation under grant 2004036130. The authors are gratefully acknowledged to the anonymous reviewers for their constructive suggestions that make this paper a more valuable form. And thanks also to the authors that appeared in the references.

References

1. Adleman, L.: Molecular computation of solution to combinatorial problems. Science, 266(1994)1021-1024
2. Hartmanis, J.: On the weight of computations. Bulletin of the European Association for Theoretical Computer Science, 55(1995)136-138
3. Ogihara, M.: Breadth first search 3-SAT algorithms for DNA computers. Technical Report TR 629, University of Rochester, Department of Computer Science, Rochester(1996)
4. Bänk, T., Kok, J. and Rozenberg, G.: Cross-Fertilization between Evolutionary Computation and DNA-based Computing, *Proceedings of the IEEE Congress on Evolutionary Computing*(1999)980-987
5. Chen, J., Wood, D. H.: Computation with Biomolecules, PNAS, 97(2000)1328–1330
6. Deaton, R., Murphy, R. C., Rose, J. A., Garzon, M., Franceschetti, D. R. and Stevens, S. E.: A DNA based Implementation of an Evolutionary Search for Good Encodings for DNA Computation. *Proceedings of the IEEE International Conference on Evolutionary Computation* (1997)267-272
7. Wood, D., Bi, H., Kimbrough, S. O., Wu, D., Chen, J.: DNA Starts to Learn Poker. *Proceedings of the 7th International Meeting on DNA-based Computers* (2001)23–32
8. Rose, J. A., Hagiya, M., Deaton, R., Suyama, A.: A DNA-based in vitro Genetic Program. Journal of Biological Physics. 28(2002): 493–498
9. Paun. G., Rozenberg G., Salomaa A., DNA computing: New Computing Paradigms, Spring-Verlag Berlin Heidelberg(1998)
10. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
11. Dove, A.: From bits to bases: Computing with DNA. Nature Biotechnology, 16(1998) 830-832
12. Pool, R.: Forget silicon, try DNA. New Scientist. 151(1996) 26-31
13. Willem P. C. Stemmer.: The evolution of molecular computation. Science, 270(1995)1510-1518
14. Head, T., Kaolan, P. D., Bladergroen, R. R.: Computing with DNA by operating on plasmids. Biosystem, 57(2000)87-93
15. Lipton, R. J.: DNA Solution of Hard Computation Problems. Science, 268 (1995)542-545
16. Roweis, S., Winfree, E., Burgoyne, R., Chelyapov, N., Goodman, M., Rothemund, P., Adleman.: A sticker based architecture for DNA computation. Journal Computational Biology, 5(1998):615-29,.
17. Wood, D., Chen, J.: Physical Separation of DNA According to Royal Road Fitness. *Proceedings of The IEEE Conference on Evolutionary Computation*, Dantsin, Voronkov (1998)

18. Chen, J., Antipov, E., Lemieux, B., Cedeno, W., Wood, D. H.: A Design for DNA Computation of the OneMax Problem. Soft Computing, 5(2001)19-24
19. Bäck, Th., Kok, J.N., and Rozenberg, G.: Evolutionary computation as a paradigm for DNA-based computing. Natural Computing Series, Springer (2003)15-40
20. Nielsen, P., Egholm, M., Berg, R., Buchardt, O.: Peptide nucleic acid (PNA). Nucleic Acids Research, 21(1992)197–200

DNA Computing Model of Graph Isomorphism Based on Three Dimensional DNA Graph Structures

Zhixiang Yin[1], Jianzhong Cui[1], Jing Yang[1], and Guangwu Liu[2]

[1] Department of Mathematics and Physics,
AnHui University of Science and Technology, Huainan City 232001, China
zxyin66@163.com
[2] Department of Mathematics,
Wuhan University of Technology, Wuhan City 430063, China
zxyin66@163.com

Abstract. An DNA computing model of solving the graph isomorphism problem with 3-D DNA structures is proposed in this paper. The k-armed branched junction molecules are used to encode k-degree vertices. Double stranded molecules are used to encode edges. These molecules are to be mixed in a test tube to be ligated. The reaction product can be detected by gel electrophoresis. The time complexity of the algorithm is $o(n^2)$, where n is the number of vertices of the graph.

1 Introduction

Computation problem is always the central or innermost part in science and technology field. About the methods and theories of computation problem, human had experienced arithmetic theory, numerical computation theory, optimization theory and bionic computation theory developed in the past half century. According to the computation tools, human has experienced manual computation, abacus, calculator, and the present computer epoch. The electronic computer was born in 1946. The development of electronic computer undergoes 5 phase: vacuum tube, transistor, small and medium scale integrated circuit and large scale integrated circuit (LSI). It is well known that computer has act as the promoter action in the development of human society. Graph theory and operations research theory are greatly promoted by the development of electronic computer. However, with the development of the technology and human society, and because of the limitation of the electronic computer in manufacturing techniques, the shortage of the memory and the relatively low operating speed, scientist are now considering develop other types of computers, thus, some new computing patterns, such as bionic computer, quantum computer, photon computer and etc. are springing up. With the advance of the bioscience, especially, the implementation of the human genome plan, both give great impetus to the development of the bionic computing, in which artificial neural network (ANN) and DNA computing are two main bionic computing models.

Research on DNA computing was initialized in 1994, when Adleman proposed a method of solving a small instance of the Hamiltonian Path problem by a laboratory

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 90 – 98, 2005.

experiment involving DNA molecules[1]. Ever since then, substantial efforts have been invested into this newly initiated field mainly because new structures and new operations on them make possible that problems which are intractable for electronic computers can be solved in the new framework. One of the major achievements of computer science in the last two decades is to understand that many important computational search problems are NP-complete, taking SAT problem into consideration, thus are unlikely to have efficient algorithm implemented on silicon-based computer. Hence, DNA computing was proposed as an alternative computing paradigm to electronic computers for solving NP-complete problems. So, the research on DNA computing has become hot- topic issue at the crossroad of mathematics, biology, chemistry and computer science. Furthermore, the research on DNA computing model for solving SAT problem gradually becomes the mainstream of the research efforts, considered that it is a NP-complete problem.

Later, Lipton demonstrated how a large class of NP-complete problems could be solved by encoding the problem in DNA molecules[2]. In particular, Lipton showed one famous NP problem, the so-called “satisfiability” problem(SAT) and subsequently the other NP-problems could be encoded and solved using molecules. [2]. Cukras developed the theory of RNA computing and a method for solving the ‘Knight Problem’ [3]. In 1997, Ouyang et al. presented a molecular biology based experimental solution to the maximal clique problem[4]. Sakamoto exploited and encoded constraints of Boolean operation into the hairpin formation of single-stranded DNA molecules to solve an instance of the SAT problem[5]. Liu introduced a surface-based computing strategy for SAT problem which made a further step toward the development from theoretical DNA computing to practical DNA computer [6]. Wu analyzed and made some improvement in the terms of combinatorial encodings of Boolean variables as well as simplifying the biological operation[7]. In 2001, Benenson et al. designed a programmable and autonomous computing machine made of biomolecules[8]. In 2002, Braich accomplished finding the unique solution of a 20-variable 3-SAT problem after an exhaustive search of more than 1 million (2^{20}) possibilities. The computational complexity of this size has been considered to be the largest in the history of unaided human computation [9].

The advantages of DNA computing are its massive parallelism and enormous information storage capacity. A model has been proposed by Smith et al. to perform computation on DNA strands attached to a surface. This method is called surface-based computing[10].

For the above mentioned algorithms, the general approach is to treat the DNA molecules as linear strings where much of the information content is encoded in the order of nucleotides that make up the DNA. However, Jonoska[11] et al. demonstrated that 3-D structures could be used to solve the 3-SAT problem in constant number of steps regardless of the size of the graph, and they found that the use of 3-D DNA structures could significantly reduce the time and steps needed to identify a solution.

The algorithm for solving the graph isomorphism problem using a 3-D DNA structure is proposed in this paper. We present procedures for solving graph isomorphism.

The number of steps required for the algorithm is $O(n^2)$, where n is the number of vertices of the graph.

Details of some terms and denotations can be seen in Refs. [12],[13] and [14].

2 Structure and Operations of DNA

DNA computing is a newly emerging computing methodology in which DNA molecules are used as means for computation. The DNA molecule is one of the most compact supports of information and has a very important feature from a computational point of view. We present briefly the composition and structure of DNA molecules here.

DNA molecule is a high-molecular weight compound whose basic unit is nucleotide. Every nucleotide consists of one phosphate group, one deoxyribose sugar and one nitrogen-containing base. There are four types of nucleotides which differ in the chemical group: Adenine (A), Guanine (G), Cytosine(C) and Thymine (T), respectively.

DNA molecule also has a regular form. DNA molecule has double helix structure. Base pairing must comply with the Watson-Crick complementarily. That is, A pairs with T, and C with G (figure.1). Otherwise, it is called mispairing. Although there are only four types of bases and two types of base pairing, DNA molecules vary a lot due to the different sequences of these base pairings.

This is the Watson-Crick complementarily rule, and when a single-stranded DNA molecules pair with its complement, a stable double helix is formed.

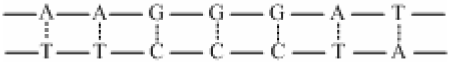


Fig. 1. A single-stranded DNA molecule pairs with its complement, the dots in figure show weak hydrogen bonding between A and T and G and C

Since the precise structure of DNA molecule was underpinned, many biotechniques have been invented including cutting, ligating, electrophoresis, polymerase chain reaction (PCR) , heating, annealing, and so on, to manipulate DNA molecules for computational efforts. These bio-techniques can also be used to help us to figure out the mechanisms of information storage and output.

Various models of DNA computing are based on different combinations of the following biological operations on DNA strands. We briefly summarize them here:

1. Synthesis: Synthesis of a desired DNA strand by means of synthetic technique

2. Annealing and melting: the hydrogen bonding between two complementary sequences is weaker than the one that links nucleotides of the same sequences. It is possible to pair two anti-parallel and complementary single strands, and it is possible to separate (melt) them, obtaining two single strands from a double one. These operations are realized by creating adequate conditions of temperature, pH, etc.

3. Amplifying: make copies of DNA strands by means of the Polymerase Chain Reaction

4. Separate: Separation of the strands according to their different physical or chemical properties by means of gel electrophoresis

5. Ligation: paste DNA strands with complementary sticky ends by means of ligases.

6. Detection: check and acquire the object DNA sequence from products of reaction

All models of DNA computation are based upon the biological operation over a set of specific sequence of DNA molecules. Note that some operations confine only to certain models of DNA computing.

2.1 Synthesis

Oligonucleotides can be synthesized in laboratory. The synthesizer is supplied with the four nucleotide bases in solution, which are combined according to the sequence designated by the user. The instrument synthesizes millions of copies of the required oligonucleotide and places them in solution in a small vial.

2.2 Denaturing, Annealing and Ligation

Double-stranded DNA may be dissolved into single strands(denatured) by heating the solution to a temperature determined by the composition of the strand[15]. Annealing is the reverse process of the denaturing. When a solution of single strands is gradually cooled, single stranded DNA molecules will find their complements and bind together. In double stranded DNA molecules, if either one of the single strands contains a "nick", this "nick" can be sealed by ligase and process is called ligation. This technique allows us to produce a unified molecule from individual single strands bound together by their corresponding complements.

2.3 Hybridization Separation

The hydrogen bonding between two complementary sequences makes it possible to pair two antiparallel and complementary single strands. This operation can be realized by creating adequate conditions of temperature, PH, etc.

2.4 Gel Electrophoresis

Gel electrophoresis is an important technique for sorting DNA strands. Electrophoresis is the movement of charged molecules in an electric field. Since DNA molecules carry negative charge, when put in and electric field they tend to migrate towards the positive pole. The rate of migration of a molecule in an aqueous solution depends on its shape and electric charge. Since DNA molecules have the same charge per unit length, molecules in equal length migrate at the same rate in an aqueous solution. However, if electrophoresis is carried out in a gel(usually made of agarose, polyacrylamide or a combination of the two), the migration rate of a molecule is also affected by its size.

2.5 PCR

Another useful method of manipulating DNA molecule is the Polymerase Chain Reaction(PCR). PCR is a process that can quickly amplify the amount of DNA in a given solution. Each cycle of the reaction doubles the quantity of each strand, leading to an exponential growth in the number of strands.

2.6 Restriction Enzymes

Restriction endonucleases (often referred to as restriction enzymes) recognize a specific sequence of DNA , known as a restriction site. Any double stranded DNA that contains the restriction site within its sequence can be cut by the enzyme at that position.

3 Graph Isomorphism and Its Algorithm

Graphs in this paper are referred to simple graphs. Let $V(G)$ and $E(G)$ be the vertex set and edge set of a graph, respectively.

3.1 DNA Computing Model About Graph Theory and Combinatorial Optimization Problem

At present, the research of DNA computer is still at “experiment " stage, none of universal computer model has been provided so far. After Adleman reported his algorithm for solving Directed Hamilton Path Problem based on DNA computing, many scholars issued some DNA computing models of NP-problems in graph theory[16] [17] [18]. Some Chinese scholars utilize the advantage of graph and combinatorial optimization, adopt various DNA computing model, set up some DNA computing models of NP-complete problem and hardly computing problem. What deserves to be mentioned is, they have set up initially DNA computing models of such problems as planning problem and Chinese Postman etc. Main results are as follows:

Liu Y. et al set up DNA computing models of covering problem in graph theory corresponding to different algorithm [19] [20] [21]. Yin et al Set up initially surface based DNA computing model of 0-1 integer programming problem and DNA computing model of Chinese Postman problem[22] [23]. Liu W. et al provided DNA computing model of Hamilton Path or Hamilton Cycle problem in a weighted digraph or graph, and Set up DNA computing model of 3-SAT problem[24] [25] [26]. Wang et al gave DNA computing model of bipartite graphs for Maximum Matching problem[27]. Pan et al set up initially surface based DNA computing model of minimal vertex cover Problem[28]. However, the model of isomorphic problem in graph theory DNA computing has never been studied. In this paper, An algorithm of solving the graph isomorphism problem with 3-D DNA structures is proposed. The k-armed branched junction molecules are used to encode k-degree vertices. Double stranded molecules are used to encode edges. These molecules are to be mixed in a test tube to be ligated. The reaction product can be detected by gel electrophoresis.

3.2 Graph Isomorphism

Two graphs G and H are said to be isomorphic (denoted by $G \cong H$) if there are bijections, θ: $V(G) \to V(H)$ and ϕ: $E(G) \to E(H)$, such that $\psi_G(e) = uv$ if and only if $\psi_H(\phi(e)) = \theta(u)\theta(v)$. Such a pair (θ, ϕ) of mapping is called an isomorphism between G and H. The graph isomorphism problem is another standard NP-complete problem in combinatorial optimization. It has abroad applications, for example, isomorphism of model in system modelling, comparability of molecule in chemistry, and rationality of component installation in mechanics.

3.3 Regular Algorithm for Graph Isomorphism

Step1: give out the degree sequences of the graphs. If they are not the same as each other, stop. And at this point, we can conclude that the two graphs are not isomorphic. Otherwise, go to step 2.

Step2: make out all possible corresponding relations of vertices with the same degree in two graphs.

Step3: check the relations above, if there are bijections, θ: $V(G) \to V(H)$ and ϕ: $E(G) \to E(H)$, such that $\psi_G(e) = uv$ if and only if $\psi_H(\phi(e)) = \theta(u)\theta(v)$, stop. The two graphs are said to be isomorphic. Otherwise, they are not isomorphic.

The execution of this algorithm is extremely hard for electric computers. Because the time required for identifying a solution increases exponentially with the sizes of the graphs.

4 Algorithm with 3-D DNA Graph Structure

4.1 Construction of *k*-Armed Vertex DNA Molecule

For a given *k*-degree vertex of the graph, a *k*-armed molecule is used to denote it. 3' end of the molecule need to be extended. If adjacent vertices of such a vertex (*k* degree) are $d_1, d_2, \cdots, d_k$, respectively, then the extending length of DNA segments at 3' ends of *k* arms of the *k*-armed molecule are correspondingly $10 \times d_1, 10 \times d_2, \cdots, 10 \times d_k$. The *k*-armed DNA molecule constructed by this method can distinguish the connecting relation of the *k*-degree vertex.

4.2 Construction of the Edge DNA Molecule

Each edge isrepresented by a regular double stranded DNA molecule. The two 3' ends of it also need to be extended. The extended part must be the complementary strand of anyone of the 3' end extension part of the vertex associated with the *k*-degree vertex. Such construction method can ensure the formation of the 3-D DNA graph as well as preserve the connection characteristic of the graph. A *k*-armed molecule can be synthesized by *k* single stranded DNA molecules through biologic method. Fig. 2 is examples of the 2-armed, 3-armed and 4-armed branched DNA molecules.

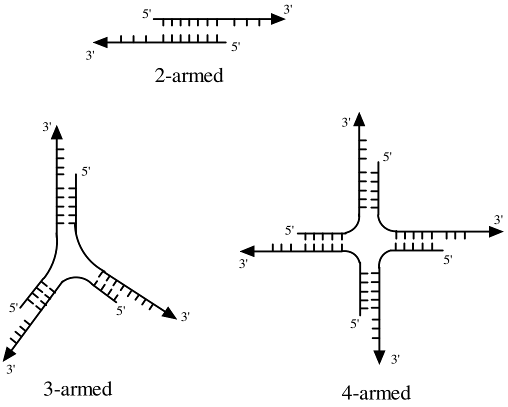


Fig. 2. *k*-armed DNA molecules for the *k*-degree vertices, this shows a figure consisting of 2-armed, 3-armed, 4-armed DNA molecular

4.3 Biology Operations for the Isomorphism Problem

For the two given graphs, we construct the vertex and edge DNA molecules by the method described in sections 3.1 and 3.2. Put these DNA molecules into two different test tubes. The same DNA primer and ligase were added into each test tube. This enables the vertex molecules and edge molecules to ligate together to form double strands at the 3' ends. Then the same exonuclease was added in the two tubes. The single stranded parts of the unmatched DNA molecules will be hybridized. In this operation, all possible connection will appear because of the randomicity of DNA molecules' ligation. In fact, all the corresponding relation of the vertices in the graph will be formed, i.e., all the corresponding relation of the vertices will be found with following operations. Finally, we can see whether or not the two graphs are isomorphic by gel electrophoresis. If the gel electrophoresis figures are different, two graphs depicted by the figures are not isomorphic.

For two graphs which are not isomorphic, there is the possibility that they have the same degree sequence or degrees of vertices adjacent to the vertex which have the same degree are the same. The operations mentioned above can not solve the graph isomorphism problem, i.e., gel electrophoresis figures of the two graphs which are not isomorphic may be the same according to the operations described above. For this reason, we add the following operations:

For the convenience of description, we use G_1 and G_2 to represent two graphs. For G_1, delete a k-degree vertex and its conjuncted edges. In the biologic operation procedure, we do not add DNA molecules of this vertex and its conjuncted edge molecules in the test tube. Mark the tube with T_{G_1}.

For G_2, delete anyone of the k-degree vertices and its conjuncted edges. Its biologic operation is the same as (1). Suppose that the number of k-degree vertices is s, we should construct s test tubes corresponding to the k-degree vertices. Mark them with $T_{1G_2}, T_{2G_2,} \cdots, T_{sG_2}$, respectively.

Add DNA primer and ligase in T_{G_1}, this will ensure the vertex molecules and edge molecules to be ligated to form double strands at the 3' ends extension part. Then add the same exonuclease in each tube to hybridized single stranded parts of the unmatched DNA molecules. Gel electrophoresis is carried out with the products and the electrophoresis figures are preserved.

Add the same DNA primer and ligase as (3) in $T_{1G_2}, T_{2G_2,}, \cdots, T_{sG_2}$, which will ensure the vertex molecules and edge molecules to be ligated to form double strands at the 3' ends extension part. Then add the same exonuclease as (3) in each tube to hybridized single stranded parts of the unmatched DNA molecules. Gel electrophoresis is carried out with the products and the electrophoresis figures are preserved.

Compare the gel electrophoresis figures of (3) and (4). If the two figures are not the same, the two graphs are not isomorphic, stop. Otherwise, go to (1) and repeat steps (1) to (5).

5 Conclusion

Let the number of vertices of the two graphs G_1 and G_2 be n, and their number of edges be m. We use $s_k (k = 2,3,\cdots,t)$ to denote the *k*-degree vertices, then $\sum_{k=2}^{t} s_k = n \cdot$ We only need construct $m + s_k$ regular double stranded DNA molecules and s_k *k*-armed ($k = 3,\cdots,t$) molecules. The maximum number of steps of our algorithm is $s_2{}^2 + s_3{}^2 + \cdots + s_k{}^2$. It is obvious that $s_2{}^2 + s_3{}^2 + \cdots + s_k{}^2 < (s_2 + s_3 + \cdots + s_k)^2 = n^2$, therefore, the complexity of our algorithm increases linearly with the size of the problem.

Acknowledgment

The project supported by CNSF (Grant number: 60274026, 60373089 and 30370356), China Postdoctoral Science Foundation (2004035196), Postdoctoral Foundation of Huazhong University of Science and Technology and Doctoral Foundation of Anhui University of Science and Technology.

References

1. Adleman L.M.: Molecular computation of solutions to combinatorial problems. Science, 266 (1994) 1021-1024
2. Lipton R.J.: DNA solution of hard computation problem. Science, 268 (1995) 583-585
3. Cukras A.R, Faulhammer D, Lipton R.J et al.: Chess games: A model for RNA-based computation. Biosystems, 52(1999) 35-45
4. Ouyang Q.: DNA solution of the maximal clique problem. Science. 278 (1997) 446-449
5. Sakamoto K, Gouzu H, Komiya K et al.: Molecular Computation by DNA Hairpin Formation. Science, 288 (2000) 1223-1226
6. Liu Q.H.: DNA computing on surfaces. Nature, 403 (2000) 175-179

7. Wu H.Y.: An improved surface-based method for DNA computation. Boisystems, 59 (2001) 1-5
8. Benenson Y., Paz-Elizur T., Adar, R., et al.: Programmable and autonomous computing machine made of biomolecules. Nature, 414 (2001) 430-434
9. Braich RS, Chelyapov N, Johnson C et al.: Solution of a 20-Variable 3-SAT Problem on a DNA Computer, Science, 296 (2002) 499-502
10. Smith L. M., Corn R. M., Condon A. E., et al.: A surface-based approach to DNA computation. Journal of Computational Biology, 5 (1998) 255-267
11. Jonoska, N., Karl, S. A., Saito, M.: Three dimensional DNA structures in computing. In: Preceeding of the 4th DNA Based Computing Workshop, Philadephia: Springer-Verlag, (1998)
12. Yin Z. X., Zhang F. Y., Xu J.: The general 0-1 programming problem based on DNA computing. Biosystems, 70 (2003) 73-79
13. Paun G., Grzegorz R., Arto S.: DNA Computing: New Computing Paradigms, Springer-Verlag, Berlin Heidelberg New York (1998)
14. Bondy J.A., Murty U.S.R.: Graph Theory with Applications. The Macmillan Press LTD, New York (1976)
15. Breslauer K. J., Frank R., Blocker H., et al.: Predicting DNA duplex stability from the base sequence. Proc.Natl. Acad. Sci., 83 (1986) 3745-3750
16. Pancoska P, Moravek Z, Moll U.M.: Rational design of DNA sequences for nanotechnology, microarrays and molecular computers using Eulerian graphs. Nucleic Acids Research, 15 (2004) 4630-4645
17. Beigel R., Fu B.: Molecular Approximation Algorithm for NP Optimization Problems. 3rd DIMACS Meeting on DNA Based Computers, Univ. of Penns (1997)
18. Kari L., Gloor G., Yu S.: Using DNA to solve the Bounded Post Correspondence Problem. Theoretical Computer Science, 2(2000) 193-203
19. Liu Y., Xu J., Pan L. et al.: DNA solution of a graph coloring problem. Journal of Chemical Information and Computer Science, 42(2002) 524-528
20. Liu Y., Guo X., Xu J.et al.: Some Notes on 2-D Graphical Representation of DNA Sequence. Journal of Chemical Information and Computer Science, 42(2002) 529-533
21. Liu W. B., Xu J.: A DNA Algorithm for the Graph Coloring Problem. Journal of Chemical Information and Computers, 42 (2002) 1176-1178
22. Yin Z., Zhang F., Xu J.: A Chinese postman problem based on DNA computing. Journal of Chemical Information and Computer Sciences, 42(2002) 222-224
23. Yin Z. X., Zhang F. Y., Xu J.: A General 0-1 Programming Problem Based On DNA Computing. Biosystem, 70(2003)73-78
24. Liu W. B., Wang S. D., Xu J.: DNA Sequence Design Based on Template Strategy. J. Chem. Inf. Comput. Sci., 43(2003) 1876-1881
25. Liu W. B., Wang S. D., Xu J.: Solving the 3-SAT Problem Based on DNA Computing. J. Chem. Inf. Comput. Sci., 11 (2003) 1942-1946
26. Liu W. B., Wang S. D., Xu J.: The Hamiltonian Cycle Problem Based on DNA Computing. Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution And Learning, 1 (2002) 313-317
27. Wang S.: DNA Computing of Bipartite Graphs for Maximum Matching. Journal of Mathematical Chemistry, 31(2002) 271-279
28. Pan L. Q., Xu J.: A Surface-Based DNA Algorithm for the minimal vertex cover Problem. Progress in Natural Science, 13 (2003) 81-84

A DNA-Based Genetic Algorithm Implementation for Graph Coloring Problem

Xiaoming Liu[1], Jianwei Yin[1], Jung-Sing Jwo[1], Zhilin Feng[1,2], and Jinxiang Dong[1]

[1] Department of Computer Science and Technology,
Zhejiang University, HangZhou 310027, China
[2] College of Zhijiang,
Zhejiang University of Technology, Hangzhou 310024, China
{liuxiaoming, zjuyjw, djx}@cs.zju.edu.cn

Abstract. This paper presents an implementation of Croitoru's genetic algorithm for graph coloring problem, and some necessary modification and simplifying are made by using DNA operations. In this algorithm, each vertex and edge is encoded with a series of encodings incorporating position information, and the initial diverse candidate population is generated using POA. One crossover operator, two mutation operators, evaluation and selection operators are all implemented using basic operations on DNA. It is shown that the algorithm can be implemented with space complexity much decreased and time complexity $O(mn^2)$ to get a new generation, where n is the number of vertices and m is the number of edges. Moreover, borrowing ideas from the above implementation, an algorithm for Maximal Clique problem is also presented.

1 Introduction

The graph coloring problem is, for a given graph G=(V,E), coloring the vertices with the property that for any pair vertices connected by an edge, they are colored differently, the minimum number of color required is called the chromatic number of G and is denoted by X(G). It is well known that it is NP-complete to decide whether for a given graph G and an integer k, there exists a k-coloring of G. There are many optimal methods for the problem, for example simulated annealing in [1] and genetic algorithm in [2].

In 1996 Amos *et al.* proposed a DNA algorithm to solve the 3-coloring problem based on a test tube [3]. In another approach, Bach et al. solving the 3-coloring problems through generating all subsets of vertices V whose size are less than n/3 [4].

Developing efficient DNA-based algorithms for solving NP-complete problems is one of the most important issues of DNA-based computing. Adleman [5] and many researchers have explored the possibility of solving NP-complete problems using DNA, but it is required to generate all the solution candidates for most algorithms, which means the quantity of DNA used grows exponentially with the size of problems. Thus, the size of instances that can be solved with these algorithms is quite limited. Several suggestions [4] have been proposed to reduce the amount of DNA to be used. Bach *et al.* [4] proposed methods for solving Clique and Independent Set, where solution candidates to the instance are generated by combining (in all possible

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 99 – 108, 2005.

combinations) solutions for its sub-problems. Cai *et al.*[6] proposed a surface-based method, and Sergio *et al.*[7] proposed an implementation of a random walk method for solving k-SAT problems with the space complexity $O((2-2/k)^n)$. Since the beginning of DNA computing concept was proposed, there have been calls [8] to consider carrying out evolutionary computation using DNA, but only a few implementations have been presented. In 1999 Chen et al. [9] proposed an implementation of genetic algorithm for Max-1 problems. Rose et al [10] proposed a genetic method for producing massively parallel, recombined gene libraries encoding for proteins.

The purpose of this paper is to study a possible implementation of a genetic algorithm proposed by Croitoru *et al* [2]. for graph coloring problem. The paper is organized as follows. In section 2 we give a brief description of Croitoru's algorithm. In section 3 we describe a DNA-based implementation of the algorithm. In section 4, we present a solution for the maximal clique problem. In section 5 we conclude the paper.

2 Brief Description of Croitoru's Algorithm

Let G=(V,E) be a graph with the vertices set V={1,2…n}. The set of all n! permutations on the set V is denoted by S_n. For an element $v=v_1 \ldots v_i \ldots v_j \ldots v_n$ of S_n, if $e=v_i v_j$, i<j and $v_{i+1} \ldots v_{j-1}$ is a stable set (including empty set Ø, see [2] for definition), then e is called a bad edge with respect to v, and denoting by b(v) the number of all bad edges in G with respect to v. There is the following theorem:

$$\chi(G) = 1 + \min_{v \in S_n} b(v)$$

The theorem says, for a graph, the problem of obtaining an optimal coloring is equivalent to find an ordering with minimum number of bad edges. Croitoru's algorithm is based on the theorem.

Croitoru's algorithm is a genetic algorithm with 1 crossover operator and 4 mutation operators. A chromosome represents an ordering of the vertices of the graph. If the graph has n vertices, the chromosome will be a vector: $chrom=(v_1,v_2 \ldots v_n)$, where $v_i \in \{1,2 \ldots n\}$, $v_i \neq v_j$, $i \neq j$.

Crossover: Let $parent^1=(v_1^1 v_2^1 \ldots v_n^1)$ and $parent^2=(v_1^2 v_2^2 \ldots v_n^2)$ be two parent chromosomes. Two cutting points $C^1, C^2 \in \{1,2 \ldots n-1\}$ are generated randomly for $parent^1$ and $parent^2$ respectively. One offspring is obtained by keeping unaltered genetic information from $parent^1$ before C^1, the vertices after C^1 from the first parent are rearranged using the ordering defined by the second parent. The second offspring is constructed similarly.

For example: $\left\{\begin{matrix} parent^1 = 316 \downarrow 254 \\ parent^2 = 52 \downarrow 4136 \end{matrix}\right. \Rightarrow \left.\begin{matrix} offspring^1 = 316524 \\ offspring^2 = 523164 \end{matrix}\right\}$

Order Mutation: Let $parent=(v_1 v_2 \ldots v_n)$ be a parent chromosome, the offspring is obtained by exchange 2 vertices located in 2 randomly generated positions.

For example, $parent = 31\overset{\downarrow}{6}25\overset{\downarrow}{4} \Rightarrow offspring = 314256$

Block Mutation: The operator translates blocks of k successive vertices (k is randomly generated). Let $parent=(v_1 v_2 \ldots v_n)$ be a parent chromosome. If k=2 and i,j∈[1,n-1] are randomly generated, the block mutation yields offspring=($v_1 \ldots v_{i-1}v_{i+2} \ldots v_jv_iv_{i+1}v_{j+1} \ldots$).

Color Spread Mutation and Bad Edge Stretch Mutation are not considered in this paper, so they are omitted here. Two evaluations are proposed in their algorithm: Evaluation by Bad Edge and Evaluation by heuristics. We consider only the first evaluation. The selection is rank-based and the number of ranks is equivalent to the number of fitness values.

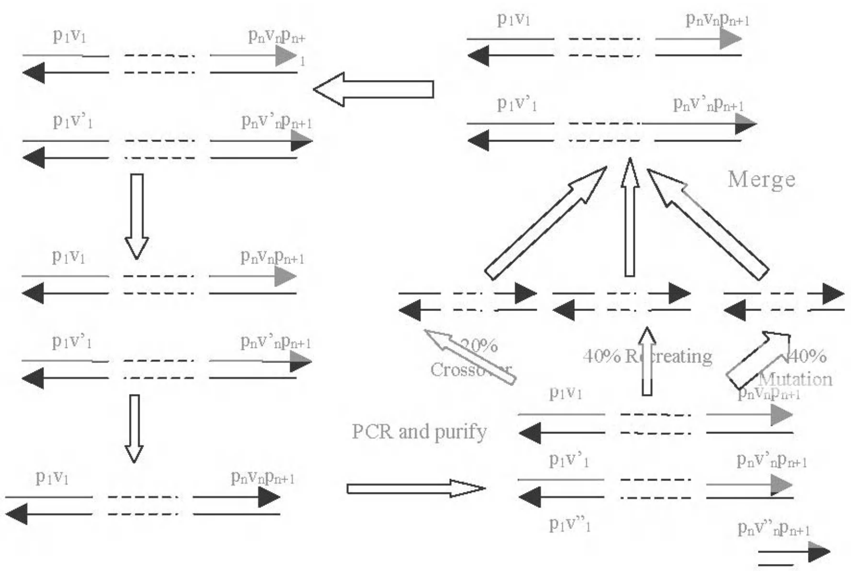


Fig. 1. DNA Implementation Cornelius Genetic Algorithm

3 DNA Implementation of Croitoru's Algorithm

The implementation of Croitoru's algorithm based on DNA is outlined below (see Fig.1 also).

Begin with a diverse initial population of candidates.

1. Evaluate the fitness of the candidates.
2. Select and purify more fit candidates.
3. Amplify fit candidates with PCR.
4. Reserve some, crossover a part and mutate others.
5. Combine all the candidates from step 4, and obtaine a new generation. Repeat.

3.1 Operations Used in the Implementation

To implement the algorithm, we permit the following operations on DNA. For the implementation of To-Single-Stranded and To-Double-Stranded, please see [11]. For convenience of description, we extend the extract operation.

1. Merge: mixing the contents of two or more test tube into one, denote by T←Merge(T_1,T_2…T_n).
2. Detect: testing whether a test tube contains a DNA strand.
3. Cut: cutting DNA strands at specific restriction sites, denote by Cut(T,s_1|s_2).
4. Length: separating DNA strands according to their base length, denoted by T←length(T_0,l).
5. Extract: extracting all strands containing certain subsequences, denoted by T←Extract(T_0,(s_1,s_2…s_n)).
6. To-Single-Stranded: denature each dsDNA in tube and remove one ssDNA, denoted by T←To-Single-Stranded(T_0).
7. To-Double-Stranded: making ssDNA to dsDNA, denoted by T←To-Double-Stranded(T_0).

3.2 Encoding Scheme and Generating of Initial Candidate Pool

In order to implement the algorithm, we employ a specific DNA encoding scheme (expanding the encoding scheme of [12]). We use dsDNA to encode the permutation of vertices and the encoding looks like $p_1v_1p_2v_2…p_nv_np_{n+1}$. Let G=(V,E) be a graph with |V|=n, |E|=m. For each vertex v_i, we use a series of encoding p_1v_i, $p_2v_i…p_nv_i$ to denote it, where p_j ($1\leq j\leq n$) means the order of vertex v_i in a specific permutation is j. we can choose the lengths of v_i and p_j ($1\leq i\leq n$, $1\leq j\leq n+1$) to be both l, then the length of a proper permutation will be (2n+1)l. n^2 Kinds of encodings are needed to encode n vertices, but in fact, we need only 2n+1 different encoding segments for the encodings of p_j ($1\leq j\leq n+1$) in each vertex's encoding are the same. For a DNA strand representing a permutation of vertices, there are n value sections (v_1 to v_n) sandwiched between n+1 position sections (p_1 to p_{n+1}) (see Fig.2). The last position section p_{n+1} is needed for the operation of To-Double-Stranded and PCR amplification. To encode an edge e=v_iv_j, we use a series of encoding $v_ip_kv_j$ ($1\leq k\leq n$) in the similar way.

To generate an initial candidate pool, we can use POA (Parallel Overlap Assembly) [12]. One thing to mention here is that we do not need to generate all the solution candidates. To get legal permutations of vertices, we need extracting strands with length of 2(n+1)l and having $v_1…v_n$ as subsequences, that is T_0←Length(T_0,(2n+1)l) and T_0←Extract(T_0,(v_1,v_2…v_n)) if we denote the initial tube as T_0.

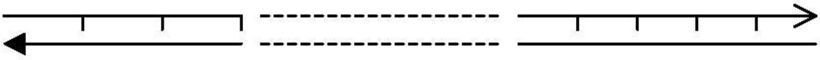


Fig. 2. Encoding for vertices ordering

3.3 Implementation of Genetic Operators

3.3.1 Implementation of Crossover

The crossover operator is designed to propagate and exchange information regarding stable sets defined in the parents throughout evolution. There is a cut point for each parent chromosome and the crossover operation is performed separately on each parent chromosome in original Croitoru's algorithm. We can carry out the operation on all the chromosomes simultaneously and consider only single point crossover since there are a large number of DNA strands presented in a tube when it is implemented using DNA.

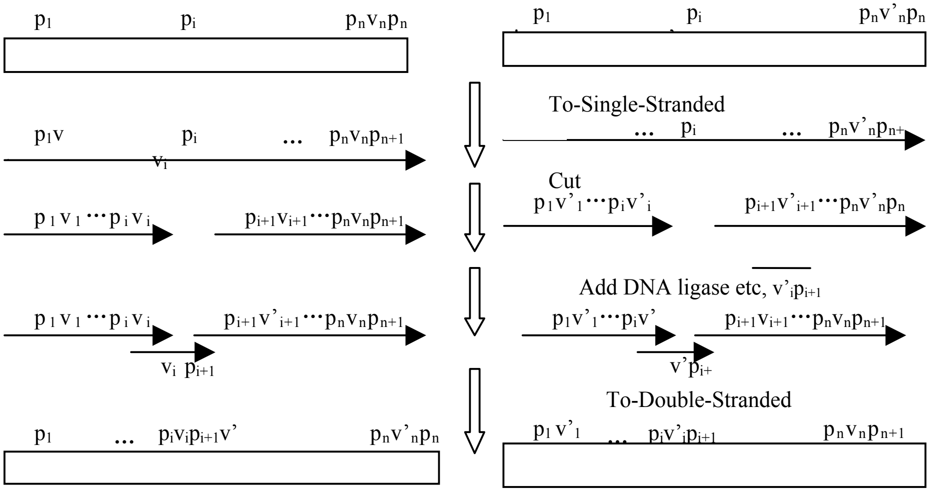

p1 pi pnvnpn
p1v pi ... pnvnpn+1
vi
p1v1···pivi pi+1vi+1···pnvnpn+1
p1v1···pivi pi+1v'i+1···pnvnpn+1
vi pi+1
p1 ... pivipi+1v' pnv'npn
p1 pi pnv'npn
To-Single-Stranded
... pi ... pnv'npn+1
Cut
p1v'1···piv'i pi+1v'i+1···pnv'npn
Add DNA ligase etc, v'ipi+1
p1v'1···piv' pi+1vi+1···pnvnpn+1
v'pi+
To-Double-Stranded
p1v'1 ... piv'ipi+1 pnvnpn+1

Fig. 3. Crossover implementation process

Let the current test tube be T_1 and a randomly generated cut point be i ($1 \le i \le n$). We get tube T_2 after performing the To-Single-Stranded operation. Single strands in T_2 are cut between v_i and p_{i+1}, then ligase enzyme and $\overline{v_i' p_{i+1}} (v_i' \in \{\overline{v_1}, \overline{v_2}, \cdots, \overline{v_n}\})$ are added into tube T_2. We get dsDNA after extracting proper length ssDNA and performing To-Double-Stranded operation. The algorithm is outlined below (see Fig.3. also).

```
1. $T_2 \leftarrow To-Single-Stranded(T_1)$
2. For each $v_i \in \{v_1,\cdots,v_n\}$, $T_2 \leftarrow Cut(T_2, v_i \mid p_{i+1})$
3. For j=1 to n add($T_2, \overline{v_j p_{i+1}}$) [In Parallel]
4. $Ligase(T_2), T_2 \leftarrow Extract(T_2,(v_1,v_2,\cdots,v_n))$,
   $T_2 \leftarrow Length(T_2,(2n+1)l)$
5. $T_2 \leftarrow To-Double-Stranded(T_2)$
```

In step 4, a single p_i ($1 \le i \le n$) extraction can be used instead, but for robustness of the algorithm, we extract all n p_i here. After many $\overline{v_j p_{i+1}}$ are added in step 3, the cut ssDNA containing p_1 and $p_i v_j$ will hybridize with $\overline{v_j p_{i+1}}$ and $\overline{v_j p_{i+1}}$ will hybridize with cut ssDNA containing p_{i+1}. The ssDNA formed now may not be proper permutation of vertices (some may appear more than once while some other may not appear), so the extracting operation is performed here. Meanwhile, some ssDNA may be of not proper length so the length operation is performed also. It is guaranteed to get proper permutation of vertices in step 5 after two operations mentioned above are performed.

3.3.2 Implementation of Mutation Operation

A permutation operator is aimed at sampling the search space in a neighborhood of a chromosome. There are 4 mutation operations in Croitoru's algorithm and we give 2 of them here.

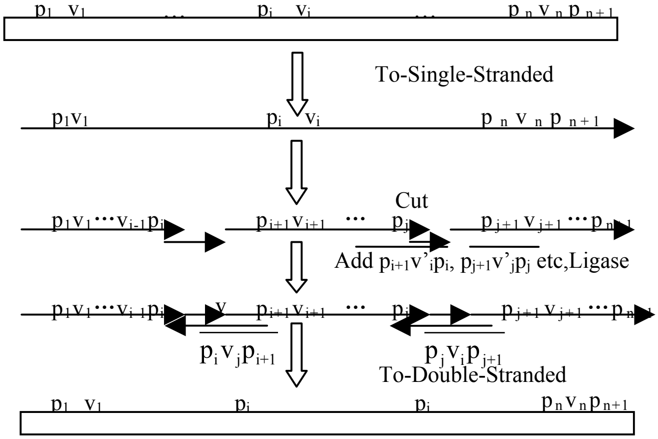

p1 v1 ... pi vi ... pn vn pn+1
To-Single-Stranded
p1v1 pi vi pn vn pn+1
Cut
p1v1···vi-1pi pi+1vi+1 ··· pj pj+1vj+1···pn+1
Add pi+1v'ipi, pj+1v'jpj etc,Ligase
p1v1···vi-1pi v pi+1vi+1 ··· pj pj+1vj+1···pn+1
pivjpi+1
pjvipj+1
To-Double-Stranded
p1 v1 pi pj pnvnpn+1

Fig. 4. Order mutation implementation process

Order Mutation: the operator is used to change the order of some vertices in permutation. For two randomly generated number i, j ($1 \le i < j \le n$), the operator will exchange vertices placed in p_i and p_j.

Let current test tube be T_1. The To-Single-Stranded operation is performed firstly, then, we cut ssDNA between p_i and v_i, between v_i and p_{i+1}, between p_j and v_j, between v_j and p_{j+1}, that is we cut down v_i and v_j placed in p_i and p_j. Reordered ssDNA (with very short double strand structure) will form after ligase, $\overline{p_{i+1}v_i'p_i}$ and $\overline{p_{j+1}v_j'p_j}$ are added. The operation completes after extracting proper permutation of vertices and performing To-Double-Stranded operation. We denote the function by OM(p_i,p_j) where p_i and p_j are two positions. The algorithm is outlined below (see Fig.4. also).

```
1.  $T_1 \leftarrow To-Single-Stranded(T_1)$
    For $v_i, v_j \in \{v_1, \cdots, v_n\}$ [In Parallel]
2.          $Cut(T_1, p_i \mid v_i)$          $Cut(T_1, v_i \mid p_{i+1})$
            $Cut(T_1, p_j \mid v_j)$          $Cut(T_1, v_j \mid p_{j+1})$
    End For
3.  $Add(T_1, \overline{p_{i+1}v_i'p_i}, \overline{p_{j+1}v_j'p_j}), Ligase(T_1)$
4.  $Extract(T_1, (v_1, v_2, \cdots, v_n)), To-Double-Strand(T_1)$
```

Block Mutation: The operator is used to translate blocks of successive vertices. This operation can be implemented by many order mutations. For example, for parent chromosome parent=$p_1v_1 \ldots p_iv_i \ldots p_jv_j \ldots p_nv_np_{n+1}$ and k=2, the aim is to transfer v_iv_{i+1} to

v_j behind, that is to get child chromosome child=$p_1v_1\ldots p_{i-1}v_{i-1}p_iv_{i+2}\ldots p_{j-2}v_jp_{j-1}v_ip_jv_{i+1}\ldots p_nv_np_{n+1}$. We can get it by the following operations, $v_i\leftrightarrow v_{i+2}$ to place v_{i+2} in the destination place, in the same way, $v_{i+1}\leftrightarrow v_{i+3}$, $v_{i+4}\leftrightarrow v_i\ldots v_{j-1}\leftrightarrow v_i$,$v_j\leftrightarrow v_{i+1}$. One thing to mention is that p_i ($1 \le i \le n$) does not change during the process. For general i, j, k, we have the following algorithm.

```
For r = 1 to k
        For s = 1 to $j - i$
                $OM(p_{i+s-1}, p_{i+s})$
        End For
End For
```

Color Spread Mutation and Bad Edge Stretch Mutation are both aimed to be carried on single chromosome separately, when implemented with DNA, we think there is not much meaning, so we do not consider them here. But since they are concerned with dealing with edges, it is a good place to show how to extract an edge. The extraction is a little difficult for our encoding for permutation contains position information of vertices. To extract an edge $e=v_iv_j$, we need n-1 extraction operations (Let current test tube be T_1).

```
1)  Perpare $n$ empty test tubes, Label $T_{1,2}, T_{1,3}, \cdots, T_{1,n-1}$ and $T_2$
2)  $T_1 \leftarrow To-Single-Stranded(T_1)$
    For s=2 to n-1
3)      $T_{1,s} \leftarrow Extract(T_1, v_i p_s v_j)$
        $T_2 \leftarrow Merge(T_2, T_{1,s})$
    End For
4)  $T_2 \leftarrow To-Double-Stranded(T_2)$
```

After performing operations above, tube T_2 consists of dsDNA containing bad edge $e=v_iv_j$, and we denote the operation by (T_2,T_1)←Extract(T_1,e) (a little different from the previous definition).

3.3.3 Implementation of Evaluation

Evaluation by Bad Edges: according to the theorem mentioned before, each stable set consists of successive vertices in the ordering, which means we can get the number of bad edges by accounting pairs of vertices connected by an edge are successive in an ordering. Our aim is to separate chromosomes by their fitness, for the rank of evaluation, we can choose |E|+1(0, 1,…, |E|), the number of bad edges in an ordering. Let current test tube be T, edge set be {e_1,…,e_m}. we test an edge one time. Performing (T_1,T_0)←Extract(T,e_1) firstly, the result is that T_1 consists of orderings containing bad edge e_1 and T_0 consists of the rest of T. Then testing for e_2 is carried out on T_0 and T_1 simultaneously, (T_{00},T_{01})←Extract(T_0,e_2) and (T_{10},T_{11})←Extract(T_1,e_2). Repeating the operations until all the edges has been tested. At end, we get 2^m test tube, $T_{e1e2\ldots em}$, where e1…em$\in\{0,1\}$and $\sum_{j=1}^{m} e_j$ means the number of bad edges in chromosomes in tube. The algorithm is outlined below (let current test tube be T).

```
1)  prepare $2^m$ empty test tubes, Labels $T_{e_1 e_2 \cdots e_m}, e_1, \cdots, e_m \in \{0,1\}$
    For i=1 to m
2)      $(T_{e_1 \cdots e_{i-1} 0}, T_{e_1 \cdots e_{i-1} 1}) \leftarrow Extract(T_{e_1 \cdots e_{i-1}}, e_i)$
    End For
    Prepare m+1 test tubes, Label with $W_k, k \in \{0, \cdots, m\}$, do the following
      For each $T_{e_1 \cdots e_m}$ [In Parallel]
3)        If $\sum_{j=1}^{m} e_j = k(e_j \in \{0,1\})$ ,$W_k \leftarrow \mathrm{Merge}(W_k, T_{e_1 \cdots e_m})$
      End For
```

We get k+1 DNA strands sets of different ranks when the algorithm completes. We use 2^m test tubes in step 2 and m+1 test tube in step 3, but, in fact, the number of test tubes can be reduced. For example, during the process of computing, some tube may become empty (through detect operation), then the extract operation for these tubes can be terminated. Besides, we can also control the degree of ranks of fitness. For example, during the first a few generations of the whole genetic algorithm, we can limit the fitness be k'(k'<m) since the difference of fitness is large and the number of chromosome of each rank is about the same, for chromosomes with bad edges number larger than k', we can simply discard them. Further, in step 3, we can reuse test tubes used in step 2. In this way, the number of test tubes used here can be greatly decreased.

3.4 Implementation of Selection and Note on the Completion of Algorithm

Selection is used to keep fit parent chromosomes in child generation and let less fit chromosomes to die. The implementation of selection is easy to achieve after the operation of evaluation.

For a given tube W_k ($0 \le k \le m$) obtained at the end of evaluation, the chromosomes in it have the same fitness rank. To let less fit candidates die, we can take a threshold k'(k'<m) and discard chromosomes in tube W_k where k>k'. To embody the difference of fitness among reserved chromosomes, we can perform different times of PCR for them, making the more fit chromosomes to breed more offspring.

Normal completion rules of genetic algorithm can be used in algorithms based on DNA also. For the problem whether a graph can be k-colored, the algorithm may complete after evaluating operation. If after the evaluating, some test tubes W_i ($0 \le i \le k-1$) are not empty, that is the number of bad edges in some ordering is less than k-1, then we can say that graph G can be k-colored and answer “yes”. Further, we can get concrete coloring schemes by decoding these DNA sequences. Another way to complete is that after a specific number of generations, we still can not say “yes”, than we say “no”.

3.5 Running Time Analysis

We take the operation of extracting as the most important criterion here and take the complexity of extracting n parts to be n. The initial generation of candidate data pool can be done in O(1) steps. To complete the crossover operation, the number of extracting operations required is O(n). Order mutation requires O(n) extracting operations in a similar way. The number of extracting operations required for block mutation is proportioned to the randomly generated k and the value of j subtracting i, so it is $O(n^3)$. Evaluation of candidate based on bad edges is proportional to O(mn). Selection can be done in O(1). So the total running time complexity to get a new generation is $O(n^3+mn)$.

As for space complexity, it is hard to estimate how many candidates are enough. But since it is implemented as a genetic algorithm, we can guess the number of enough candidates should be much less than the overall search space which is n!.

4 A Genetic Algorithm for Resolving the Maximal Clique Problem

The maximal clique problem is for a given graph G=(V,E), finding a maximal vertices set in which any two vertices are connected by an edge in G. The problem is also NP-complete. For a DNA solution to the problem, please see [12].

Adopting the same encoding scheme as [12], we encode a chromosome representing a possible clique as $p_1v_1p_2v_2\ldots p_nv_np_{n+1}$ where p_i is denoting position i and $vi\in\{0,1\}$, v_i equals 1 means that vertex i belongs to the clique, equals 0 means it does not belong to the clique. The number of 1 in a legal chromosome can be chosen as evaluation criterion, while legal means that two vertices connected by an edge in the complementary graph do not appear as 1 in a chromosome [12]. To get a legal chromosome, do the following operations for an edge $e=v_iv_j$ in a complementary graph G'.

Let the current test tube be T_0, $(T_0,T_1)\leftarrow Extract(T_0,p_i1)$. After the operation, T_0 consists of DNA strands containing p_i1 and T_1 consists of strands containing p_i0. In the same way, $(T_0,T_2)\leftarrow Extract(T_0,p_j1)$, T_2 consists of strands containing $\overline{p_i1}$ and p_j0. $T\leftarrow Merge(T_1,T_2)$, then T consists of strands does not contain p_i1 and p_j1 at the same time, that is, strands that do not contain $e=v_iv_j$.

5 Conclusion and Discussion

We present a primary implementation of a genetic algorithm proposed by Croitoru et al. with some simplifying in this paper. Their experiments showed that the algorithm is efficient for graph coloring problem. The implementation is based on available operations, but for the imperfect of operations, the feasibility of this implementation need to be further studied. Using DNA to implement genetic algorithm have a lot of advantages: high parallelism, high information density, the crossover operation is easy to implement and more tolerance to error. The paper shows further the possibility of implementing genetic algorithm on DNA.

Acknowledgement

The work has been supported by the National High-Tech. R&D Program for CIMS, China (No. 2003AA411021) and the National Natural Science Foundation, China (No. 60273056).

References

1. D.S. Johnson, C.R. Aragon, L.A. McGeoch, and C. Schevon. Optimization by simulated annealing: An experimental evaluation: Part II, graph coloring and number partitioning. Operations Research, 39(3):378–406, 1991
2. C. Croitoru, H. Luchian, O. Gheorghies, A. Apetrei. A New Genetic Graph Coloring Heuristic, COLOR02, Ithaca, NY
3. M. Amos, A. Gibbons. Error-resistant Implementation of DNA Computations. Proceedings of the Second Annual Meeting on DNA Based Computers, 1996, Vol.44, 151-168
4. E. Bach, A. Condon, E. Glaser, and C. Tanguay. DNA models and algorithms for NP-complete problems. In Proceedings of 11th Conference on Computational Complexity, 290-299. IEEE Computer Society Press, Los Alamitos, CA, 1996
5. L. Adleman. Molecular computation of solution to combinatorial problems. Science, 266:1021-1024, 1994
6. W. Cai, A. Condon, R. Corn, E. Glaser, Z.Fei, T. Frutos, Z. Guo, M. Lagally, Q. Liu, L. Smith, and A. Thiel. The power of surface-based DNA computation. In Proceedings of 1st International Conference on Computational Molecular Biology,67-74, ACM Press,1997
7. S. Diaz, J. L. Esteban, and M. Ogihara, A DNA-based random walk method for solving *k*-SAT. Proceedings of the Sixth International Workshop on DNA-based Computers, 209-220, Springer-Verlag Lecture Notes in Computer Science, 2001
8. W.P.C. Stemmer. The evolution of molecular computation. Science, 270:1510-1510, December 1, 1995
9. J. Chen, E. Antipov, B. Lemieux, W. Cedeno, and D. H. Wood. DNA computing implementing genetic algorithms. In L. F. Landweber, E. Winfree, R. Lipton, and S. Freeland, editors, Evolution as Computation, pages 39--49, New York, 1999. Springer Verlag
10. J. Rose, M. Takano and A. Suyama. A PNA-mediated Whiplash PCR-based Program for In Vitro Protein Evolution. Proceedings of the Eighth International Workshop on DNA-based Computers, 47-60, Springer-Verlag Lecture Notes in Computer Science, 2003
11. K. Chen, V Ramachandran. A Space Efficient Randomized DNA Algorithm. Proceedings of the Sixth International Workshop on DNA-based Computers,199-208, Springer-Verlag Lecture Notes in Computer Science, 2001
12. Q. Ouyang, P.D. Kaplan, S.Liu, A. Libechabe. DNA Solution of the Maximal Clique Problem. Science 1997,278,446-449

Studies on the Minimum Initial Marking of a Class of Hybrid Timed Petri Nets

Huaping Dai

State Key Lab of Industrial Control Technology, Zhejiang University,
Hangzhou 310027, P.R. China
hpdai@iipc.zju.edu.cn

Abstract. For the minimum initial marking (MIM) problem is one of minimum resource allocation problems, it is significant to study the MIM problem for a class of hybrid timed Petri nets, called a hybrid timed event graph (HTEG). An HTEG has additional continuous places and continuous transitions than a timed event graph (TEG). By the construction of a new dioid endowed with the pointwise minimum as addition and the composition of functions as multiplication, a linear min-plus algebraic model of HTEG was derived. Based on the min-plus algebra and its properties, the MIM problem for HTEG was studied in the text.

1 Introduction

Petri net is one of the most important modeling and analyzing methods in computer and control science. A timed event graph (TEG) is such a Petri net that each place has only (no more than) one input arc or one output arc. A linear model was derived for a TEG by using a max-plus algebra endowed with the maximum as addition and the conventional addition as multiplication [1].

But TEG can't describe quantitative relations, e.g., one frame and two wheels constructing a bicycle. As an extension of TEG, a timed event multigraph (TEMG) was introduced by [1]. A TEMG has an integer weight assigned to its each arc. The authors of [3] derived a max-plus algebraic model to study TEMG. Farther, a TEMG can't describe real quantities, e.g., 1.5 kilogram of rubber and one steel circle constructing a wheel. A wider extension of a TEMG, called a fluid timed event graphs with multipliers (FTEGM), was introduced by [2]. The multipliers and marking can take real values, i.e., an FTEGM has additional continuous places than a TEMG. Its min-plus algebraic model was derived in [2]. But it can't describe continuous events, e.g., a cool water flow at 10 °C and 0.6 litre/s mixed with a hot water flow at 90 °C and 0.4 litre/s making a warm water flow at 42 °C and 1.0 litre/s. A wider extension of an FTEGM, called a hybrid timed event graph (HTEG) will be presented in this paper. An HTEG has additional continuous transitions than an FTEGM.

The order of set containment among TEG, TEMG, FTEGM and HTEG is as follows: TEG $\subseteq$ TEMG $\subseteq$ FTEGM $\subseteq$ HTEG.

The minimum initial marking (MIM) problem is one of minimum resource allocation problems and is defined as follows: given a firing count vector X (with each

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 109 – 118, 2005.

component X (q) denoting the total firing number of a transition q), to find a minimum initial marking M_0 such that there is a firing sequence δ and each transition q appears exactly X (q) times in δ, the first transition is firable on M_0 and the rest can be fired one by one subsequently [5]. For an HTEG has time constraint, the MIM problem for HTEG is something different from [5]. This paper will propose a novel min-plus algebra based method to study the MIM problem for HTEG.

The outline of this paper is as follows: the definition of HTEG and enabling conditions of transitions will be given in Section 2, the min-plus algebraic model of HTEG will be discussed in Section 3, some properties of the min-plus algebra will be given in Section 4, the main results will be summarized in Section 5.

2 HTEG

Denote the real set by **R**, the nonnegative real set by $\mathbf{R}^+$, the nonnegative integer set by **N**. Hybrid Petri nets were first put forward by the authors of [6], a similar definition is given for HTEG as follows:

Definition 1: An HTEG is defined by a 7-tuplet = ⟨P, Q, R, W, *Tempo*, V, M_0⟩, where P is the set of places including a discrete place set (denoted by P_D) and a continuous place set (denoted by P_C); Q is the set of transitions including a discrete transition set (denoted by Q_D) and a continuous transition set (denoted by Q_C); R is the set of arcs, excluding the relations between Q_C and P_D, but including the hybrid relations $R_{Q_DP_C} \bigcup R_{P_CQ_D}$, $R \subseteq P \times Q \bigcup Q \times P$; W is the weight of R, W: $R_{Q_CP_C} \bigcup R_{Q_DP_C} \bigcup R_{P_CQ_C} \bigcup R_{P_CQ_D} \rightarrow \mathbf{R}^+$, $R_{Q_DP_D} \bigcup R_{P_DQ_D} \rightarrow \mathbf{N}$; V is the restricted velocity of continuous transitions, V: $\mathrm{Q}_\mathrm{C} \rightarrow \mathbf{R}^+$; *Tempo* is the time delay of discrete transitions, *Tempo:* $Q_D \rightarrow \mathbf{R}^+$; M_0 is the initial marking of places, M_0: $P_C \rightarrow \mathbf{R}^+$, $P_D \rightarrow \mathbf{N}$; An HTEG demands that each place has no more than one input arc or one output arc.

For the number of servers at a discrete transition can be obviously expressed [6], the number of servers is not defined in the 7-tuplet. If an enabled transition has one server, it can fire as soon as it finishes the last time. If an enabled transition has infinite servers, it can fire instantaneously. The number of servers at a continuous transition takes one as the default.

An HTEG can be drawn by a directed graph where a discrete place is represented by O , a continuous place by ◎ , a discrete transition by | , a continuous transition by, ▯ an arc by →.

Denote the marking of a place p by *Mark*(p), the input place set of a transition q by ${}^{\circ}q$, the output place set of q by q°, the input transition set of a place p by ${}^{\circ}p$, the output transition set of p by p°. Seeing Fig.1, ${}^{\circ}p = \{q'\}$, $p^{\circ} = \{q\}$.

Definition 2: The enabling conditions of a transition q is defined as the following cases:

(1) For a discrete transition $q \in Q_D$, if $\forall p \in {}^{\circ}q, Mark(p) \geq W(p,q)$, or if q has no input places, q is enabled. This case is the same as the traditional definition.

(2) For a continuous transition $q \in Q_C$,

(2.1) If $\forall p \in {}^{\circ}q, Mark(p) > 0$ or if q has no input places, q is enabled and the actual firing velocity equals $V(q)$.

(2.2) If $\exists p \in {}^{\circ}q, Mark(p) = 0$, q may be weakly-enabled. As shown in Fig.1, let $Q_0 = \{q' \in {}^{\circ}p \mid Mark(p) = 0, p \in {}^{\circ}q\} \cap Q_C$, the actual firing velocity of q, denoted by $V'(q)$, equals min{ $\{V'(q') \times W(q', p) / W(p,q) \mid q' \in Q_0\}$, $V(q)$}where $V'(q')$ is the actual velocity of q'. If there is a loop composed of empty places, those transitions in the loop are not enabled. If $\{q' \in {}^{\circ}p \mid Mark(p) = 0, p \in {}^{\circ}q\} \cap Q_D \neq \emptyset$, q is not enabled.

Remark 1. (2.1) For a continuous transition q, if its each input place is not empty, q fires at the highest velocity $V(q)$. (2.2) If one input place p of q is empty, the input transition q' of p should be considered. The actual firing velocity of q' determines the marking increase rate of p. The marking increase rate of p affects the firing velocity of q. Recursively consider the actual firing velocity of q'.

A discrete transition q can fire if q is enabled and q finishes the last time. The firing process will cost *Tempo*(q) units of time. A continuous transition q can run at the velocity $V(q)$ if q is enabled, or at the velocity $V'(q)$ if q is weakly-enabled. A fiing transition consumes some tokens in its each input places and produces some tokens in its each output places. The token change is like the traditional case.

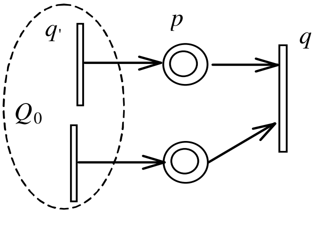


Fig. 1. The weakly-enabling case

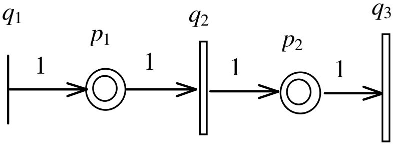


Fig. 2. A simple HTEG example

As shown in Fig. 2, a simple HTEG example demonstrates the enabling conditions of transitions. Given $M_0(p_1)= M_0(p_2) = 0$, $Tempo(q_1)=0.4$, $V(q_2)=2$, $V(q_3)=3$. q_1 is always enabled. When $t = 0$, q_2 and q_3 are not enabled; when $t \geqslant 0.4$, q_2 is enabled, q_3 is weakly enabled. $V'(q_3) = \min\{V(q_2), V(q_3)\} = \min\{2, 3\} =2$.

3 Min-plus Algebraic Model of HTEG

First we define counter variable associated with each place and each transition. With a place p, a counter variable $M(p,t)$ is associated. It denotes the cumulated number of tokens which have entered into p from instant 0 to instant t. Seeing Fig.2,

$M(p_1,t) = \lfloor t/0.4 \rfloor$ where $\lfloor x \rfloor = \sup\{n\in N \mid n \le x\}$,

$$M(p_2,t) = \begin{cases} 0, t < 0.4 \\ 2(t-0.4), t \ge 0.4 \end{cases}.$$

With a discrete transition $q \in Q_D$, a counter variable $M(q,t)$ is associated. It denotes the cumulated firing times of q from instant 0 to instant t. It is an integer. Seeing Fig.2, $M(q_1,t) = \lfloor t/0.4 \rfloor$+1.

For a continuous transition $q \in Q_C$, q is discretized by using a sampling period η, then a counter variable $M(q,t)$ is associated with the discretized transition q. It also represents the cumulated firing times of q from instant 0 to instant t. It may be a non-negative real. This case is like the numerical computing method for differential equations.

With the independent variable time t, $M(\cdot,t)$ has three meanings for a place, a discrete transition and a continuous transition. $M(\cdot,t) \ge 0$. Suppose that the initial instant = 0, $M(\cdot,t) = 0$ if $t < 0$.

Now consider two cases: (1) a discrete transition case and (2) a continuous transition case.

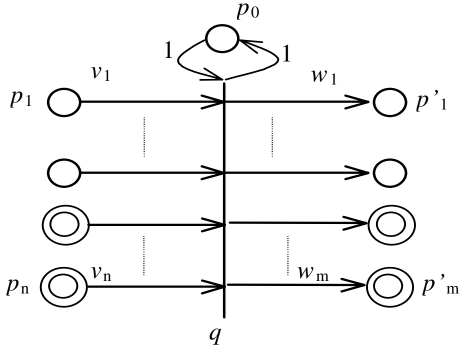


Fig. 3. A discrete transition case

(1) Consider a discrete transition q and the token increase of its output places. As shown in Fig.3, given ${}^{\circ}q=\{p_0,p_1,...,p_n\}$ and $q^{\circ}=\{p'_0,...,p'_m\}$ where $p'_0=p_0$, ${}^{\circ}q$ and q° maybe include discrete and continuous places, *Tempo*(q) = d, $W(p_{\mathrm{i}}, q)$ = v_{i}, $W(q, p'_j)$ = w_{j}, v_0 = w_0 =1, $i=0,1,...,n$, $j=0,1,...,m$. $M_0(p_0)$ represents the number of servers at q.

Proposition 1: $M(q,t)$ = min{ $\lfloor M(p_i,t)/v_i\rfloor, i=0,1,2,...,n$ }. The cumulated number of tokens of the output place p'_j of q, $M(p'_j,t)=M_0(p'_j)+M(q,t-d)\times w_j$, where M_0(p'_j) denotes the initial marking of p'_j. After eliminating $M(q,t)$, we have

$$M(p'_j,t)=M_0(p'_j)+w_j\times\sum_{i=0}^{n}{}_{\oplus}\lfloor M(p_i,t-d)/v_i\rfloor \cdot \qquad (1)$$

where $\oplus$ denotes the minimum.

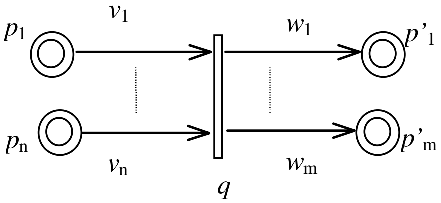


Fig. 4. A continuous transition case

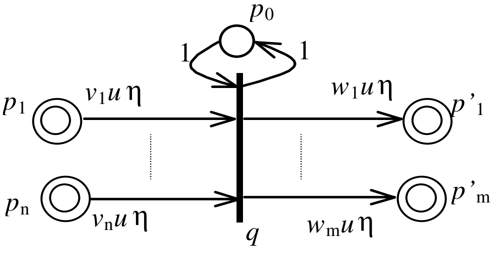


Fig. 5. A discretized transition of Fig.4.

(2) Consider a continuous transition q and the token increase of its output places. As shown in Fig.4, given ${}^{\circ}q=\{p_1,...,p_n\}$ and $q^{\circ}=\{p'_1,...,p'_m\}$ which include only continuous places, $W(p_{\mathrm{i}}, q)$ = v_{i}, $W(q, p'_j)$ = w_{j}, $i=1,2,...,n$, $j=1,2,...,m$.. The velocity of q is denoted by u for the sake of brevity, i.e., $V(q)$ = u.

After q is discretized, Fig.5 is gotten where *Tempo*(q) $=\eta$, ${}^{\circ}q=\{p_0,p_1,...,p_n\}$, $q^{\circ}=\{p_0,p'_1,...,p'_m\}$, $W(p_i, q)$ = $v_i u\eta$, $W(q, p'_j)$ = $w_j u\eta$, $i=1,2,...,n$, $j=1,2,...,m$, $W(p_0, q)=W(q, p_0)=1$, $M_0(p_0)=1$ as the default, i.e., q has one server.

Proposition 2: $M(q,t)=$ min{ $M(p_i,t)/W(p_i,q), i=0,1,2,...,n$ }. The cumulated tokens of p'_j, $M(p'_j,t)=M_0(p'_j)+M(q,t-\eta)\times W(q,p'_j)$, $j=0,1,...,m$. After $M(q,t)$ and p_0 are eliminated,

$$\begin{aligned}&[M(p'_j,t)-M_0(p'_j)]\\&=([M(p'_j,t-\eta)-M_0(p'_j)]+w_ju\eta)\oplus w_j\times\sum_{i=1}^{n}{}_{\oplus}[M(P_i,t\text{-}\eta)/v_i]\end{aligned}\,. \tag{2}$$

Remark 2. (1) η may be taken as a sampling period. (2)When an appropriate value is assigned to η, the equation is approximately computed. When η is infinitesimal, Equation(2) is transformed to a differential equation. (3) η is small enough that the fraction of $M(q,t)$ can be neglected. So Equation(2) does not have the function $\lfloor x\rfloor$ while Equation(1) does.

A formal definition of min-plus algebra is stated as follows.

Since Equation(1) and (2) do not have $M(q,t)$, the system variables only include $M(p,t)$. For the sake of brevity, $M(p,t)$ is denoted by $s(t)$, $s(t)\geq 0$. $s(t)$ is a nondecreasing curve. All curves like $s(t)$ form a set, denoted by S. A minimum operation $\oplus$ in S is defined as follows: $\forall s_1,s_2\in S$, $(s_1\oplus s_2)(t)=s_1(t)\oplus s_2(t)$ =min{ $s_1(t)$, $s_2(t)$ }. Define $\varepsilon(t)=+\infty$, denoted by ε for the sake of brevity, let $\varepsilon\in S$. ε is a special element.

Consider Equation(1) and (2), the coefficients of $M(p,t)$ are taken as functions on S. Four families of functions are defined as follows, $x\in$ $\mathbf{R}^+$, $s(t)\in S$, $\circ$ is an operator between the defined function and $s(t)$,

Definition 3: (1)A counting shifting function U^x : $U^x\circ s(t)=s(t)+x$,

(2)A time shifting function Z^{-x} : $Z^{-x}\circ s(t)=s(t-x)$,

(3)A scaling function K^x : $K^x\circ s(t)=x\times s(t)$,

(4)A Gaussian function I : $I\circ s(t)=\lfloor s(t)\rfloor$.

Let G = { $U^x\mid x\in R^+$ } $\bigcup$ { $Z^{-x}\mid x\in R^+$ } $\bigcup$ { $K^x\mid x\in R^+$ } $\bigcup$ { I }. For these four families of functions are lower-semicontinuous(defined in [1]), we have

Proposition 3: $< G; \oplus, \otimes >$ can be induced into a dioid by using the following definition: $\forall f, g \in G, s \in S$, $(f \oplus g) \circ s(t) = (f \circ s(t)) \oplus (g \circ s(t))$, $(f \otimes g) \circ s(t) = f \circ (g \circ s(t))$.

The four families of functions were first defined in [2]. The set containing *G* and the composition of functions with $\oplus$ and $\otimes$ is denoted by $\overline{G}$. $< \overline{G}; \oplus, \otimes >$ is a dioid. For the sake of brevity, $\overline{G}$ is also denoted by *G*. $< G; \oplus, \otimes >$ is the needed min-plus algebra.

Theorem 1. The dynamics of an HTEG can be represented by $< G; \oplus, \otimes >$.

Proof: Equation(1) can be rewritten as

$$M(p'_j, t) = U^{M_0(p'_j)} \circ K^{w_j} \circ \sum_{i=0}^{n}{}_{\oplus} (I \circ K^{1/v_i} \circ Z^{-d} \circ M(p_i, t)). \tag{3}$$

Equation(2) can be rewritten as

$$[M(p'_j, t) - M_0(p'_j)] = K^{w_j u \eta} \circ Z^{-\eta} \circ \\ ([M(p'_j, t) - M_0(p'_j)]) \oplus K^{w_j} \circ \sum_{i=1}^{n}{}_{\oplus} (K^{1/v_i} \circ Z^{-\eta} \circ M(P_i, t)). \tag{4}$$

Those system variables $M(p, t)$ of the places without input transitions are taken as being known and stacked as a known vector *b*, the other system variables are stacked as an unknown vector *x*, the coefficients of system variables are ranked as a known matrix *A*, those equations like(3) and (4) are transformed into

$$x = Ax \oplus b. \tag{5}$$

The solution to Equation(5) is

$$x = A^* b . \tag{6}$$

where $A^* = \sum_{i=0}^{\infty} {}_{\oplus} A^i$.

Now consider a simple example shown in Fig.6. Given that *Tempo*(q_1)=1, *V*(q_2)=2, *Tempo*(q_3)=3, $M_0(p_1)$=0, $M_0(p_2)$=0, $M_0(p_3)$=12, the transition q_3 has infinite servers, i.e., q_3 can fire instantaneously, the other transitions have only one server each. After q_2 is discretized, Fig.6 is transformed to Fig.7 where *Tempo*(q_2) $= \eta$. Notice that $M_0(p_2)$+ $M_0(p_3)$ equals the capacity limit of p_2.

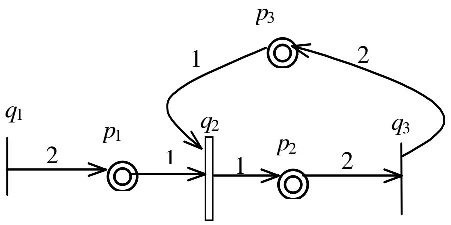


Fig. 6. A simple HTEG example

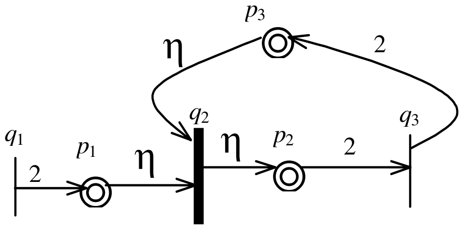


Fig. 7. After q_2 is discretized

The following three equations are gotten,

$$M(p_1,t) = M(q_1,t-1)+0 = 2\lfloor t \rfloor \ . \tag{7}$$

$$M(p_3,t) = 2\lfloor M(p_2,t-3)/2 \rfloor + 12 \ . \tag{8}$$

$$M(p_2,t) = (M(p_2,t-\eta)+2\eta) \oplus M(p_1,t-\eta) \oplus M(p_3,t-\eta) \ . \tag{9}$$

It is computed that $M(p_2,t) = a(t)$ where $a(t)$ □ $\begin{cases} 2(t-1), t \geq 1 \\ 0, t < 1 \end{cases}$.

4 Properties of $< G; \oplus, \otimes >$

Some properties of $< G; \oplus, \otimes >$ are mentioned bellow:

Property 1: (1)The identity element of $< G; \otimes >$ is $U^0 = Z^0 = K^1$, denoted by $1_\otimes$.

(2) $\forall x \in$ **R**, the inverse of U^x is U^{-x}, the inverse of Z^{-x} is Z^x, the inverse of K^x is $K^{1/x}$ if $x \neq 0$.

(3) $\forall x, y \in$ **R,** $U^x \otimes U^y = U^{x+y}$, $Z^{-x} \otimes Z^{-y} = Z^{-(x+y)}$, $K^x \otimes K^y = K^{xy}$.

(4) U^x is commutative with Z^{-x}, K^x is commutative with Z^{-x}. K^x is not commutative with U^x, but $K^y \circ U^x = U^{xy} \circ K^y$.

(5) I is not commutative with U^x and K^x, but I is idempotent, i.e., $I \circ I = I$.

(6) ε is the zero element of $< G; \otimes >$, the identity element of $< G; \oplus >$.

For a curve $x(t) \in S$, for any instant $t, x(t) \in$ **R**$^+$, therefore S is an ordered set like **R**. Similarly, G is also an ordered set as follows: $x, y \in G, x \ge y \Leftrightarrow \forall s \in S, x \circ s \ge y \circ s$.

Property 2: $\forall x, y \in$ **R**$^+$, (1)$U^x \ge U^y \Leftrightarrow x \ge y$; (2) $K^x \ge K^y \Leftrightarrow x \ge y$.

Remark 3. (1) $\forall s \in S$, $U^x \circ s(t) = s(t) + x$, $U^y \circ s(t) = s(t) + y$, $s(t) + x \ge s(t) + y \Leftrightarrow x \ge y$, therefore the inequation(1) holds.

(2) $\forall s \in S$, by the definition of $s(t)$, it is known that $s(t) \ge 0$, $K^x \circ s(t) = xs(t)$, $K^y \circ s(t) = ys(t)$, therefore $xs(t) \ge ys(t) \Leftrightarrow x \ge y$, the inequation(2) holds.

Property 3: $\forall x, y, z \in S$, $\forall a \in$ **R**, $\forall b \in$ **R**$^+$, $b \ne 0$, the following four equations hold,

(1) $y \oplus x \ge z \Leftrightarrow x \ge z$ and $y \ge z$;

(2)$U^a \circ y \ge z \Leftrightarrow y \ge U^{-a} \circ z$;

(3)$K^b \circ y \ge z \Leftrightarrow y \ge K^{1/b} \circ z$;

(4)$I \circ y \ge z \Rightarrow y \ge z$.

Remark 4. (1) For the minimum of x and y is larger than z, both x and y is larger than z; (2) holds because U^a is inverse; (3) holds because K^b is inverse; (4) holds because $y \ge I \circ y$.

Definition 4: Consider a map f, f is monotone if and only if $\forall x, y \in S$, $x \ge y \Rightarrow f(x) \ge f(y)$.

Theorem 2: Obviously, $\forall f \in \overline{G}$, f is monotone.

Remark 5. Since each atomic function of G is monotone, and the composition of monotone functions is monotone, therefore, **Theorem** 2 holds.

Consider the example in Fig.6 again. Here let us compute the minimum initial marking of p_3 such that q_2 can fire fastest. If $M_0(p_3) = +\infty$, q_2 can fire fastest and $M(p_2, t) = a(t)$. Equation(9) implies that $M(p_2, t) \le M(p_3, t)$, i.e., $a(t) \le 2\lfloor t - 4 \rfloor + M_0(p_3)$. Then, $M_0(p_3) \ge 8$, i.e., the minimum of $M_0(p_3)$ is 8.

As shown in Fig.6, p_3 is a dual place of p_2, $M_0(p_2)$+ $M_0(p_3)$ represents the capacity of p_2, this example gave a solution to the minimum capacity of p_2. Notice that the result does not relate to the variable η. For the transition of $^\circ p$ (e.g. $^\circ p_2 = \{q_2\}$) may be continuous and its firing counter is meaningless, so the MIM problem considered here gave a known cumulated tokens into a place *p*(e.g. $M(p_2, t)$).

5 Conclusions

As a rational extension of a TEG, an HTEG is such a proper class of hybrid timed Petri nets that it has continuous weights, places and transitions to model quantitative relations and continuous events. It is lucky that a min-plus algebra, denoted by $< G; \oplus, \otimes >$, was derived to study an HTEG. Until now we do not know whether a more general hybrid timed Petri net than an HTEG can have a max-plus algebraic representation.

Intuitively, the system described by an HTEG is a dynamical system with time variables for discrete transitions and speed variables for continuous transitions. An HTEG may describe complex synchronization relations among continuous and discrete events occurring in manufacturing systems, computer network, transportation systems and other man-made systems.

The min-plus algebra $< G; \oplus, \otimes >$ is endowed with two operators: $\oplus$ and $\otimes$. $\oplus$ represents the minimum and $\otimes$ represents the composition of functions. G and S are ordered sets like the real set **R**. Since the initial marking of a place of an HTEG was represented by a parameter in Equation(5), the MIM problem for an HTEG was studied by using the properties of $< G; \oplus, \otimes >$. For the initial marking represents the configuration of a system including initial resources, buffer capacity and the number of servers, etc., the MIM problem can be taken as one of minimum resource allocation problems.

Acknowledgement

This work is supported by the Chinese National Natural Science Foundation under grant 60304018.

References

1. Baccelli, F., Cohen, G., Olsder, G.J., Quadrat, J.P.: Synchronization and Linearity: An Algebra for Discrete Event Systems. Wiley, New York(1992)
2. Cohen, G., Gaubert, S., Quadrat, J.P.: Timed-events graphs with multipliers and homogeneous Min-plus systems. IEEE Trans. on Automat. Contr., 43(1998)1296-1302
3. Huaping, D., Youxian, S.: An Algebraic Model for Performance Evaluation of Timed Event Multigraphs[J], IEEE Trans. On Automat. Contr., 48(2003)1227-1230
4. Cofer, D.D., Garg, V.K.: Supervisory Control of Real-Time Discrete-Event Systems Using Lattice Theory, IEEE Trans. on Automat. Contr., 41(1996)199-209
5. Nishi, S., Taoka, S., Watanabe, T.: A new heuristic method for solving the minimum initial marking problem of Petri nets. 2000 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 5. (2000)3218–3223
6. David, R., Alla, H.: Petri Nets Grafcet Tools for modeling discrete event systems. Hermes, France(1992)

A Fuzzy Neural Network System Based on Generalized Class Cover and Particle Swarm Optimization

Yanxin Huang, Yan Wang, Wengang Zhou, Zhezhou Yu, and Chunguang Zhou

College of Computer Science and Technology, Jilin University, Key Laboratory for Symbol
Computation and Knowledge Engineering of the National Education Ministry,
Changchun 130012, China
wy666@tom.com, zhouwengang@email.jlu.edu.cn
{huangyx, yuzz, cgzhou}@jlu.edu.cn

Abstract. A voting-mechanism-based fuzzy neural network system is proposed in this paper. When constructing the network structure, a generalized class cover problem is presented and its two solving algorithm, an improved greedy algorithm and a binary particle swarm optimization algorithm, are proposed to get the class covers with relatively even radii, which are used to partition fuzzy input space and extract fewer robust fuzzy IF-THEN rules. Meanwhile, a weighted Mamdani inference mechanism is adopted to improve the efficiency of the system output and a real-valued particle swarm optimization-based algorithm is used to refine the system parameters. Experimental results show that the system is feasible and effective.

1 Introduction

The first step in designing a fuzzy inference system is partition of input space. A good partition method can implement a small rule base with robust rules[1]. Adam Cannon and Lenore Cowen[2] presented the class cover problem (CCP) firstly, and proved the CCP is the NP-hard. Based on the CCP, a generalized class cover problem (GCCP) and its two solution algorithm, an improved greedy algorithm and a binary particle swarm optimization (bPSO) algorithm, which are used to partition input space and extract fewer robust fuzzy IF-THEN rules, are proposed in this paper. Then a voting-mechanism-based fuzzy neural network based on the obtained fuzzy IF-THEN rules and the real-valued particle swarm optimization (PSO) is constructed. Experimental results identifying 11 kinds of mineral waters by its taste signals show that the system is the feasible and effective.

2 Generalized Class Cover Problem

Adam Cannon and Lenore Cowen defined the CCP as follows[2]: Let B be the set of points in class one, and R be the set of points in class two, with |R|+|B|=n. Then the CCP is:

Minimize K

$$s.t. \max_{v\in B}\{d(v,S)\} < \min_{w\in R}\{d(w,S)\}$$

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 119 – 128, 2005.

where $S \subseteq B$,$|S|=K$, $d(.,.)$ denotes the distance between two points or a point to a set. Carey E. Priebe *et al*[3] presented a greedy algorithm (we refer to it as the original greedy algorithm below) for the CCP.

Aiming at designing the fuzzy neural network system architecture, we propose the generalized class cover problem (GCCP) as follows:

Minimize *K, Var*

$$s.t. \begin{cases} r_\beta(v,R) \le \gamma MaxD_\beta, \forall v \in S \\ |\{w \mid w \in B, and\ \exists v \in S, d(v,w) \le r_\beta(v,R)\}| \ge \alpha |B| \end{cases}$$

where $S \subseteq B$, $|S|=K$, $Var = \sqrt{\frac{1}{|S|-1}\sum_{v\in S}(r_\beta(v,R)-\bar{r})^2}$, $\bar{r} = \frac{1}{|S|}\sum_{v\in S} r_\beta(v,R)$, $\gamma,\alpha,\beta \in [0,1]$, $r_\beta(v,R)$ denotes the cover radius centered on the point $v \in S$, and *MaxD*$_\beta$ denotes the most cover radius in the points in B. The parameter γ is used to control producing the class covers with relatively even radii, which is beneficial to extracting the robust fuzzy IF-THEN rules, while the parameters α and β are used to make the system more noise-resistant, in which, α indicates that at least $\alpha|B|$ points in B must be covered by the class covers produced from *S*, and β indicates that a class cover centered on a point in B is permitted to cover at most $\beta|R|$ points in *R*. So, the GCCP is to find a minimum cardinality set of covering balls, with center points in *S* and relatively radii, whose union contains at least $\alpha|B|$ points in B and each covering ball contains at most $\beta|R|$ points in *R*. The CCP can be regarded as a special case of the GCCP with $\gamma=1, \alpha=1, \beta=0$.

3 Improved Greedy Algorithm

Toward the GCCP, we propose an improved greedy algorithm as follows.

Let $S=\phi$, and $C=B$:

(1) $\forall x \in B$, computing $d_\beta(x,R)$, which equals to the $\beta|R|+1$th smallest distance from *x* to the points in *R*, Let $MaxD_\beta = \max_{x\in B}\{d_\beta(x,R)\}$, then $\forall x \in B$ computing $r_\beta(x,R)$ as: $\begin{cases} r_\beta(x,R) = \gamma MaxD_\beta, if\ d_\beta(x,R) > \gamma MaxD_\beta \\ r_\beta(x,R) = d_\beta(x,R), if\ d_\beta(x,R) \le \gamma MaxD_\beta \end{cases}$;

(2)producing digraph $G=(B,E)$ as follows: $\forall x \in B$, for all $y \in B$, if $d(x,y) \le r_\beta(x,R)$, then producing a edge from x to y , namely $(x,y) \in E$;

(3) $\forall x \in C$, computing $\mathrm{cov}er(x) = \{y \mid y \in C, and\ (x,y) \in E\}$;

(4)taking $z \in C$, and $\mathrm{cov}er(z) = \max_{x\in C}\{\mathrm{cov}er(x)\}$, if $|C| < (1-\alpha)|B|$, then output S, and end the algorithm, else Let $S = S \cup \{z\}$, $C = C - \{x \mid x \in C, (z,x) \in E\}$, and go to (3)Let $|B|=N$,$|R|=M$, and $|S|=\rho(S)$, then the algorithm has the time complexity of $O(NM+N(N-1)+N\rho(S))$. Suppose $N\approx M$, and $\rho(S)<<N$, then the algorithm has the time complexity of $O(N^2)$ approximately.

4 bPSO Algorithm

J. Kennedy et al. invented the real-valued particle swarm optimization (PSO) model in 1995 [4]. Now, the real-valued PSO demonstrates good performance in many optimization problems[5]. In 1997, J. Kennedy et al. presented the binary particle swarm optimization (bPSO) model for solving discrete optimization problem[6], but the bPSO model still needs further research till now[5]. The improved greedy algorithm runs at a faster speed, which is its strongest character. However, it is hard to get best optimal solutions. Therefore we propose a bPSO algorithm for the GCCP in this section.

4.1 Data Pretreatment

In order to meet the requirement of our following bPSO steps, the data are pretreated firstly. The data points in R are stored in the array of $Array\_R$. And the data points in B are stored in the array of $Array\_B$ with size of *N*, where the nearer data points in the Euclid space in B are stored as nearer as possible in $Array\_B$. After the data pretreatment procedure, the class cover relational matrix of B relative to R is constructed as follows.

(1) Produce a digraph $G=(B,E)$ as same as the improved greedy algorithm, and store the cover radii of the points in $Array\_B$ into an array with size of *N* accordingly.

(2) Based on the digraph $G=(B,E)$, the 0-1 class cover relational matrix M_G of B relative to R with size of $N\times N$ is calculated, where $M_G(i,j)=1$ or 0 indicates there is a edge from *i* to *j* or not, $i,j\in[1,2,...,N]$.

4.2 Binary Encoding and Fitness Function of Particles

Let *L* be the population size of the particle swarm, then a particle I_k can be encoded as $I_k=b_1^{(k)}b_2^{(k)}...b_N^{(k)}$, where $b_i^{(k)}=1$ or 0 indicates the point $Array\_B[i]$ is included in *S* or not, $i\in[1,2,...,N]$, $k\in[1,2,...,L]$.

According to the definition of GCCP, the fitness of the particle I_k mainly depends on three factors: (1) the cardinality $C(I_k)$ of *S* corresponding to the particle I_k, where $C(I_k)=b_1^{(k)}+b_2^{(k)}+...+b_N^{(k)}$; (2) the number $R(I_k)$, which indicates the number of points in B which are covered by the class covers corresponding to particle I_k. Especially, when $R(I_k)\geq\alpha|B|$, which satisfies the constrained conditions of GCCP, a reward should be given to particle I_k; (3) the sample standard deviation $Var(I_k)$ of the class cover radii corresponding to particle I_k.

According to these factors mentioned above, we define the fitness of the particle I_k as:

$$F(I_k)=\begin{cases}\lambda_1\times\frac{N-C(I_k)}{N}+\lambda_2\times\frac{R(I_k)}{N}+\lambda_3\times\frac{Q-Var(I_k)}{Q}, if\ \ R(I_k)<\alpha N\\ \lambda_1\times\frac{N-C(I_k)}{N}+\lambda_2\times\frac{R(I_k)}{N}+\lambda_3\times\frac{Q-Var(I_k)}{Q}+0.2, if\ \ R(I_k)\ge\alpha N\end{cases}\qquad(1)$$

where λ_1 , λ_2 , and λ_3 are weight coefficients, which denote the importance of the three factors on scaling the fitness of a particle. Q is the least upper bound of the function $Var(I_k)$ which is calculated by the following theorem 4.1.

Theorem 4.1:

Let $a_1,a_2,...,a_n$ be a sequence of real numbers, and $a_{i_1},a_{i_2},...,a_{i_k}$ be its any subsequence, $i_1,i_2,...,i_k\in\{1,2,...,n\}$, $2\le k\le n$, then $Var(a_{i_1},a_{i_2},...,a_{i_k})\le Var(a_{\min},a_{\max})$, where $Var(a_{i_1},a_{i_2},...,a_{i_k})=\sqrt{\frac{1}{k-1}\sum_{j=1}^{k}a_{i_j}-\bar{a}_k)^2}$, $\bar{a}_k=\frac{1}{k}\sum_{j}^{1}a_{i_j}$, $a_{\min}=\min\{a_1,a_2,...,a_n\}$, $a_{max}=\max\{a_1,a_2,...,a_n\}$.

Proof:

Let $f(a_{i_1},a_{i_2},...,a_{i_k})=Var(a_{i_1},a_{i_2},...,a_{i_k})^2=\frac{1}{k-1}\sum_{j=1}^{k}(a_{i_j}-\bar{a}_k)^2$, then calculate $\frac{\partial f}{\partial a_{i_j}}=0$, where $a_{\min}\le a_{i_j}\le a_{\max}$, $j=1,2,...,k$, we know that the minimum of function f is zero at $a_{i_1}=a_{i_2}=...=a_{i_k}$ within the closed region: $a_{\min}\le a_{i_j}\le a_{\max}$, $j=1,2,...,k$, $2\le k\le n$. Next we calculate the function value of f on the boundary of the closed region. Without losing generality, suppose the subsequence $a_{i_1},a_{i_2},...,a_{i_k}$ has p a_{max} and k-p $a_{\min}$ $p\in Z$, $1\le p\le k-1$, then $f=\frac{1}{k-1}\frac{1}{k^2}.(a_{\max}-a_{\min})^2\ \ .((k-p)^2p+p^2(k-p))$, regarding p as an independent variable, we know that the maximum of function f is gotten in $p=\frac{k}{2}$, and the function f is a unimodal function. Because p is an integer, when k is an even number, we obtain $f=\frac{1}{k-1}\frac{1}{k^2}.(a_{\max}-a_{\min})^2.((k-\frac{k}{2})^2\frac{k}{2}+(\frac{k}{2})^2(k-\frac{k}{2}))\ =\frac{1}{4(k-1)}(a_{\max}-a_{\min})^2$, and because $k\ge2$, we get $f\le\frac{1}{2}(a_{\max}-a_{\min})^2\ =Var(a_{\max}-a_{\min})^2$. When k is an odd number, we obtain $f=\ \frac{1}{k-1}\frac{1}{k^2}.(a_{\max}-a_{\min})^2.((k-\frac{k-1}{2})^2\frac{k-1}{2}\ +(\frac{k-1}{2})^2(k-\frac{k-1}{2}))$ $=\frac{k+1}{4k}(a_{\max}-a_{\min})^2$, and because $k\ge2$, we get $f\le\frac{1}{2}(a_{\max}-a_{\min})^2$ $=Var(a_{\max}-a_{\min})^2$. Sum up, we come to the conclusion that $Var(a_{i_1},a_{i_2},...,a_{i_k})\le Var(a_{\min},a_{\max})$. The proof is completed.

According to theorem 4.1, we can set $Q = Var(R_{\min}, R_{\max})$, such that $0 \le \frac{Q - Var(I_k)}{Q} \le 1$, where $R_{\min}$ and $R_{\max}$ denote the least radius and the most radius in the class covers obtained from S respectively.

4.3 Recursive Equations of bPSO

Let L be the population size of particle swarm (generally set L=20[5]), then each particle presents a candidate solution in the search space. Each particle has four state variables: $\vec{v}(t)$, $\vec{x}(t)$, $\vec{x}^{(p)}(t)$ and $\vec{x}^{(g)}(t)$, which present its current velocity, current position, previous best position and the best position of all the particles, respectively. The velocity and position of the ith particle are updated with the following formulae:[6]

$$v_{ij}(t+1) = wv_{ij}\ (t)\ +\ c_1 r_{1j}(t)(\ x_{ij}{}^{(p)}(t)\ \text{-}\ x_{ij}\ (t))\ +\ c_2 r_{2j}(t)(x_j{}^{(g)}(t)\ \text{-}\ x_{ij}\ (t)) \tag{2}$$

$$x_{ij}(t+1) = \begin{cases} 0, if & \rho \ge Sig(v_{ij}(t+1)) \\ 1, if & \rho < Sig(v_{ij}(t+1)) \end{cases} \tag{3}$$

where $i = 1,2,...,L$; $j \in \{1,2,...,N\}$, j represents the jth element of N-dimensional vector, $r_{1j}(t) \sim U[0,1], r_{2j}(t) \sim U[0,1]$, w, c_1 and c_2 are *acceleration coefficients*, $\rho \sim U[0,1]$, and $Sig(x) = 1/(1+\exp(-x))$ [6]. When solving the GCCP with the bPSO algorithm, the position of one particle is initialized by the solution gotten from the improved greedy algorithm, and all the other particles are initialized randomly, which means a bit of binary code of a particle takes 1 with 0.2 probability, and 0 with 0.8 probability. The initial velocities of the particles are set by random numbers uniformly distributed on [-0.4, +0.4]. In order to enhance local search near the best position of all the particles, the parameters of the recursive equation of particles is determined as follows: $w = 1$, $c_1 = 0.9$, $c_2 = 1.2$. The weight coefficients are set for $\lambda_1 = 0.55$, $\lambda_2 = 0.1$, $\lambda_3 = 0.15$.

5 Voting-Mechanism-Based Fuzzy Neural Network Model

5.1 Fuzzy Neural Network Architecture

By projecting the class covers onto each input coordinate axis, the fuzzy subsets (linguistic terms), which are used to produce the initial fuzzy IF-THEN rules, can be obtained[1]. The Gauss function defined as Eq. (4) is used as the membership functions in this paper.

$$\mu = \exp(-(x-c)^2/\sigma^2) \tag{4}$$

Based on the fact that an object with unknown class tag is generally close to those samples whose class tags are the same with the one while far from the samples that have different class tags, a voting-mechanism-based fuzzy neural network system (VMFNN) is proposed as follows.

Different fuzzy inference mechanism can be distinguished by the consequents of the fuzzy IF-THEN rules[7], such as Mamdani and Takagi-Sugeno inference system. The Mamdani inference mechanism is adopted in this paper. Let $v^{(1)}, v^{(2)}, ..., v^{(r)}$ be the class tags of the training samples. Assume the number of the fuzzy IF-THEN rules with the consequent $y\ is\ v^{(j)}$ is a_j, $j = 1, 2, ..., r$, then those fuzzy rules can be represented as:

$$R_{i,j}: if\ x_1\ is\ A_{1,i}\ and\ x_2\ is\ A_{2,i}\ and...and\ x_s\ is\ A_{s,i}\ then\ y\ is\ v^{(j)}$$

where $(\overline{x}, y)$ is a training sample, and $\overline{x} = (x_1, x_2, ...x_s)$ is the input feature vector and y is its class tag, $i = 1, 2, ..., a_j$, $j = 1, 2, ..., r$. The matching degree of the fuzzy rule antecedent for the training sample $\overline{x}$ is computed by the multiplication T-norm[8] as:

$$A_T(R_{i,j}) = \prod_{k=1}^{s} A_{k,i}(x_k) \tag{5}$$

Then a subsystem S_j can be constructed using all $R_{i,j}$, $i = 1, 2, ..., a_j$. Its output is defined as:

$$O_j = 1 - \exp(-\sum_{i=1}^{a_j} A_T(R_{i,j})) \tag{6}$$

Finally, the output of the fuzzy neural network system is defined as:

$$Y = v^{(j)},\ \text{Subject to}\ O_j = \max(O_1, O_2, ..., O_r) \tag{7}$$

According to the above, the VMFNN model is derived as Fig.1. Note that the links between nodes in different layers indicate the direction of signal flow, and they have no weights. The nodes in [B] layer have the node parameters, while the ones in other layers have none. The node functions in the same layers are of the same form as described follows:

[A] Input layer. $o_j = I_j = x_j$, $j = 1, 2, ..., s$.

[B] Fuzzification layer. The node parameters include the centers and the radii of the membership functions in the fuzzy rule antecedent. The output of the nodes in the layer are the membership degree calculated by Eq.(4).

[C] Fuzzy rule layer. By Eq.(5), every circle node in the rectangles multiplies the incoming signals and sends the product out. The outputs of the layer make up the input parts of the corresponding subsystems.

[D] Subsystem output layer. The outputs of the layer are calculated by Eq.(6).

[E] Voting output layer. The output of the layer is calculated by Eq.(7).

5.2 Optimization for the Fuzzy Neural Network System

Assume the system has *r* subsystems, subsystem *j* contains a_j fuzzy IF-THEN rules, $j = 1, 2, ..., r$, and each fuzzy IF-THEN rule contains 2*s* parameters (i.e., the center

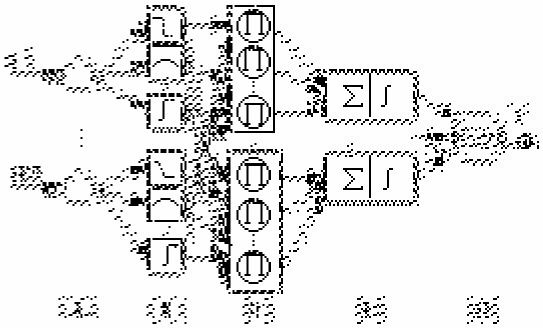


Fig. 1. The VMFNN architecture

and the radius in the Gaussian function), where *s* is the dimensionality of input feature vectors, then the number of the system parameters is totally $M = 2s\sum_{j=1}^{r} a_j$.

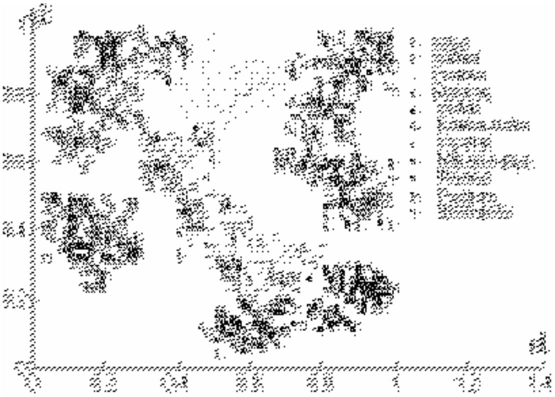


Fig. 2. 11 kinds of taste signals of mineral waters

The real-valued PSO is used to refine the system parameters, and the real-valued PSO algorithm flow can refer to [4][5]. Let *L* be the size of the particle swarm (generally set *L*=20). The initial velocity of the particles are initialized by random numbers uniformly distributed on [-0.3, +0.3], and the initial positions of the particles are initialized by the initial system parameters with 15% noise. Taking use of information included by the particle *i*, a fuzzy system as Fig.1 can be constructed. The misclassifycation rates of the system constructed by particle *i* are defined as: $\mathrm{E}_i = \frac{err_i}{n}$, where $i = 1, 2, ..., L$, *n* is the number of the training samples, and err_i is the misclassification number of the system constructed by particle *i*. Then the fitness of the particle *i* can be defined as : $f_i = 1 - \mathrm{E}_i$.

Set acceleration coefficients w=0.7298 , c_1=1.42 and c_2=1.57 , which satisfy the convergence condition of the particles: $w > (c_1 + c_2)/2 - 1$ [5]. Since $c_2 > c_1$, the particles will faster converge to the global optimal position of the swarm than the local optimal position of each particle. To avoid the premature convergence of the particles, an inactive particle, whose position unchanged in consecutive S epochs (set S=10 in this paper), will be reinitialized.

6 Experimental Results

The taste signals of 11 kinds of mineral waters[1] are used as experimental data in this paper. The experimental data consist of 1100 points with 100 points for each taste signal as shown in Fig. 2. When computing the class covers of a taste signal, let the taste signal be the set of B and all the others the set of R.

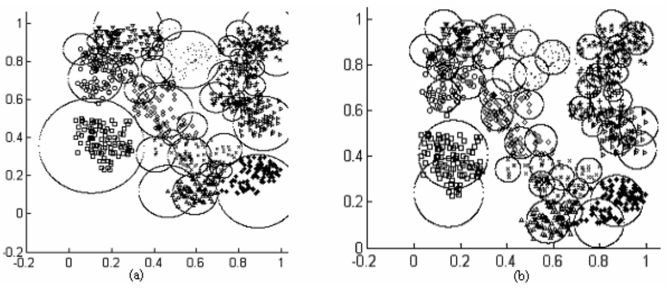


Fig. 3. (a)The class covers of the taste signals obtained by the original greedy algorithm; (b)The class covers of the taste signals obtained by the improved greedy algorithm ($\gamma = 0.5, \alpha = 0.96, \beta = 0.01$)

The class covers of the taste signals obtained by the original greedy algorithm and the improved greedy algorithm ($\gamma = 0.5, \alpha = 0.96, \beta = 0.01$) are shown in Fig 3(a) and Fig 3 (b), respectively. From Fig. 3(a), we can see that the obtained class covers are not even in size, and in which many with the radii of about zero (we call them outliers below) distribute over boundary areas between the taste signals, and that are unsuited for extracting robust fuzzy IF-THEN rules. From Fig. 3(b), we can obviously see that the obtained class covers are relatively even in size and have no outliers, that are suited for extracting fewer robust fuzzy IF-THEN rules. Comparison of the original greedy algorithm, the improved greedy algorithm and the bPSO algorithm in terms of K, Var and running time is listed in Table. 1.

From Table.1, we can obviously see that the bPSO algorithm can get better results than the improved greedy algorithm, but it needs more running time.

Then the fuzzy neural network system I and II (below which are abridged as Sys I and Sys II) can be constructed using the class covers obtained by the original greedy algorithm and the improved greedy algorithm ($\gamma = 0.5, \alpha = 0.96, \beta = 0.01$), respectively. The original taste signals are used for training the systems, and the taste signals

Table 1. Comparison of the three algorithms for the GCCP in terms of K, Var and Runing Time

	α	β	γ	K	Var	Runing Time
Original Greedy Algorithm				70	0.002617	8.503sec.
Improved Greedy Algorithm	0.97	0	0.4	72	0.000393	7.995sec.
	0.96	0.01	0.5	47	0.000647	8.477sec.
	0.95	0.02	0.6	36	0.000875	7.987sec.
bPSO Algorithm	0.97	0	0.4	67	0.000293	124.14sec.
	0.96	0.01	0.5	44	0.000447	123.58sec.
	0.95	0.02	0.6	36	0.000575	123.22sec.

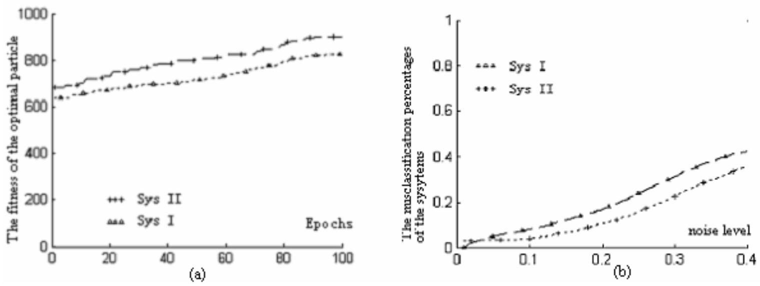


Fig. 4. (a)The curves of fitness of optimal particle in Sys I and II with respect to epochs; (b)The curves of the misclassification percentages in Sys I and II with respect to noise level

polluted by the noise are used for identification experiment. Note that, assume original signal is A, then the signal polluted by the noise is set $A' = A + A \times \eta \times rand$, where $0 \le \eta \le 1$, is the noise level, and *rand* is a random number uniformly distributed on [-1, +1].

The real-valued PSO is used to refine the system parameters of Sys I and Sys II. The curves of the fitness of the optimal particle of Sys I and Sys II with respect to training epochs and the curves of the misclassification percentages of Sys I and Sys II with respect to noise level are plotted in Fig.4 (a) and Fig.4 (b), respectively. Obviously, from Fig. 4(a), we can see Sys II shows better learning capability than Sys I, and from Fig. 4 (b), Sys II shows stronger noise-resistance capability than Sys I. On the other hand, because Sys II is constructed by fewer fuzzy IF-THEN rules than Sys I, it needs 71.363 sec. per 10 training epochs, while Sys I needs 106.146 sec. accordingly. Therefore Sys II is better than Sys I in terms of learning capability, error tolerance and running speed.

7 Conclusions and Discussions

By choosing proper parameters of γ, α and β in the improved greedy algorithm and the bPSO algorithm, fewer robust fuzzy IF-THEN rules can be obtained. Then those rules are used to construct the fuzzy neural network system, which has many perfect characteristics, such as better robustness, learning capability, simplicity and running speed. In our experiences, we set $\gamma = 0.3 \sim 0.7, \alpha = 0.90 \sim 0.97, \beta = 0 \sim 0.05$ as general choices.

Acknowledgement

This paper is supported by the National Natural Science Foundation of China under Grant No. 60433020.

References

1. Yanxin, H., Chunguang, Z.: Recognizing Taste Signals Using A Clustering-based Fuzzy Neural Network. Chinese Journal of Electronics, 14(1) (2005) 21-25
2. Cannon, A, Cowen, L.: Approximation Algorithms for the Class Cover Problem. Annals of Mathematics and Artificial Intelligence, .40(3) (2004) 215-223
3. Priebe, C.E., Marchette, D.J., DeVinney, J., Socolinsky, D.: Classification using Class Cover Catch Digraphs. Journal of Classification, 20(1) (2003) 3-23
4. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In Proceeding of IEEE International Conference on Neural Networks, Volume IV, Perth, Australia: IEEE Press. (1995) 1942-1948
5. Van, Den, Bergh, F.: An Analysis of Particle Swarm Optimizers [PH.D thesis]. Pretoria: Natural and Agricultural Science Department, University of Pretoria, 2001
6. Kennedy, J., Eberhart, R.C.: A Discrete Binary Version of the Particle Swarm Algorithm. Proceedings of the 1997 Conference on Systems, Man, and Cybernetics. Piscataway, NJ, IEEE Press (1997) 4104-4109
7. Jyh-Shing, R.J., Chuen-Tsai, S., Eiji, M.: Neuro-Fuzzy and Soft computing. Xi An: Xi An Jiaotong University Press, Feb. 2000
8. Ludmila, I.K.: How Good are Fuzzy If-Then Classifiers? IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 30(4) (2000) 501~509

Simulation-Based Optimization of Singularly Perturbed Markov Reward Processes with States Aggregation*

Dali Zhang, Hongsheng Xi, and Baoqun Yin

Network Communication System and Control Laboratory,
Department of Automation, University of Science and Technology of China
Hefei, Anhui, China 230027
zhangdl@mail.ustc.edu.cn, xihs@ustc.edu.cn

Abstract. We present a simulation-based algorithm to compute the average reward of singulary perturbed Markov Reward Processes (SPMRPs) with large scale state spaces, which depend on some sets of parameters. Compared with the original algorithm applied on these problems of general Markov Reward Processes (MRPs), our algorithm aims to obtain a faster pace in singularly perturbed cases. This algorithm relies on the special structure of singularly perturbed Markov processes, evolves along a single sample path, and hence can be applied on-line.

1 Introduction

Many models in communication network, finance, operations research, and other fields can be formulated by Markov models. Meanwhile, Markov Decision Processes (MDPs) are introduced to solve optimization problems in these models. The theory of Markov Decision Processes (MDPs) is a mathematical framework for modelling sequential decision tasks that had become very popular in the field of Intelligent Computing. And one of popular algorithms for MDPs is based on the simulation of a single sample path.

However, dimensions of state spaces in these models are often too large for our normal algorithms, and the computing of optimization for these decision problems will waste a lot of time and memory saving. Obtaining desired optimal control parameters or policies will be quite intensive and the way to solve the Markov Reward Process (MRP) or MDP with a large scale state space is a challenging issue at present in this field. But some of these large scale state space problems can be appropriately simplified. The aim of our paper is to create an algorithm to optimize a class of MRPs with large state spaces, in which underlying Markov chains have hierarchical structures and are called singularly perturbed Markov processes (SPMPs). The asymptotical properties of these reward processes are studied by G.Yin and Q.Zhang [2]and [5],and M.Abbad and

* This work was supported by National Natural Science Foundation of China under Grant(60274012) and Natural Science Foundation of Anhui Province under Grant(01042308).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 129–138, 2005.

J.A.Filar [3] and [4],respectively. From our simulation in Section 4, we will find that in some segments of a single sample path, the occurrence of states will be far more frequent than others. With the convenience of this property, we will introduce a new simulation-based algorithm to optimize an averaging reward problem of the original SPMRP. Differing from other methods, our algorithm is based on the aggregated states, which directly cuts down the steps of computing, and balances the frequency of every states taken in computing.

To concentrate the paper on methods based on policy parameterization and gradient improvement, here, we bring the conception of Markov Reward Processes, created and described a popular simulation-based method, displayed the whole schedule of this algorithm, and modified this algorithm into a new way with faster updating.

2 The Gradient of the Performance Metric

2.1 Singularly Perturbed Markov Reward Processes

Using the same notation as in [1], we consider a discrete time, finite-state Markov chain $\{i_n\}$ with state space $S = \{1, \cdots, N\}$, and its transition probabilities depend on a parameter vector $\theta \in \mathbb{R}^k$. Here, transition probabilities can be denoted by

$$P(j\,|i;\theta) = P(i_n = j \mid i_{n-1} = i;\, \theta)\ . \tag{1}$$

When the current state is i, the process receives a one-stage reward, which also depends on the parameter θ and denoted by $g_i(\theta)$. Transforming them into a matrix form $P(\theta) = [P(j\,|i;\theta)]_{N\times N}$. Henceforth, such an MRP can be expressed as the four-tuple $\Gamma =< S, \{\theta,\, \theta \in \mathbb{R}^k\}, \{g_i(\theta), i \in S\}, \{P(j\,|i;\theta), i, j \in S\} >$.

The performance metric used to compute different parameter θs is the average reward criterion $\eta(\theta)$, and defined as:

$$\eta(\theta) = \lim_{T\to+\infty} \frac{1}{T} E_\theta\left[\sum_{k=0}^{T} g_{i_k}(\theta)\right]\ . \tag{2}$$

where the i_k is the state at time k. If the transition probabilities matrix $P(\theta)$ is aperiodic, the average reward $\eta(\theta)$ is well defined, and does not depend on the initial state. The average reward can be rewritten as $\eta(\theta) = \sum_{i=1}^{N} \pi_i(\theta) g_i(\theta)$, where the steady state probability vector $\pi(\theta) = (\pi_1(\theta), \cdots, \pi_N(\theta))$ is the unique solution of the balance equations: $\pi P = \pi$ and $\pi e = 1$. If the matrix $P(\theta)$ here is aperiodic and the functions $P(j\,|i;\theta)$ and $g_i(\theta)$ are bounded, twice differentiable, and have bounded first and second derivatives, then $\pi(\theta)$ and $\eta(\theta)$ are also twice differentiable, and have bounded first and seconded derivatives.

The main objective of this paper aims to the properties of the singularly perturbed Markov Reward Process(SPMRP).In fact(see, for example,[2]), any transition probability matrix of a finite-state Markov chain without transient states can be of the form

$$P(\theta) = diag\{P_1(\theta),\, P_2(\theta), \cdots, P_n(\theta)\}\ . \tag{3}$$

where each $P_\alpha(\theta)$, $\alpha \in \{1, \cdots, n\}$ is a transition matrix within the αth recurrent class for $\alpha \leq n$. Here, we denote the steady state distribution corresponding to $P_\alpha(\theta)$ as $v_\alpha(\theta) = (v_\alpha^1(\theta), \cdots, v_\alpha^i(\theta), \cdots, v_\alpha^{m_\alpha}(\theta)) \in \mathbb{R}^{1\times m_\alpha}$, where $i \in S_\alpha$. First of all, we look into a general Markov chain with its states and the transition probabilities satisfy the following assumption:

A1. $S = \bigcup_{i=1}^n S_i$, where $S_i \bigcap S_j = \emptyset$, if $i \neq j$ and $|S_i| = m_i$, $m_1 + \cdots + m_n = 1$.
A2. $p\{s^{'}|s, \theta\} = 0$ whenever $s \in S_i$ and $s^{'} \in S_j$, $i \neq j$.
A3. For every $i = 1, 2, \cdots, n$, and for all $\theta \in \mathbb{R}^k$ the matrix $P_i(\theta)$ in matrix (3) is irreducible .

Then we consider the situation where the transition probabilities of Γ are perturbed slightly. We define the disturbance set $D = \{d(s^{'}|s, \theta)|s^{'}, s \in S, \theta \in \mathbb{R}^k\}$, and the elements of the set $D(\theta)$ satisfy: $\sum_{s^{'}\in S} d(s^{'}|s, \theta) = 0$. Now, we transfer these elements into a matrix form $D(\theta) = [d(s^{'}|s, \theta)]_{N\times N}$, where $D(\theta)$ can be seen as a generator. We shall also require that there exists $\epsilon_0 > 0$ such that $\forall\, \theta \in \mathbb{R}^k$

$$G^\epsilon(\theta) = G(\theta) + \epsilon D(\theta) \ . \tag{4}$$

is a generator of a Markov chain for any $0 < \epsilon < \epsilon_0$. Then we shift our aim to the perturbed Markov chain. Suppose that $\{i_n^\epsilon\}$ is a singularly perturbed Markov chain depending on a small disturbance parameter $\epsilon > 0$, and having a finite-state space $S = \{1, 2, \cdots, N\}$. The form of the transition probabilities matrix satisfies

$$P^\epsilon(\theta) = P(\theta) + \epsilon D(\theta) \ . \tag{5}$$

where $P^\epsilon(\theta)$ and $P(\theta)$ are the transition probability matrices, and $D(\theta)$ is a disturbance matrix. Here, singularly perturbed Markov chain are required to be also irreducible whenever $\forall \epsilon > 0$ and the assumptions are all satisfied,e.g.no matter how close ϵ approaches zero, every sub-state space will occur eventually.

One-step rewards of the SPMRP are the same as those in the general process, still denoted by $g_i(\theta)$, $i \in S$. Under the condition of the singularly perturbed Markov chain, we rewritten the SPMRP as Γ_ϵ, where $\epsilon \in (0, \epsilon_0]$, in the form of a four-tuple $\Gamma_\epsilon =< S, \{\theta, \theta \in \mathbb{R}^k\}, \{g_i(\theta), i \in S\}, \{P^\epsilon(j|i;\theta), i, j \in S\} >$.

Denoting the steady state distribution corresponding to P^ϵ by $\pi_\epsilon(\theta)$ and one step reward by $g(\theta) = (g_1(\theta), \cdots, g_N(\theta))$. We define the average reward as we discussed above

$$\eta_\epsilon(\theta) = \pi_\epsilon(\theta) g^T(\theta) \ . \tag{6}$$

And the optimal value function $\eta_\epsilon(\theta)$ corresponding to SPMRP is given by

$$\eta_\epsilon^* = \max_{\theta\in\mathbb{R}^k} [\pi_\epsilon(\theta) g^T(\theta)] \ . \tag{7}$$

To deal with singularly perturbed Markov reward problems, we will construct a new Markov reward process through states aggregation, which has an optimization objective asymptotically converging to the original one. Here, we denoted the aggregated state space as $\Omega = \{S_1, S_2, \cdots, S_n\}$, for more simple denotation as $\Omega = \{1, 2, \cdots, \alpha \cdots, n\}$. To proceed, we define a matrix $\widetilde{\mathbf{1}}$ as

$$\widetilde{\mathbf{1}} = diag\{\mathbf{1}_{m_1}, \mathbf{1}_{m_2}, \cdots, \mathbf{1}_{m_n}\} \ .$$

where the entry vector $\mathbf{1}_{m_\alpha} = (1, 1, \cdots, 1)^T \in \mathrm{I\!R}^{m_\alpha}, \forall \alpha \in \Omega$. From [2] and [3], we can construct a new Markov chain as $\{\alpha_k\}$, generated by the generator

$$\overline{Q}(\theta) = diag\{v_1(\theta), \cdots, v_\alpha(\theta), \cdots, v_{m_n}(\theta)\}D(\theta)\widetilde{\mathbf{1}} \ . \tag{8}$$

Let $\overline{\pi}(\theta)$ be the steady state distribution vector of the transition matrix $\overline{P}(\theta) = \overline{Q}(\theta) + \mathbf{I}_{n\times n}$. Directly from the Lemma 2.1 in [4], we assume there is another optimization problem named Aggregated Limiting Problem as

$$\overline{\Gamma}_\epsilon =< \Omega, \{\theta, \theta \in \mathrm{I\!R}^k\}, \{\overline{g}_\alpha(\theta), \alpha \in \Omega\}, \{\overline{P}(\alpha \,|\, \beta; \theta), \ \alpha, \beta \in \Omega\} > \ . \tag{9}$$

where vectors $\widehat{g}_\alpha(\theta) = (g_{\alpha_1}(\theta), \cdots, g_{\alpha_i}(\theta), \cdots, g_{\alpha_{m_\alpha}}(\theta)), \forall \alpha_i \in S_\alpha$, $\overline{g}_\alpha(\theta) = v_\alpha(\theta)\widehat{g}_\alpha^T(\theta), \forall \alpha \in \Omega$ and $\overline{g}(\theta) = (\overline{g}_1(\theta), \overline{g}_2(\theta), \cdots, \overline{g}_n(\theta))$. And its average reward function is defined as

$$\overline{\eta}(\theta) = \overline{\pi}(\theta)\overline{g}^T(\theta) \ . \tag{10}$$

Lemma 1. *Assume assumption A1)-A3),we have*

$$\lim_{\epsilon\to 0 \ \epsilon>0} |\max_{\theta\in\mathrm{I\!R}^k}[\overline{\pi}(\theta)\overline{g}^T(\theta)] - \max_{\theta\in\mathrm{I\!R}^k}[\pi^\epsilon(\theta)g^T(\theta)]| = 0 \ . \tag{11}$$

or in a more exact form as

$$|\max_{\theta\in\mathrm{I\!R}^k}[\overline{\pi}(\theta)\overline{g}^T(\theta)] - \max_{\theta\in\mathrm{I\!R}^k}[\pi^\epsilon(\theta)g^T(\theta)]| = O(\epsilon) \ . \tag{12}$$

which can be directly proved from the [2].

So any maximizing parameter θ for Aggregated Problem is also a maximizing parameter for θ and vice-versa. In the next section, we will resolve gradient of aggregated performance function to take place of original one. Hence, we can easily prove that $|\nabla\overline{\eta}(\theta) - \nabla\eta(\theta)| = O(\epsilon)$. If we optimize the sample path generated original perturbed process along the gradient of $\overline{\eta}(\theta)$, we also can obtain the optimized value of $\eta(\theta)$.

2.2 Properties of the Gradient of the Performance Metric

For any $\theta \in \mathrm{I\!R}^k$ and $\alpha \in \Omega$, we define the differential reward $D_\alpha(\theta)$ of an aggregated state $\alpha \in \Omega$ by

$$D_\alpha(\theta) = E_\theta\left[\sum_{k=0}^{T^{'}-1}(\overline{g}_{\alpha_k}(\theta) - \overline{\eta}(\theta))|\alpha_0 = \alpha\right] \ . \tag{13}$$

where α_k is the aggregated state at time t_k, the kth epoch with the transition between two different subspace in set Ω, which can be seen as an index for a certain segment of the sample path dominated by states in some S_α, and $T^{'} = \min\{k > 0|\alpha_k = \alpha^*\}$, where we have a general assumption $i^* \in S_*$. We also have properties that $D_{\alpha^*}(\theta) = 0$ and that the vector $\widetilde{D}(\theta) = (D_1(\theta)), \cdots, D_n(\theta))$ is a solution to Poisson equation

$$\overline{g}(\theta) = \widetilde{D}(\theta) + \overline{\eta}(\theta)\mathbf{1}_n^T - \overline{P}(\theta)\widetilde{D}(\theta) \ . \tag{14}$$

We will display a theorem which gives us an expression for the gradient of the average reward $\overline{\eta}(\theta)$, with respect to θ. Before we start our theorem, we first bring some assumptions here:

A4. The Markov chain corresponding to $\overline{P}(\theta)$ is aperiodic and irreducible. That is to say, there is a state α^* which is recurrent for the chain.This point can be directly derived from the aperiodicity and irreducibility of $P^\epsilon(\theta)$, and at least we can divide S into some S_α with $i^* \in S_{\alpha^*}$, which α_* is an index of some aggregated state.

A5. For every i, $j \in S$, the function $p^\epsilon_{ij}(\theta)$ and $g_i(\theta)$ are bounded, twice differentiable, and have bounded first and second derivatives. Hence, for every α, $\beta \in \Omega$, the functions $\overline{P}_{\alpha\beta}(\theta)$ and $\overline{g}_\alpha(\theta)$, as the linear combinations of the $p^\epsilon_{ij}(\theta)$ and $g_i(\theta)$, are also bounded, twice differentiable, and have bounded first and second derivatives.

Theorem 1. *Let A4), A5)hold. Then the gradient of the aggregated limiting average reward is*

$$\nabla\overline{\eta}(\theta) = \sum_{\alpha\in\Omega} \overline{\pi}_\alpha(\theta)\Bigg(\nabla g_\alpha(\theta)v^T_\alpha(\theta) + g_\alpha(\theta)\nabla v^T_\alpha(\theta) + \sum_{\beta\in\Omega}\Bigg[\sum_{i\in S_\alpha}\Bigg(\sum_{j\in S_\beta}\nabla d_{ij}(\theta)\Bigg)v^i_\alpha(\theta)$$

$$+ \sum_{i\in S_\alpha}\Bigg(\sum_{j\in S_\beta} d_{ij}(\theta)\Bigg)\nabla v^i_\alpha(\theta)\Bigg]D_\beta(\theta)\Bigg) \ . \tag{15}$$

Specially, when the disturbance matrix $D(\theta)$ is irrelative with the parameter θ, we have

Corollary 1. *Let A4), A5)hold, and the disturbances are independent of the parameter θ. Then*

$$\nabla\overline{\eta}(\theta) = \sum_{\alpha\in\Omega} \overline{\pi}_\alpha(\theta)\Bigg(\nabla g_\alpha(\theta)v^T_\alpha(\theta) + g_\alpha(\theta)\nabla v^T_\alpha(\theta) +$$

$$\sum_{\beta\in\Omega}\Bigg[\sum_{i\in S_\alpha}\Bigg(\sum_{j\in S_\beta} d_{ij}\Bigg)\nabla v^i_\alpha(\theta)\Bigg]D_\beta(\theta)\Bigg) \ . \tag{16}$$

The expressions given by Theorem 1 and Corollary 1 involve no terms of the steady state distribution of $P^\epsilon(\theta)$, but only involve with those of $P_\alpha(\theta)$, $\alpha \in \Omega$.

3 The Simulation-Based Optimization

In this section, we propose a simulation-based method to compute the gradient of $\eta(\theta)$. Moreover, we will routine the whole process of the algorithm for a disturbance-controlled case. As we proposed in Section 2, the extra estimators such as $\nabla v_\alpha(\theta)$ and $v_\alpha(\theta)$ are required before our estimating $\nabla\overline{\eta}(\theta)$ from $D_\beta(\theta)$.

3.1 Estimators of $\nabla v_\alpha(\theta)$

Compared with the algorithm without states aggregation(such as in [1]), $\nabla v_\alpha(\theta)$ can not be neglect here, so we should design a method to estimate them. However, these terms are not directly generated through our sample pathes with transition matrix $P^{\epsilon}(\theta)$. To obtain them, we should first look into sample pathes generated by the matrices $P_{\alpha}(\theta)$ to estimate $\nabla v_{\alpha}(\theta)$. $v_{\alpha}(\theta)$ is the steady state distribution of $P_{\alpha}(\theta)$, which can be easily estimated from counting every states along the sample path , and we also have balance equations: $v_{\alpha}(\theta)P_{\alpha}(\theta) = v_{\alpha}(\theta)$ and $v_{\alpha}(\theta)\mathbf{1}_{m_{\alpha}} = 1$. Here, we will propose a method based on a single sample path generated by $P^{\epsilon}(\theta)$ to approach results of a theoretical one. Now we take one of steady state probabilities $v_{\alpha}^{i}(\theta)$, $i \in S_{\alpha}$ as a new performance evolving with matrix $P_{\alpha}(\theta)$. To compute this performance, we set one step reward vector $\phi_{\alpha}^{i} = (\phi_{\alpha}^{i1}, \cdots, \phi_{\alpha}^{ij}, \cdots, \phi_{\alpha}^{im_{\alpha}}) \in \mathbb{R}^{1\times m_{\alpha}}$ as:

$$\phi_{\alpha}^{ii} = 1 \qquad \phi_{\alpha}^{ij} = 0 \qquad \forall j \neq i \qquad i\,, j \in S_{\alpha} \ . \tag{17}$$

As the performance metric introduced in Section 2 as equation (2), we first set the average reward criterion defined by

$$\mu(\theta) = \lim_{T\to+\infty} \frac{1}{T} E_{\theta}\left[\sum_{k=0}^{T} \phi_{\alpha}^{i\,i_k}(\theta)\right] \ .$$

where the whole process is generated by some $P_{\alpha}(\theta)$, so $i_k \in S_{\alpha}$, and we can easily find that $\mu(\theta) = v_{\alpha}^{i}(\theta)$. When we reform the average performance metric as $\mu(\theta) = v_{\alpha}(\theta)\phi_{\alpha}^{i} = v_{\alpha}^{i}(\theta)$, we obtain that our average reward is equal to the estimator for $v_{\alpha}^{i}(\theta)$. Similarly as computing the gradient of performance, from Theorem 1, we have

$$\nabla v_{\alpha}^{i}(\theta) = \sum_{i\in S_{\alpha}} v_{\alpha}^{i}(\theta) \sum_{j\in S_{\alpha}} \nabla p_{ij}(\theta) d_{j}^{i}(\theta) \ . \tag{18}$$

where $p_{ij}(\theta)$ is an entry in $P(\theta)$ in (3), and $d_{j}^{i}(\theta)$ is the differential reward corresponding to this partial problem to compute $\nabla v_{\alpha}^{i}(\theta) = \nabla\mu(\theta)$, defined by

$$d_{j}^{i}(\theta) = E_{\theta}\left[\sum_{k=0}^{T-1} (\phi_{\alpha}^{i\,i_k} - v_{\alpha}^{i}(\theta)) | i_0 = j\right] \ . \tag{19}$$

where i, i_k, $j \in S_{\alpha}$ and $T = \min\{k > 0 | i_k = i^{*}\}$ is the first future epoch state i^{*} is visited. Through an approximation we can connect these theoretical results with our simulation sample path as follows.

Lemma 2. *For any $\alpha_i\,, \alpha_j \in S_{\alpha}$, we have $|\nabla \widetilde{v}_{\alpha}^{i}(\theta) - \nabla v_{\alpha}^{i}(\theta)| = O(\epsilon)$ where*

$$\nabla \widetilde{v}_{\alpha}^{i}(\theta) = \sum_{i\in S_{\alpha}} v_{\alpha}^{i}(\theta) \sum_{j\in S_{\alpha}} \nabla p_{\alpha_i\alpha_j}^{\epsilon}(\theta) d_{j}^{i}(\theta) \ . \tag{20}$$

where α_i is the ith state in the set S_{α}.

we can proof this problem in a short way as

$$\nabla v_{\alpha}^{i}(\theta) = \sum_{i\in S_{\alpha}} v_{\alpha}^{i}(\theta) \sum_{j\in S_{\alpha}} \nabla[p_{\alpha_i\alpha_j}^{\epsilon}(\theta) - \epsilon q_{\alpha_i\alpha_j}(\theta)] d_{j}^{i}(\theta) \ .$$

So we can use sample pathes generated by $P^{\epsilon}(\theta)$ to approach the theoretical results, and as $\epsilon \to 0$ we have $\nabla \widetilde{v}_{\alpha}^{i}(\theta) = \nabla v_{\alpha}^{i}(\theta)$.

From Theorem 3.2 in [5], for any $i_t \in S$, $\alpha_t \in \Omega$, $\alpha_j \in S_\alpha$, $\alpha \in \Omega$ we have

$$\sup_{t\in[0,T]} E_\theta \left| \epsilon \sum_{t=0}^{k-1} (1_{\{i_t = S_{\alpha_j}\}} - v_{\alpha}^{\alpha_j} 1_{\{\alpha_t=\alpha\}}) \right|^2 = O(\epsilon) \ . \tag{21}$$

So we can use the states generated from the sample path to approximate the $v_{\alpha}^{i}(\theta)$ by

$$v_{\alpha}^{i}(\theta) = \frac{v_{\epsilon}^{i}(\theta)}{\sum_{j\in S_\alpha} v_{\epsilon}^{j}(\theta)} + O(\epsilon) = \frac{\sum_{t=0}^{k-1} 1_{\{i_t = S_{\alpha_j}\}}}{\sum_{t=0}^{k-1} 1_{\{\alpha_t=\alpha\}}} + O(\epsilon) \ . \tag{22}$$

where $v_{\epsilon}^{i}(\theta)$ is the ith element of $v_{\epsilon}(\theta)$ which is the steady state distribution vector of $P^{\epsilon}(\theta)$. When we obtain the estimator of $\nabla v_{\alpha}(\theta)$, we complete equation (15) in Theorem 1, and hence obtain $\nabla\overline{\eta}(\theta)$ from it. Following the direction of $\nabla\overline{\eta}(\theta)$, we can complete our optimization.

3.2 The Optimization for $D(\theta)$-Controlled Model

We will take a special but useful case into consideration in this section, that the matrix $P(\theta)$ is irrelative with θ, i.e. only the behavior of perturbation factor $D(\theta)$ is controlled by parameters. So the equation (5) can be written as $P^{\epsilon}(\theta) = P + \epsilon D(\theta)$. We call it $D(\theta)$-controlled model, and have $\nabla v_{\alpha}(\theta) = 0, \forall \alpha \in \Omega$, and

$$\nabla\overline{\eta}(\theta) = \sum_{\alpha\in\Omega} \overline{\pi}_{\alpha}(\theta) \left(\nabla g_{\alpha}(\theta) v_{\alpha}^{T} + \sum_{\beta\in\Omega} \left[\sum_{i\in S_\alpha} \left(\sum_{j\in S_\beta} \nabla d_{ij}(\theta) \right) v_{\alpha}^{i} \right] D_{\beta}(\theta) \right) \ .$$

And to finish the theoretical support of our algorithm, the following lemma is necessary:

Lemma 3. *The aggregated chain associated with the generator $\overline{Q}(\theta)$ is a Markov chain on Ω defined by $\alpha_n = i_{T_n}$, where the T_n are the successive epoches at which the chain enter another subspace: $i_{T_n} \in S_{\alpha_n}$ and $i_{T_n - 1} \notin S_{\alpha_n}$.*

This can be directly obtained from the Lemma 2 in [6], and the proof is omitted here. From (23), we can find out that the gradient of the performance is irrelative with $\nabla v_{\alpha}(\theta)$, so that only some counters are required to record the steady state distribution vectors. Meanwhile, a long term decision series for $\nabla v_{\alpha}(\theta)$ are not necessary. Here, we will display the algorithm of this case in Algorithm 1, and From the expression of this algorithm, we can easily find out that no matter how many states in the original state space S, the algorithm only evolves with the aggregated state space Ω.

4 Simulation and Results

In this section, we will give some examples to illustrate our algorithm, which will update the performance and parameters along the sample path generated by a

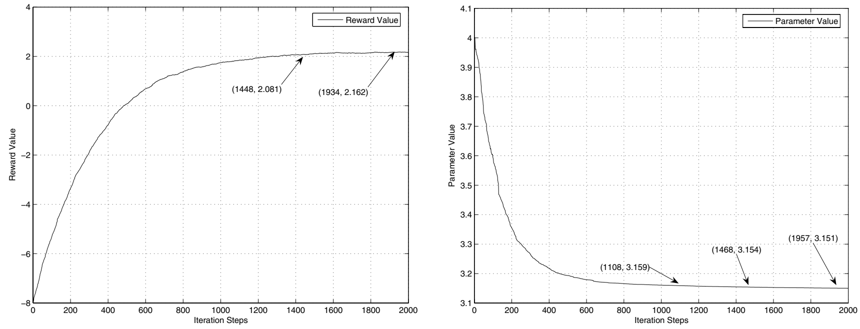
Algorithm 1. optimization algorithm based on states aggregation

```
0: Compute $\overline{g}_\alpha(\theta)$, and $D^\beta(\theta)=\sum_{j\in\beta} d_{ij}(\theta)$ for all $\alpha$, $\beta \in \Omega$ and $i \in S$.
1: Estimate the steady state distributions of $P_\alpha(\theta)$ as:
       $\widehat{\chi}_{\alpha_i}(T)=\sum_{k=0}^{T} 1_{\{i_k=\alpha_i\}}$ ,   $\widehat{\chi}_\alpha(T)=\sum_{k=0}^{T} 1_{\{i_k\in S_\alpha\}}$ ,   $\widehat{v}_\alpha^{\alpha_i}(T)=\frac{\widehat{\chi}_{\alpha_i}(T)}{\widehat{\chi}_\alpha(T)}$ .
   until the condition, for some $\epsilon$ $|\widehat{v}_\alpha^{\alpha_i}(T+1)-\widehat{v}_\alpha^{\alpha_i}(T)|<\varepsilon$, is satisfied.
2: Compute some factors:
       $\widehat{F}_1^{\alpha}(\theta)=\widehat{v}_\alpha\overline{g}_\alpha^T(\theta)$ ,   $\widehat{F}_2^{\alpha\beta}(\theta)=\widehat{v}_\alpha[D^\alpha(\theta)]^T+1$ .
   where $\widehat{v}_\alpha=(\widehat{v}_\alpha^{i_1},\cdots,\widehat{v}_\alpha^{i_{m_\alpha}})\in \mathrm{I\!R}^{1\times m_\alpha}$, $D^\alpha(\theta)=(D^{i_1}(\theta),\cdots,D^{i_{m_\alpha}}(\theta))$,
   and $i_1,\cdots,i_{m_\alpha}\in S_\alpha$.
3: Recursive in every epoch $k$ where $\alpha_k=i_{T_k}\neq i_{T_k-1}$, and compute following
   until the recurrent aggregated state $S_*$ or $\alpha_*$ is first visited in future
       $\widehat{D}_{i_n}(\theta,\widehat{\eta}(\theta_m))=\sum_{k=n}^{t_{m+1}-1}(\widehat{F}_1^{\alpha_k}(\theta_m)-\widehat{\eta}(\theta_m))$ ,
       $F_m(\theta_m,\widehat{\eta}(\theta_m))=\sum_{n=t_m}^{t_{m+1}-1}\left(\nabla\widehat{F}_1^{\alpha_n}(\theta_m)+\widehat{D}_{i_n}(\theta_m,\widehat{\eta}(\theta_m))\frac{\nabla\widehat{F}_2^{\alpha_{n-1}\alpha_n}(\theta_m)}{\widehat{F}_2^{\alpha_{n-1}\alpha_n}(\theta_m)}\right)$ ,
       $\theta_{m+1}=\theta_m+\gamma_m F_m(\theta_m,\widehat{\eta}(\theta_m))$ ,
       $\widehat{\eta}_{m+1}(\theta)=\widehat{\eta}_m(\theta)+\lambda\gamma_m\sum_{n=t_m}^{t_{m+1}-1}(\overline{g}_{\alpha_n}(\theta_m)-\widehat{\eta}_m(\theta_m))$ .
   where $\gamma_m$ is a positive step size sequence, $\lambda>0$ allows to scale the step size
   for updating $\widehat{\eta}_m(\theta)$ by a positive constant, and $\alpha_k$ is the aggregated state.
4: When $\alpha_*$ is revisited, return to step 1, unless $|\overline{\eta}(\theta_{t_{m+1}})-\overline{\eta}(\theta_{t_m})|<\delta$ , where
   $\delta>0$ is small enough,or $\{\overline{\eta}(\theta_{t_{m+k}}),\ k=1,2,\cdots,l\}$ enter some stable domain.
```

singularly perturbed Markov process. Here, we name every epoch of the data updating as iteration, and iteration steps is an important measurement to value an algorithm. The sample path generated by the transition probability matrix with the form of:

$$P^\epsilon(\theta)=\begin{pmatrix} P_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & P_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & P_3 \end{pmatrix}+\epsilon\times\begin{pmatrix} D_{1,1}(\theta) & D_{1,2}(\theta) & D_{1,3}(\theta) \\ D_{2,1}(\theta) & D_{2,2}(\theta) & D_{2,3}(\theta) \\ D_{3,1}(\theta) & D_{3,2}(\theta) & D_{3,3}(\theta) \end{pmatrix} .$$

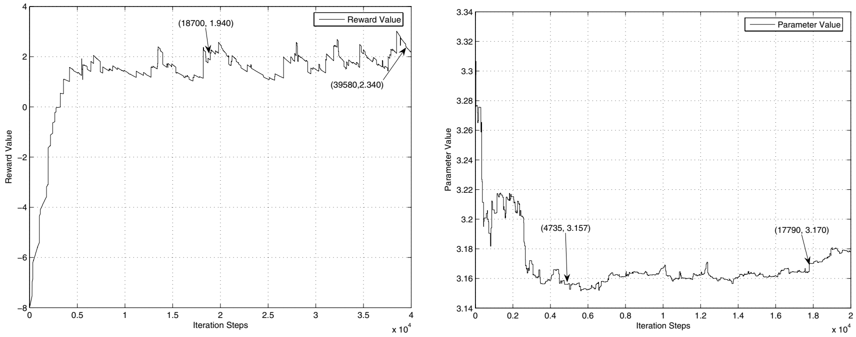
where the second term of equation is the disturbance matrix with all entries in $D_{\alpha,\beta}(\theta)$ larger than $0, \forall\theta\in\mathrm{I\!R}^k$, for any $\alpha\neq\beta$, and $\epsilon=0.001$. All of the sub-matrix $P_\alpha, \alpha=1,2,3$ are with a large scale state space. The recurrent state i^* is assumed in the subset S_1 corresponding to the transition matrix P_1, so the aggregated state $\alpha^*=S_1$ is the recurrent state after the aggregation.



(a) Iteration for Performance in The Algorithm with States Aggregation

(b) Iteration for Parameter in The Algorithm with States Aggregation

Fig. 1. Algorithm with States Aggregation



(a) Iteration for Performance in The Algorithm without States Aggregation

(b) Iteration for Parameter in The Algorithm without States Aggregation

Fig. 2. Algorithm without States Aggregation

In our algorithm, we view the sample path as the evolution of aggregated states. So the recurrent state should also be defined in the sense of aggregated state, and we denote it as α^*. The other details in our simulation are omitted here. Obviously, from these figures, we can find that algorithm with state aggregation will lead to a smoother optimization. Furthermore, the optimization will be faster than that without state aggregation.The results in Fig.1. show that the convergence in algorithm with state aggregation will be better than the other. It is just because of the sample path of the singularly perturbed Markov chain, which always transits between long segments, which are dominated by a certain subset $S_i, i = 1, 2, 3$, so our iterations in an algorithm without states aggregation are constrained as a local one with states in S_α until the single sample path arrives some new segment. We can easily find these shortcomings from the Fig.2, for example, the undulation of the optimized performance in the Fig.2(a). With states aggregation, our algorithm cut down all of these shortcomings.

5 Conclusion

In this paper, we obtain the gradient of performance of singularly perturbed Markov reward processes based on aggregated states, and design an algorithm which can save iteration steps and the mount of computing during the optimization. Our designment also can optimize the objective on-line with the evolution of practical processes.

References

1. Marbach, P., Tsitsiklis, J.N.: Simulation-Based Optimization of Markov Reward Processes. IEEE Transactions on Automatic Control, Vol. 46, No. 2, February (2001) 191-209
2. Yin, G., Zhang, Q.: Singularly Peturbed Discrete-Time Markov Chains. SIAM Journal appl. math. Vol. 61, No. 3(2000) 834-854
3. Abbad, M., Filar, J.A.: Perturbation and Stability Theory for Markov Control Problems. IEEE Transactions on Automatic Control, Vol. 37, No. 9, September(1992) 1415-1420
4. Abbad, M., Filar, J.A.: Algorithms for Singularly Perturbed Limiting Average Markov Control Problem. IEEE Transactions on Automatic Control, Vol. 37, No. 9, September(1992) 1421-1425
5. Yin, G., Zhang, Q., Badowski, G.: Discrete-Time Singularly Perturbed Markov Chain: Aggregation, Occupation Measures, and Switching Diffusion Limit. Adv. Appl. Prob. Vol. 35(2003) 449-476
6. Delebecque, F.: A Reduction Process for Perturbed Markov Chains. SIAM Journal Appl.Math. Vol. 43, No. 2, April(1983) 325-350
7. Liu, R.H., Zhang, Q., Yin, G.: Singularly Perturbed Markov Decision Processes in Discrete Time. Decision and Control, Proceedings of the 40th IEEE Conference on, Vol. 3, December(2001) 2119 - 2124
8. Hao, T., Xi, H.S., Yin, B.Q.: Performance Optimization of Continuous-Time Markov Control Processes Based on Performance Potentials. International Journal of Systems Science, 34(1)(2003) 63-71

Semi-active Control for Eccentric Structures with MR Damper by Hybrid Intelligent Algorithm

Hong-Nan Li[1], Zhiguo Chang[2], and Jun Li[1]

[1] Department of Civil Engineering, Dalian University of Technology,
116024 Dalian, China
hnli@dlut.edu.cn
civilli@163.com
[2] Department of Civil Engineering, Tongji University, 200092 Shanghai, China
zgchang@tongji.edu.cn

Abstract. In this paper, an application of hybrid intelligent control algorithm to semi-active Control of the MR Damper is presented for engineering structures. The control signal is optimized directly by the μGA approach to obtain the numerical relation between the control signal and the system output. And then, this relation is stored in the weight value of a trained artificial neural network, which can be available for another structure subjected to other seismic inputs. The results of numerical example indicate that the semi-active control of MR damper based on the hybrid algorithm can efficiently reduce the structural responses induced by earthquake.

1 Introduction

Magneto-rheological (MR) fluid is a kind of controllable rheological fluid[1]. When subjected to magnetic field, the rheological properties of the fluid can be reversibly and dramatically changed from the free-flowing Newton fluid to semi-solid fluid with certain shear strength. Thus, the MR damper is developed from MR fluid by reason that this microscopic shift can induce macroscopically the damping force of the MR damper to change greatly. Since the dynamic constitutive relation of the MR fluid is complex and nonlinear, it is rather difficult to obtain the mapping relationship between the damping force of the MR damper and structural response[2-4].

In this paper, an application of hybrid intelligent control algorithm to semi-active Control of the MR damper is presented for engineering structures. In the process of calculation, the control signal is optimized directly by the Micro-GA Approach to obtain the digital relation between the control signals and system output. And then, this relation is stored in the weight values of a trained ANN, which can be available for another structure subjected to other seismic inputs. Results of the numerical example indicate that the MR damper applied for the semi-active control can efficiently reduce the structural responses induced by an earthquake.

2 Modeling of Structural Control

An appropriate holistic model should be established for dynamic response analyses of structure under the action of earthquake. Due to the asymmetry of structure and multi-

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 139 – 148, 2005.

dimensional feature of ground motion, which induce the structural response to be multi-dimensional, seismic response of asymmetric structure will be the spatial vibration with the coupled translational-torsional motion. If only considering one-directional dynamic response of structure, the actual seismic response of asymmetric structure cannot be correctly evaluated[5].

The equation of motion for the structural system with the coupling torsional motion subjected to bi-directly horizontal seismic excitations is derived as follows:

$$[M]\{\ddot{U}\}+[C]\{\dot{U}\}+[K]\{U\}=-[M][H_g]\{\ddot{X}_g\} \tag{1}$$

where $\{U\}=\{U_{x1},U_{x2},...U_{xn},U_{y1},U_{y2},...U_{yn},U_{\theta 1},U_{\theta 2},...U_{\theta n}\}^T$ is the displacement vector of structure; $[M]$,$[C]$ and $[K]$ are the mass, damping and stiffness matrices of structure, respectively; $\{\ddot{X}_g\}=\{\ddot{X}_{Xg},\ddot{X}_{Yg}\}^T$ means the vector of biaxial horizontal seismic excitation; $[H_g]$ denotes the location matrix of seismic excitation, given by $[Hg]=\begin{Bmatrix} \{I_{n\times 1}\} & 0 \\ 0 & \{I_{n\times 1}\} \\ 0 & 0 \end{Bmatrix}$, where $\{I_{n\times 1}\}=\{1,1,..,1\}^T$.

Experimental results have shown that the relationship between damping force and relative velocity for the MR fluid is of obviously hysteretic characteristics[1]. Thereby, a more accurate model of the MR damper is accomplished in terms of the nonlinear bi-viscous model with four parameters that are the pre-yielding damping coefficient C_{pr}, post-yielding damping coefficient C_{po}, damping force F_y, and the velocity v_0 when the damping force is equal to the zero value. The relation between the damping force and relative velocity for the MR damper is illustrated in Fig. 1, the expression of which is given by:

$$F_d=\begin{cases} C_{po}\dot{x}-F_y & \dot{x}\le -v_1 & \ddot{x}>0 \\ C_{pr}\left(\dot{x}-v_0\right) & \text{-}v_1\le \dot{x}\le v_2 & \ddot{x}>0 \\ C_{po}\dot{x}+F_y & \dot{x}\ge v_2 & \ddot{x}>0 \\ C_{po}\dot{x}+F_y & \dot{x}\ge v_1 & \ddot{x}<0 \\ C_{pr}\left(\dot{x}+v_0\right) & \text{-}v_2\le \dot{x}\le v_1 & \ddot{x}<0 \\ C_{po}\dot{x}-F_y & \dot{x}\le -v_2 & \ddot{x}<0 \end{cases} \tag{2}$$

where v_1 and v_2 represent the compressive yielding velocity and stretched yielding velocity, respectively, given by:

$$v_1 = \frac{F_y - C_{pr} v_0}{C_{pr} - C_{po}},\ v_2 = \frac{F_y + C_{pr} v_0}{C_{pr} - C_{po}} \tag{3}$$

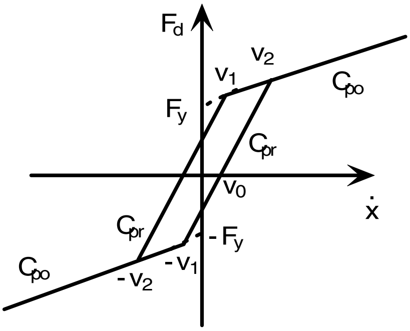


Fig. 1. Nonlinear Hysteretic Bi-Viscous Model for MR Damper

The range of the damping force obtained by Eq. (2) is $F_y \in (0, F_{y,\max})$. And the damping force can be changed in terms of imposed current as follows:

$$F_y = F_{y,\max} f_s \tag{4}$$

where $f_s = v / v_{\max}$ means the controlling signal of the damper, apparently, $f_s \in [0,1]$.

From Eqs. (2) and (4), it can be found that the damping force is a function of the damper piston velocity $\dot{x}$ and the applied controlling signal f_s, i.e.

$$F_d = F_d(\dot{x}, f_s) \tag{5}$$

Accordingly, the equation of motion for multi-story eccentric structure with dampers subjected to two-dimensional seismic excitations can be expressed as

$$[M]\{\ddot{U}\} + [C]\{\dot{U}\} + [K]\{U\} = -[M][H_g]\{\ddot{X}_g\} + [H_c]\{F_c\} \tag{6}$$

where $[H_c]$ denotes the location matrix of the dampers; $\{F_c\}$ is the control force vector of damper, every component of which is calculated according to Eq. (5).

Let the state variable be $\{X\} = \{\{U\}^T\ \{\dot{U}\}^T\}^T$. Then, if the floor displacement with the dimension of p and floor velocity with the dimension of q are defined as the observed variable $\{Y_o\}$, the dynamic system described by Eq. (6) can be transformed into state space equation:

$$\begin{cases} \{\dot{X}\} = [A]\{X\} + [B]\{F_c\} + [L]\{\ddot{X}_g\} \\ \{Y_o\} = [C_o]\{X\} \end{cases} \tag{7}$$

where

$$\{Y_o\} = \{\Delta U_{(1)}, \Delta U_{(2)}, ..., \Delta U_{(p)}, \Delta\dot{U}_{(1)}, \Delta\dot{U}_{(2)}, ..., \Delta\dot{U}_{(q)}\}^T;$$

$$[A] = \begin{Bmatrix} 0 & I \\ -M^{-1}K & -M^{-1}C \end{Bmatrix};\ [B] = \begin{Bmatrix} 0 \\ M_{-1}H_c \end{Bmatrix};\ [L] = \begin{Bmatrix} 0 \\ H_g \end{Bmatrix};$$

in which $[C_o]$ means the transformation matrix of observed variable, which is related to the location of sensors. From Eq. (5), one can obtain:

$$\{F_c\} = F_c(\{Y_c\}, F_s) \tag{8}$$

where $\{F_s\}$ is the controlling signal; $\{Y_c\}$ represents the relative velocity vector at the locations of dampers, determined by:

$$\{Y_c\} = [C_c]\{X\} \tag{9}$$

in which $[C_c]$ denotes the transform matrix of observed variable relative to the location of sensors, in which the derivation of $[C_c]$ can be found in Reference 6.

The design objective of the controlled system described by Eq. (7) is to find a mapping relationship, $\{F_s\} = F_s(\{Y_o\})$, to make a performance of the system, $J = J(\{X\}, \{Y_o\}, \{F_c\})$, meet a definite demand (maximization or minimization).

3 Hybrid Intelligent Algorithms

3.1 Basic Process of GA

The GA, which is based on natural selection conceiving of eliminating the inferiority and retaining the superiority, as it were, "survival of the fittest", transforms the possible solution space of problem into the genetic operation space to search the optimized solutions. For the design of a control system, suppose an optimization problem to be:

$$\begin{cases} f = f(\{Z\}) \\ \{Z\} \in D \end{cases} \tag{10}$$

where $\{Z\} = \{z_1, z_2, ..., z_n\}^T$ is the independent variable vector. Each vector $\{Z\}$ is a possible solution. Thus, D can be regard as not only the definition space of $\{Z\}$, but the constraints of the problem and solution space composed of solutions.

The objective of optimal problem in Eq. (10) is to search optimized parameters $\{Z_{\mathrm{m}}\} \in D$ that optimize the target function f.

Consequently, the problem comes down to the parametric optimal problem of cost function in the n-dimensional space. For the class of problem, it is not easy to derive directly the solution from the target function. Normally, the ANN or the climbing mountain searching optimal algorithm depending on gradient can be efficiently applied for some problems. Yet, for the parametric optimal problem in the n-dimensional space, the optimizing strategies may converge towards local optimal solution, but not global optimal solution. However, the GA is an effective method to solve the kind of problem.

3.2 Micro-GA Approach

The calculation of fitness function is the most time-consuming process of the above-mentioned genetic operators. Furthermore, the population scale of standard GA (ranging in size from 30 to 200) is large and computation is complicated. Accordingly, a GA with small a population, namely, Micro-GA, is developed to achieve a fast turnaround time from the generation to next generation evolution. Its essential point is that the evaluating and ceasing of the algorithm are not based on the average performance of the population, but on the best individual. However, it is a known fact that the information of the GA is generally less with very small population due to insufficient information processing and early convergence to non-optimal results. On another hand, some new individual is continuously introduced to prevent the premature convergence that is caused by the reduction of population size in this algorithm. The investigation results showed that the implement of the Micro-GA for the optimization problem looks highly feasible and rewarding[7].

3.3 Processing of ANN Training Data

Multi-layer feed-forward ANN is composed of one input layer, one output layer and one or several hidden layers. Commonly, the transformation function of the hidden layer is nonlinear, such as the S type function or hyperbolic tangent function. And the transform function of output layer determined by mapping the relation between input and output is linear or nonlinear. Theoretically, any function can be approximately used for the multi-layer feed-forward NN.

After the optimized control signals are obtained by using the above Micro-GA, the numerical relation between control signals and outputs of system will be stored in the weight value of artificial neural network to train. And the training data should be normalized before training the ANN so that the absolute value of the data is made less than 1. By the processing, the adaptation of the ANN increases.

If the domain of an ANN input component N_i is $(N_{i,\min}, N_{i,\max})$, the normalized $\overline{N}_i$ is:

$$\overline{N_i} = \frac{2(N_i - N_{i,\min})}{N_{i,\max} - N_{i,\min}} - 1 \tag{11}$$

When the ANN output function is bounded, the normalization of the training data is unnecessary. Whereas, the training data must be normalized when the output function is unbounded:

$$\overline{N_o} = \frac{2(N_o - N_{o,\min})}{N_{o,\max} - N_{o,\min}} - 1 \tag{12}$$

The output data of the trained ANN should be adjusted so as to obtain the actual output based on the following equation:

$$N_o = \frac{\overline{N_o}(N_{o,\max} - N_{o,\min}) + 1}{2} + N_{o,\min} \tag{13}$$

4 Control of Multi-story Structure Using MR Damper

In order to explore the effectiveness of the above-discussed intelligent control algorithm, let us consider a six-floor asymmetric structure shown in Fig. 2, where M_c is the center of mass and K_c is the center of stiffness. The floor-frame model is adopted to calculate the structural response. The structural parameters: the mass of every floor is 922×10^3 Kg; the inertia moment of every floor is 2.797×10^8 Kgm^2; the lateral stiffness of every column of the lowest story in x-axis and y-axis directions are both 2.836×10^7 KNm, and that of other columns are 4.504×10^7 KNm.

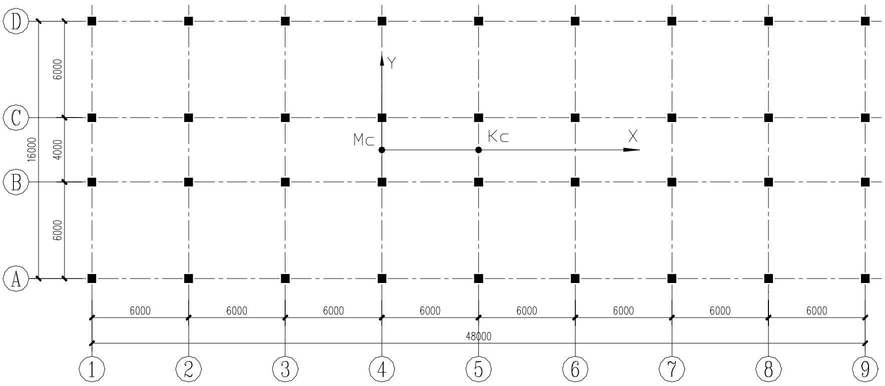


Fig. 2. Locations of columns

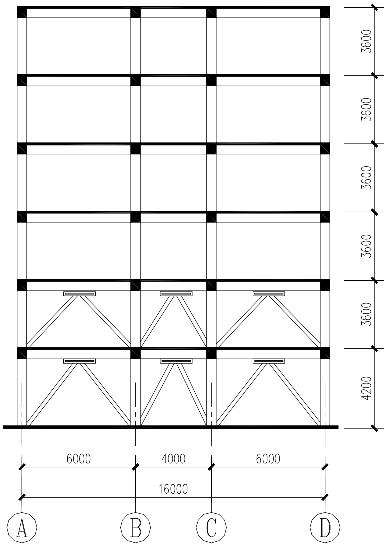


Fig. 3. Locations of dampers

The first several periods of structure based on calculations are: 0.82801s, 0.6758 s, 0.5805 s, 0.27655 s and 0.22571 s. The MR dampers made by American Lord Corporation are used as control device. Its energy consumption is 22 watt and maximal damping force is 200KN. The installed places of devices are shown in Fig. 3.

The selected seismic wave as the structural inputs are (1) the El Centro wave of USA on May 18, 1940, (2) the Seattle wave of USA on April 13, 1949 and (3) Taft wave of USA on July 21, 1952. It is convenient for comparisons that the x-component peak acceleration of each wave is adjusted to 200 cm/s^2 and y-component is adjusted to 250 cm/ s^2.

The seismic response of asymmetrical structure often shows the effect of translational motion coupled with torsion . In this instance (shown in Fig. 2), the Y-axis displacement of columns at ⑨ axis is magnified since the mass center does not coincide with the stiffness center. In order to control the torsional response, the dampers are fixed at ① and ⑨ axis along with Y direction (Fig. 3).

4.1 Optimizing Control Signals Using Micro-GA

In this section, the optimized numerical solution of Eq. (10) is calculated directly using the Micro-GA in interval $\tau = n\Delta\tau$ composing of sample periods $\Delta\tau$ with size of n . The time serial $\{F_s\}$ is transformed into space serial:

$$\{Z\} = \{\{F_s\}_0^T, \{F_s\}_{\Delta\tau}^T, \{F_s\}_{2\Delta\tau}^T, \ldots \{F_s\}_{(n-1)\Delta\tau}^T\}^T \tag{14}$$

Firstly, the optimized control signal vector $\{F_s\}$ at interval τ composing of sample periods with $\Delta\tau$ size of n is obtained by using the Eq. (2) as fitness function. And then, for the next interval τ of the simulating duration, the corresponding control signals are achieved in terms of the above control theory. It is noticeable that the initial values of the two '1/s' modules are not zeroing, but the final velocity and displacement at the last interval, respectively.

The controlled and uncontrolled peak displacements of floors and control efficiency are shown in Table 1. It can be seen from the table that the Micro-GA has the good control efficiency on optimizing control signals. Although Eq. (10), which shows the relationship between control signal and output of system, is complicatedly nonlinear, the optimized control signal can be gained directly in virtue of the optimizing ability of the Micro-GA.

Table 1. Reduction effect using GA control

Floor	Uncontrolled displacement peak (mm)	Controlled displacement peak (mm)	Reduction ratio (%)
6	103.4275	71.4949	30.8744
5	98.1718	67.5032	31.2397
4	88.1077	60.1003	31.7876
3	73.7994	50.5815	31.4607
2	55.9789	38.5027	31.2192
1	35.4568	24.3807	31.2383

4.2 Nonlinear Feedback Control Using Hybrid ANN

Although the optimized control signal can directly be obtained by using the Micro-GA, the relationship between control signal and output of system cannot be expressed by a known way. Hence, it cannot be used when subjected to other seismic wave. Just as well, the ANN has an ability to approximate the nonlinear mapping, and it can be used to learn this relationship that can be used to structural control subjected to other seismic wave.

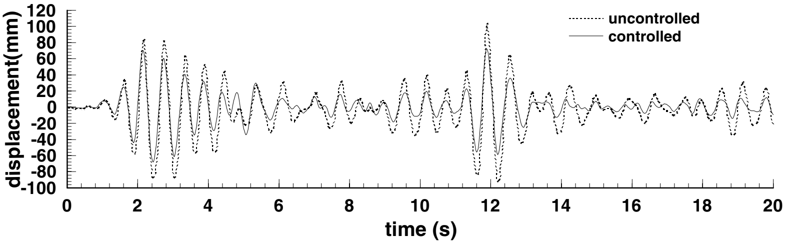


Fig. 4. Y-axis displacement reduction at 6th floor (hybrid ANN control, El-Centro wave)

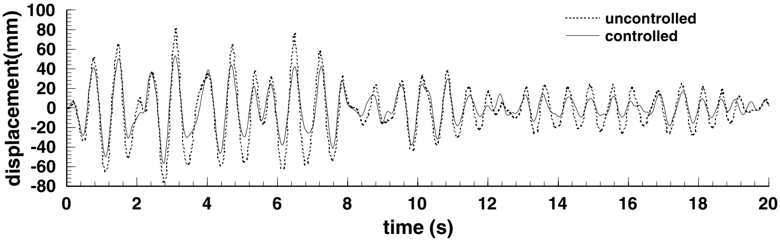


Fig. 5. Y-axis displacement reduction at 6th floor (hybrid ANN control, Seattle wave)

Firstly, a feed-forward ANN with two hidden layers is constructed. The transformation function of the hidden layers is expressed by the hyperbolic tangent function, and that of output layer with the domain [0, 1] is the Logsig function, respectively. The node numbers of two hidden layers are chosen as 20 and 10, while the output layer is selected as 3. After the optimized control signals are obtained by using the Micro-GA, the relationship between control signals and outputs of system is stored in the weight value of artificial neural network to train, which can be available for the structure subjected to other seismic waves. The data must be normalized to make the absolute value of data be less than 1 before the ANN is trained. The comparison curves of controlled and uncontrolled are shown in Fig. 4 and 5.

The controlled and uncontrolled displacement peaks and reduction ratios of the floor responses using the hybrid ANN approach are shown in Table 2. By comparing the Table 2 with Table 1, it can be concluded that the control effect using the hybrid ANN control approach that is trained in terms of the nonlinear relationship between the control signal and system output is close to that only using the μGA control. However, its advantage is that when subjected to other wave, the control effectiveness using the trained hybrid ANN approach is also efficient.

Table 2. Reduction effect using hybrid ANN with El Centro wave

Floor	Controlled displacement peak (mm)	Uncontrolled displacement peak (mm)	reduction ratio (%)
6	72.5762	103.4275	29.8289
5	68.5502	98.1718	30.1733
4	60.8727	88.1077	30.9110
3	50.6066	73.7994	31.4268
2	38.4519	55.9789	31.3100
1	35.4568	24.2834	31.5129

5 Conclusions

Although semi-active control is essentially most intricately nonlinear control, it can be efficiently solved by the intelligent control theory combined with the classic control theory. In this paper, a control strategy based on the GA and ANN for the semi-active control is proposed. Since the Linear Quadratic performance function is introduced, the proposed control strategy also belongs to the class of optimal control. The explored main contents include:

(1)The eccentricity of structure and the multidimensionality of seismic excitation are considered in the analyses. Under the action of earthquake, the torsion effect greatly impacts on the structural responses coupled with the plane and torsion motions.

(2)The nonlinear relationship between the damping force of the MR damper and control signal is considered.

(3)The hybrid intelligent control algorithm is introduced to the structural control. The control signal is optimized directly by the GA to obtain the numerical relation between the control signal and the system output. And then, this relation is stored in the weight value of a trained neural network so that it can be available for other seismic inputs.

The numerical results have shown that the strategy presented here is simple, robust, fault-tolerant and effective.

References

1. Spencer, B. F. *et al*:Phenomenological Model for Magnetorheological Dampers. ASCE, Journal of Engineering Mechanics. 123 (1996) 230-237
2. Dyke, S. J. *et al*:Modeling and Control of Magnetorheological Dampers for Seismic Response Reduction. Smart Mater. and Struct. 5 (1996) 565–575
3. Laura, M. Jansen and Shirley, J. Dyke. : Semi-active Control Strategies for MR Dampers: Comparative Study. Journal of Engineering Mechanics. ASCE, 126 (2000) 795-803
4. Norman, M. W. and Li, P.: Non-dimensional Analysis of Semi-active Electrorheological and Magnetorheological Dampers Using Approximate Parallel Plate Models. Smart Mater. and Struct. 7 (1998) 732–743
5. LI Hong-nan *et al.*: Stochastic Coupled Response Analysis of Asymmetrical Structure Due to Multidimensional Seism. Journal of Building Structures. 13 (1992) 12-20
6. CHANG Zhi-guo: Application of Intelligent Algorithm to Semi-Active Control of MR Damper. Shenyang: master's degree paper of Shenyang architectural and civil Engineering institute [Advised by Hong-nan LI] (2003)
7. LI Hong-nan, CHANG Zhi-guo: Studies on Optimization of Structural Vibration Control by GA. Earthquake Engineering and Engineer Vibration. 22 (2002) 92-96

Control Chaos in Brushless DC Motor via Piecewise Quadratic State Feedback

Hai Peng Ren[1] and Guanrong Chen[2]

[1] School of Automation and Information engineering, Xi'an University of Technology, Xi'an, China
[2] Department of Electronic Engineering, City University of Hong Kong, Hong Kong SAR, China

Abstract. The chaotic phenomenon in the brushless DC motor is revisited in this paper. For a specific real physical plant (i.e., the brushless DC motor), the main drawback of some existing chaos control methods is first analyzed. Then, a piecewise quadratic state feedback method is proposed for controlling chaos in the brushless DC motor model. In the proposed method, the direct-axis (or quadrature-axis stator voltage) is used as the control variable, and the piecewise quadratic state feedback is used as the control law. The control mechanism is illustrated and then the principle of parameters selection is discussed. Chaos in the brushless DC motor model can be satisfactorily eliminated by the proposed piecewise quadratic state feedback, and the speed of the motor can be stabilized to a constant value. This control method is simple and can be easily implemented. Simulation results verify the effectiveness of the method.

1 Introduction

Power electronics systems are highly nonlinear due to the on-off operations of the devices. Kuroe and Hayashi [1] pointed out earlier that chaos can exist in a power electronics system, and Kuroe et al. [2] showed that chaos can also be observed in some other motor-driven systems. From then on, many chaotic phenomena in typical power electronics particularly converters have been reported [3-5]. Chaos in brushless DC motors was first studied via simulations [6]. Chaos in induction motors fed by a PWM inverter was then investigated [7, 8] and chaos in DC motors was also reported and discussed [9]. Moreover, chaos in Permanent Magnet Synchronous Motors (PMSMs) was studied [10, 11] and chaos in switched reluctance drive systems was investigated [12]. It has been pointed out that the torque and the speed of a motor will oscillate in a random-like way subject to strong electromagnetic noise, so that the motor drive system is operated in a chaotic mode. This will cause degradation in performance and reliability of the system, not acceptable in industrial applications. Therefore, it is very important to control chaos in various motor-driven systems.

There are many chaos control methods suggested in theoretical research [13-17], but for a specific engineering plant there are quite few methods that can be practically used. For instance, the OGY method [17] is the most well-known method for chaos control but it is often found that there is no adjustable parameter for control when this method is applied to a real plant such as the Brushless DC Motor. Furthermore, it is

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 149 – 158, 2005.

not certain that one can actually achieve the desired control objective even though one can find an adjustable parameter because the approximation nature of the OGY method. It is well known that in a specific application, each chaos control method has some practical limitations, such as the choice of initial conditions, the feasibility of implementation, the design or computational complexity, and so on. Therefore, a specific and overall simple and effective control method for a particular application is often very desirable.

It was suggested that the entrainment and migration control method might be used to control chaos in PMSM [18]. But it requires an exogenous force to be input to the differential equation on the velocity, which is an external torque and is extremely difficult (if not impossible) to be used as a manipulating physical system variable in a real machine. In addition, the initial conditions and the target have to be very accurately computed and implemented, which are conceptually very difficult and practically impossible. Therefore, this method actually is undesirable in the PMSM control applications. It was suggested in our previous work that the time delay feedback control (TDFC) method could be used for chaos control in PMSM [19]. In doing so, the direct-axis or the quadrature-axis stator voltage can be used as the manipulating variable without requiring any exogenous force being inputted to the differential equation of the velocity. Although the TDFC method is acceptable for such applications, it has two shortcomings: first of all, it is difficult to determine the time delay constant for a specific target; secondly, the control objective must be an equilibrium point or a UPO (Unstable Periodic Orbit) of the system, which may not be the requirement of the application in interest.

In general, if a speed-adjustable motor-driven system is operated in the chaotic mode for some reason, it is desirable to move it out of the mode by eliminating chaos in the system and retaining the speed to the desired value. How can this goal be achieved? In this paper, a piecewise quadratic state feedback method is suggested for the purpose for a brushless DC motor model, where the direct-axis (or the quadrature-axis stator voltage) can be used as the manipulating variable and the control objective can be set to a constant value consistent with the requirement in the application while chaos is eliminated from the operating mode.

The piecewise quadratic state feedback control method was originally suggested for generating chaos [20, 21], with theoretical analysis performed [20] and physical circuit implemented [21]. It was found, through simulations, that the piecewise quadratic state feedback strategy can also be used to control chaos [22], where unfortunately the reason for the ability of controlling chaos was not explained. In order to apply this seemingly effective method to control physical plants, the chaos control mechanism will be analyzed in detail in this paper. Based on the analysis, control parameters can be selected according to the control objectives.

The rest of this paper is organized as follows: in section 2, chaos in a brushless DC motor model is reviewed. In section 3, the piecewise quadratic state feedback method is described and the control mechanism is analyzed in detail. In section 4, simulation results are presented to verify the correctness of the theoretical analysis. Finally, in section 5, some discussions and remarks are given to conclude the paper.

2 Chaos in a Brushless DC Motor Model

The transformed mathematical model of a brushless DC motor with smooth air gap is given by [6]

$$\begin{cases} \dot{\tilde{i}}_q = -\tilde{i}_q - \tilde{i}_d\tilde{\omega} + \rho\tilde{\omega} + \tilde{v}_q \\ \dot{\tilde{i}}_d = -\tilde{i}_d + \tilde{i}_q\tilde{\omega} - \tilde{v}_d \\ \dot{\tilde{\omega}} = \sigma\left(\tilde{i}_q - \tilde{\omega}\right) - \tilde{T}_L \end{cases} \tag{1}$$

where ρ and σ are system parameters, $\tilde{i}_d$, $\tilde{i}_q$, $\tilde{\omega}$ and $\tilde{T}_L$ are the transformed direct-axis, the transformed quadrature-axis for stator current, the transformed angular speed, and the transformed external load torque (including friction), respectively, and $\tilde{v}_d$ and $\tilde{v}_q$ are the transformed direct-axis and transformed quadrature-axis for stator voltage, respectively.

When $\tilde{v}_d = \tilde{v}_q = \tilde{T}_L = 0$, $\rho = 20$, $\sigma = 5.5$, $\tilde{i}_d(0) = 20$, $\tilde{i}_q(0) = 1$, $\tilde{\omega}(0) = -2$, system (1) has chaotic behaviors, with a chaotic attractor shown in Fig.1.

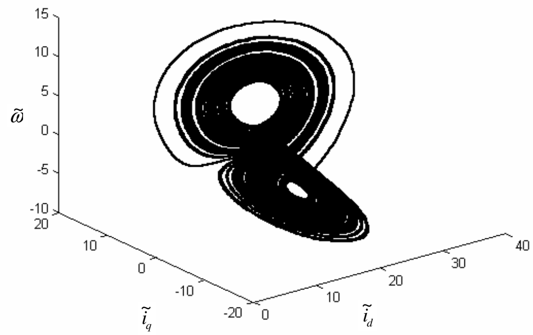


Fig. 1. The chaotic attractor in the brushless DC motor, when $\rho = 20$, $\sigma = 5.5$, $\tilde{v}_d = \tilde{v}_q = \tilde{T}_L = 0$

From Fig.1, one can see that the angular speed of the motor in the chaotic mode is oscillating in a random-like way, which is not acceptable in real applications. Details about chaos and bifurcation from this system under different conditions have been analyzed before [6], which will not be repeated here.

3 Controlling Chaos via Piecewise Quadratic State Feedback

Consider the following general n -dimensional chaotic system:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}(t), t) \tag{2}$$

where $\mathbf{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$ is the state vector and $\mathbf{F}$ is a smooth vector function. A controller $\mathbf{u}$ is designed for system (2), so that the controlled system becomes

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}(t), t) + \mathbf{u} \tag{3}$$

The following piecewise quadratic state feedback is employed as the controller:

$$\mathbf{u} = -\mathbf{Kcx}|\mathbf{cx}| \tag{4}$$

where $\mathbf{K} = [k_1 \quad k_2 \quad \cdots \quad k_n]^{\mathrm{T}}$ is the feedback gain vector and $\mathbf{c} = [c_1, c_2, \cdots, c_n]$ is the state observation vector. The objective, as discussed above, is to eliminate chaos in the brushless DC motor model (1) and stabilize the speed at a given value.

For simplicity, we first assume that the load $\tilde{T}_L = 0$ and the system parameters are fixed at $\rho = 20$ and $\sigma = 5.5$. If no control force is inputted, the equilibrium points of the system are found to be (- 4.3589, 19, - 4.3589), (0, 0, 0), (4.3589, 19, 4.3589), respectively.

Reexamine system (1). One can easily find that only $\tilde{v}_d$ and $\tilde{v}_q$ can be used as the manipulating variables for control, with free choices for $\mathbf{K}$ and $\mathbf{c}$. Note that this can be considered as one kind of output feedback control strategy, but is not as general as the case where $\boldsymbol{c}$ is a matrix. Now, we simply determine the controller structure with

$$\tilde{v}_q = 0, \tilde{v}_d = -k\tilde{i}_d|\tilde{i}_d| \tag{5}$$

Then, the controlled system becomes

$$\begin{cases} \dot{\tilde{i}}_q = -\tilde{i}_q - \tilde{i}_d\tilde{\omega} + \rho\tilde{\omega} \\ \dot{\tilde{i}}_d = -\tilde{i}_d + \tilde{i}_q\tilde{\omega} - k\tilde{i}_d|\tilde{i}_d| \\ \dot{\tilde{\omega}} = \sigma(\tilde{i}_q - \tilde{\omega}) \end{cases} \tag{6}$$

If $\tilde{i}_d^* < 0$, then there is no equilibrium in system (6); if $\tilde{i}_d^* \geq 0$, there are three equilibrium points given by

$$\left(\tilde{i}_q^*, \tilde{i}_d^*, \tilde{\omega}^*\right) = (0,0,0) \tag{7}$$

$$\begin{aligned} \tilde{i}_q^* &= \sqrt{\rho - 1 + k - 2k\rho + k\rho^2} \\ \tilde{i}_d^* &= \rho - 1 \\ \tilde{\omega}^* &= \sqrt{\rho - 1 + k - 2k\rho + k\rho^2} \end{aligned} \tag{8}$$

$$\begin{aligned} \tilde{i}_q^* &= -\sqrt{\rho - 1 + k - 2k\rho + k\rho^2} \\ \tilde{i}_d^* &= \rho - 1 \\ \tilde{\omega}^* &= -\sqrt{\rho - 1 + k - 2k\rho + k\rho^2} \end{aligned} \tag{9}$$

For equilibrium (7), one obtains the eigenvalues at equilibrium (7) as $\{8.73, -13.23, 1\}$; therefore, equilibrium (7) is unstable. For equilibrium points (8) and (9), the characteristic polynomial of the corresponding Jacobian matrix is

$$\lambda^3 + 38k\lambda^2 + 608k\lambda + 3971k + 7.5\lambda^2 + 25.5\lambda + 209 = 0 \tag{10}$$

To this end, one can study the locations of the eigenvalues versus the variation of k by using the equivalent “open-loop transfer function” and root locus. From (10), it is easy to have the equivalent “open-loop transfer function”, as follows:

$$G_o(s) = \frac{k\left(38s^2 + 608s + 3971\right)}{s^3 + 7.5s^2 + 25.5s + 209} \tag{11}$$

The root locus of (11) is shown in the left part of Fig.2. From root locus in Fig.2, it can also be seen that if the controller parameter k is greater than a critical value k_c, all eigenvalues have negative real parts; therefore, the equilibrium points (8) and (9) are both stable. According to the Routh-Hurwitz criterion, one can easily obtain $k_c \approx 0.01$. From (8) and (9), one can see that the equilibrium points are changed with the change of the feedback gain k. The speeds of the motor at the two changing equilibrium points versus the change of k are visualized in right part of Fig.2, respectively.

From Fig.2, it is clear that a stable equilibrium of the controlled system starts from the two symmetrical open-loop equilibrium points, and is changed smoothly with the increase of the feedback gain k. If $k > k_c$, the output speed will be stabilized at one of the two stable equilibrium points according to the basin in which the initial condition locates. Therefore, a proper feedback gain can be selected according to the expected speed value. However, the objective of controller (5) cannot be set as a value between – 4.3589 and 4.3589.

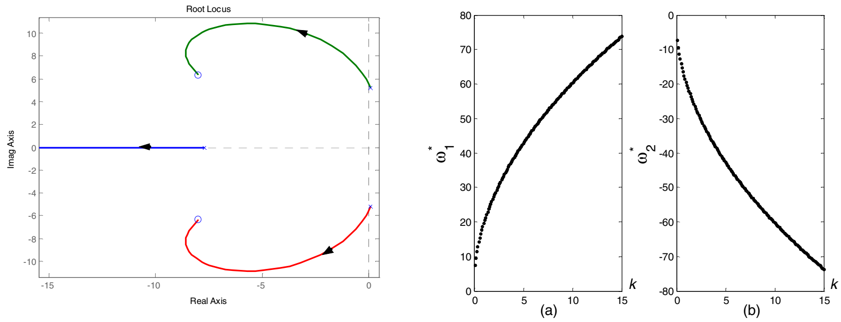


Fig. 2. The root locus of Eq.(11) and speeds of the motor at the changing equilibrium versus the change of k

To solve this problem, the following controller is used instead:

$$\tilde{v}_q = -l\tilde{i}_q\left|\tilde{i}_q\right|, \tilde{v}_d = 0 \tag{12}$$

Then, the controlled system becomes

$$\begin{cases} \dot{\tilde{i}}_q = -\tilde{i}_q - \tilde{i}_d\tilde{\omega} + \rho\tilde{\omega} - l\tilde{i}_q\left|\tilde{i}_q\right| \\ \dot{\tilde{i}}_d = -\tilde{i}_d + \tilde{i}_q\tilde{\omega} \\ \dot{\tilde{\omega}} = \sigma\left(\tilde{i}_q - \tilde{\omega}\right) \end{cases} \tag{13}$$

then the equilibrium points of the controlled system are

$$\left(\tilde{i}_q^*, \tilde{i}_d^*, \tilde{\omega}^*\right) = (0,0,0) \tag{14}$$

$$\begin{aligned} \tilde{i}_q^* &= \left(-l + \sqrt{l^2 + 4(\rho - 1)}\right)\Big/2 \\ \tilde{i}_d^* &= \left(-l + \sqrt{l^2 + 4(\rho - 1)}\right)^2\Big/4 \qquad \text{if } \tilde{i}_q^* > 0 \\ \tilde{\omega}^* &= \left(-l + \sqrt{l^2 + 4(\rho - 1)}\right)\Big/2 \end{aligned} \tag{15}$$

$$\begin{aligned} \tilde{i}_q^* &= \left(l - \sqrt{l^2 + 4(\rho - 1)}\right)\Big/2 \\ \tilde{i}_d^* &= \left(l - \sqrt{l^2 + 4(\rho - 1)}\right)^2\Big/4 \qquad \text{if } \tilde{i}_q^* < 0 \\ \tilde{\omega}^* &= \left(l - \sqrt{l^2 + 4(\rho - 1)}\right)\Big/2 \end{aligned} \tag{16}$$

It is easy to find that the equilibrium (14) is unstable. For equilibrium (15), it is also easy to derive the equivalent open-loop transfer function, as

$$G_o(s) = \frac{Q(s^2 + 3.25s - 2.75)}{s^3 + 7.5s^2 + 25.5s + 209} \text{ with } Q = l\sqrt{l^2 + 76} - l^2 \tag{17}$$

From (17), the root locus can be derived for Q, as shown in the left part of Fig.3. According to the Routh-Hurwitz criterion and some transformation, it is not difficult to derive if the following condition is satisfied:

$$l > l_c = 0.038 \tag{18}$$

Then all the eigenvalues of equilibrium (15) have negative real parts, i.e. the equilibrium becomes stable under control.

It can be easily verified that the characteristic polynomials for (16) are the same as that for (15). Therefore, the root locus for the linearized system at equilibrium (16) is same as that of (15). The speeds of the motor at the changing equilibrium versus the change of l are shown in the right part of Fig.3. Accordingly, the output speed will be stabilized to be at one of the stable equilibrium points according to the basin in which the initial condition locates, provided that $l > l_c$. At the same time, the output speed is stabilized to be either at (–4.3589, 0) or at (0, 4.3589). The three equilibrium points of the original system can never be stabilized because the feedback gain is equal to zero, so that there exists no control at these points.

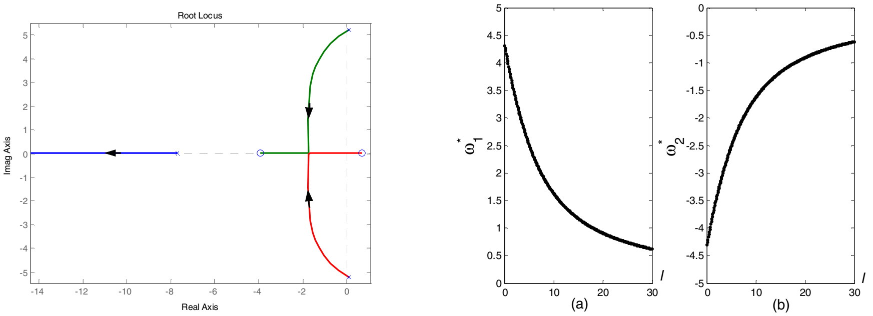


Fig. 3. The root locus of Eqs.(15) (16) and speeds of the motor at the changing equilibrium versus the change of l

When the load $\tilde{T}_L \neq 0$, the similar analysis can be conducted, which is omitted for the limited length. Some simulation will five to illustrate the method is effective when $\tilde{T}_L \neq 0$, which will be given in the following section together with the case discussed above.

4 Numerical Simulations

In order to verify the feasibility of the above-proposed control method, some simulations under different conditions have been carried out.

Firstly, the simulation on the case with $\tilde{T}_L = 0$ is reported. When the controller (5) is used, The controlled system is given by (6), with $\tilde{i}_q(0) = 10$, $\tilde{i}_d(0) = 20$, $\tilde{\omega}(0) = 0.1$, $k = 0.1$. According to the above analysis, when $\tilde{\omega}(t_0) \geq 0$, where t_0 is the moment when the controller is put into operation, the stable equilibrium is obtained from (8), at which $\tilde{\omega}^* = 7.423$; if $t_0 = 30$, the controlled output is shown in Fig.4(a). If controller (12) is used with $l = 2$, $t_0 = 40s$, and the same initial condition as above, the output angular speed is obtained as shown in Fig.4 (b). In this case, $\tilde{\omega}^* = -3.4721$. One can see that, when $\tilde{T}_L = 0$, the proposed control method can eliminate chaos effectively. The stable angular speed of the motor is relevant to the control parameter k and the initial conditions. Therefore, the parameter k can be calculated according to the control objective, and the controller can be put into operation at any time instant. Moreover, the control objective can take different constant values.

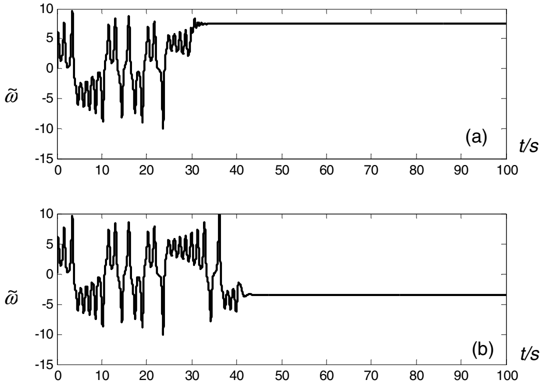


Fig. 4. The curves of the output angular speed, (a) when controller (5) is put into effect at $t_0 = 30s$, (b) when controller (11) is put into effect at $t_0 = 40s$ with $\tilde{T}_L = 0$

Secondly, simulation on the in case of $\tilde{T}_L \neq 0$ is reported. If controller (5) is used, with $\tilde{T}_L = 1.5$ and $k = 0.2$, $\tilde{i}_q(0) = -2$, $\tilde{i}_d(0) = -1$, $\tilde{\omega}(0) = 4$, and moreover if the controller (5) is put into effect at $t_0 = 50s$, then the output angular speed is obtained as shown in Fig.5 (a). In this case, $\tilde{\omega}(t_0) < 0$, and the stable output is –9.69. If controller (12) is used with $l = 3$, there exist two stable equilibrium points, with $\tilde{\omega}^* = -3.2018$ and $\tilde{\omega}^* = 3.0158$, corresponding to the initial conditions $\tilde{i}_q(t_0) < 0$ or $\tilde{i}_q(t_0) \geq 0$, respectively. In this simulation, $\tilde{i}_q(0) = 1$, $\tilde{i}_d(0) = -3$, $\tilde{\omega}(0) = -2$, $\tilde{T}_L = 0.5$, and the controller is put into effect at

$t_0 = 30s$. In this case, the output angular speed is obtained as shown in Fig.5 (b). From Fig.5, it is clear that the proposed controller is still effective even if the load is not zero.

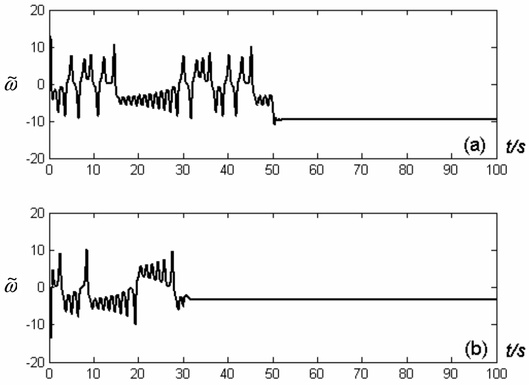


Fig. 5. The curves of the output angular speed (a) when controller (5) is put into effect at $t_0 = 50s$ with $\tilde{T}_L = 1.5$, (b) when controller (11) is put into effect at $t_0 = 30s$ with $\tilde{T}_L = 0.5$

5 Conclusions

Piecewise quadratic state feedback method has been proposed for controlling chaos in a brushless DC motor-driven model. The quadrature-axis (or the direct-axis voltage) is used as the manipulating variable, while the state variable is used in the feedback control law for control. Therefore, it is feasible for real brushless DC motors. The control objective can be a constant value. It is more reasonable considering practical requirements, as compared with the OGY and TDFC chaos control methods. Given a control objective, it is easy to calculate the controller parameters according to the given objective. The control objective is sensitive to the initial conditions of the system. This may be inconvenient for some applications therefore should be improved in the future. The equilibrium points of the uncontrolled system cannot be used as the control target, which is another drawback of the proposed method. Moreover, the points near these equilibrium points are difficult to reach because the feedback gain may be too small to have immunity to parametric perturbations and measurement noise, or too large to be realized by circuits.

References

1. J. H. B. Deane, D. C. Hamill: Instability, subharmonics and chaos in power electronics systems. IEEE Transactions on Power Electronics Specialists Conference Rec., (1989). Also in IEEE Transactions on Power Electronics, 5 (1990) 206-268
2. Y. Kuroe, S. Hayashi: Analysis of bifurcation in power electronic induction motor drive system. IEEE Power Electronics Specialists Conference Rec. (1989) 923-930

3. D. C. Hamill, J. H. B. Deane, D. J. Jefferies: Modeling of chaotic DC-DC converters by iterated nonlinear mappings. IEEE Trans. Power Electronics, 49 (1992) 25-36
4. Chi K. Tse , M. di Bernardo: Complex behavior in switching power converters. Proceeding of IEEE, 90 (2002) 768-781
5. A. Reatti, M. K. Kazimierczuk: Small-signal model of PWM converters for discontinuous conduction mode and its application for boost converter. IEEE Trans. on Circuits and Systems 1, 51 (2003) 65-73
6. N. Hemati: Strange attractors in brushless DC motors. IEEE Trans. on Circuit and Systems 1, 41 (1994) 40-45
7. Z. Suto, I. Nagy, E. Masada: Avoiding chaotic processes in current control of AC drive. IEEE Power Electronics Specialists Conference Rec. (1998) 255-261
8. Z. Cao,Z. Zheng: The chaos of nonlinear moving system for the synchronous motor. Proceeding of China Society Electronic Engineering (in Chinese), 18 (1998) 318-322
9. J. H. Chen, K. T. Chau, C. C. Chan: Analysis of chaos in current-mode-controlled DC drive systems. IEEE Trans. on Circuit and Systems 1, 47 (2000) 67-76
10. Z. Li, J. B. Park, Y. H. Joo, B. Zhang, G. Chen: Bifurcation and chaos in a permanent-magnet synchronous motor. IEEE Trans. on Circuit and Systems 1, 49 (2002) 383-387
11. Z. Jing, C. Yu , G. Chen: Complex dynamics in a permanent magnet synchronous motor model. Chaos Solution & Fractals, 22 (2004) 831-848
12. K. T. Chua, J. H. Chen: Modeling, analysis and experimentation of chaos in switched reluctance drive system. IEEE Trans. on Circuit and Systems 1, 50 (2003) 712-716
13. G. Chen, X. Dong: From Chaos to Order: Perspective, Methodology and Applications. World Scientific, Singapore, (1998)
14. H. P. Ren, D. Liu: Chaos control and anticontrol via a fuzzy neural network inverse system method. Chin. Phys. Lett., 19 (2002) 982-987
15. K. Barone, S. N. Singh: adaptive feedback linearizing control of chua's circuits. Int. J. of Bifurcation and Chaos, 12 (2002) 1599-1604
16. G Jiang, W. X. Zhang: chaos control for a class of chaotic system using PI type state observer approach. Chaos Solution & Fractals, 21 (2004) 93-99
17. E. Ott, C. Grebogi, A. Yorke: Controlling chaos. Phys Rev Lett, 64 (1990) 1196-1199
18. Z. Li, B. Zhang, Z. Mao: Entrainment and migration control of permanent magnet synchronous motor. Control Theory and Application, 19 (2002) 53-56
19. H. P. Ren, D. Liu, J. Li: Time delay feedback control of chaos in permanent magnet synchronous motor. Proceeding of China Society Electronic Engineering, 23 (2003)175-178
20. K. S. Tang, K. F. Man, G. Q. Zhang, G. Chen: Generating chaos via $x|x|$. IEEE Trans. on Circuit and Systems 1, 48 (2001) 636-641
21. G. Q. Zhang, K. F. Man and G. Chen: Generating chaos via a dynamical controller. Int. J. of Bifurcation and Chaos, 11 (2001) 865-869
22. F. H. Min, Z. Y. Xu, W. B. Xu: Controlling chaos via $x|x$. Acta Physica Sinica, 52 (2003) 1360-1365

Support Vector Machine Adaptive Control of Nonlinear Systems

Zonghai Sun[1], Liangzhi Gan[2], and Youxian Sun[1]

[1] National Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, 310027,Institute of Modern Control Engineering,Zhejiang University, Hangzhou, 310027, China
[2] Electric Engineering Department of Xuzhou Normal University,Xuzhou 221011, China
{zhsun, lzgan}@iipc.zju.edu.cn

Abstract. Support vector machine is a new and promising technique for pattern classification and regression estimation. The training of support vector machine is characterized by a convex optimization problem, which involves the determination of a few additional tuning parameters. Moreover, the model complexity follows from that of this convex optimization problem. In this paper we introduce the support vector machine adaptive control by Lyapunov function derivative estimation. The support vector machine is trained by particle filter. The support vector machine is applied to estimate the Lyapunov function derivative for affine nonlinear system, whose nonlinearities are assumed to be unknown. In order to demonstrate the availability of this new method of Lyapunov function derivative estimation, we give a simple example in the form of affine nonlinear system. The result of simulation demonstrates that the sequential training algorithm of support vector machine is effective and support vector machine adaptive control can achieve a satisfactory performance.

1 Introduction

Lyapunov-like techniques have long been used in nonlinear control. But even for completely known nonlinear systems, the construction of Lyapunov function is a challenging task. For unknown nonlinear systems, the construction of Lyapunov function has been solved only for special classes of systems. However, the problem becomes intractable for most unknown nonlinear systems. During the last decade, a vast of research effort was concentrated on problem of controlling highly uncertain and possibly unknown nonlinear dynamical systems, using the neural networks. Neural networks were applied to adaptive control by substituting the unknown system. But neural network has some disadvantages such as trapping in local minimum and worse generalization ability. Other new intelligent methods are needed to adopt to overcome these disadvantages.

Support vector machine (SVM) [1, 2] for pattern classification and regression estimation is an important methodology in the area of neural and nonlinear modeling. The theory of SVM is based on the idea of structural risk minimization (SRM). SVM maps input data into high dimensional feature space in which an optimal separating

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 159 – 168, 2005.

hyperplane is built. In many applications it provides high generalization ability and overcomes the overfitting problem suffered by the other learning technique such as neural network. The training of SVM is done by quadratic programming (QP) possessing a global solution, which overcomes the problem of local minima suffered by classical neural network. An interesting and important property of SVM's solutions is that one obtains a sparse approximation, in the sense that many elements in the QP solution vector are equal to zeros. SVM is a kernel-based approach which allows the use of linear, polynomial and radial basis function and others that satisfy Mercer's condition.

In this paper, we will discuss the SVM adaptive control of affine nonlinear systems by Lyapunov function derivative estimation. The SVM is used to estimate the derivative of an unknown Lyapunov function. In control the source of data is time-variant; we may adopt a sequential strategy for training SVM. Thus each item of data is used only once and then discarded. In the classic state-space model estimation by a Kalman filter setting, one defines a Gaussian probability density over the parameters to be estimated. The mean and covariance matrix are propagated using recursive update equations each time new data are received. The use of the matrix inversion lemma allows an elegant update of the inverse covariance matrix without actually having to invert a matrix for each sample. This is a widely studied topic, optimal in the case of linear Gaussian models. For nonlinear models the common trick is Taylor series expansion around the operating point, leading to the extended Kalman filter (EKF) [3, 4].

But Taylor series expansion leading to the EKF makes gross simplification of the probabilistic specification of the model. With the assumption of Gaussian probability density over the parameters to be estimated, the solutions of sequential SVM are not optimal. Particle filter [5-8] techniques provide a class of algorithms to address this issue. It is well suitable for applications involving on-line, nonlinear, and nongaussian signal processing. In this paper we focus on SVM, constraining ourselves to sequential training and its application in nonlinear control [9].

This paper is organized as follows. In Section 2 we discuss representation of state space for SVM. In Section 3 the particle filter realization of SVM is discussed. The SVM adaptive control of nonlinear systems by Lyapunov function derivative estimation will be discussed in Section 4. In Section 5 there is a simple example to illustrate the effectiveness of SVM adaptive control. In Section 6 we discuss the existing problems that should be solved in future for SVM control.

2 The Presentation of State Space for SVM

Suppose we have a set of training samples $\{x_i, y_i\}_{i=1}^{N}$, where $x_i \in R^n$, $y_i \in R$. For an integer i, the following regression model defines the relation between x_i and y_i,

$$y_i = w^T \varphi(x_i) + b\,,\ i = 1, \cdots, N \tag{1}$$

where b denotes the threshold value, and w denotes the weight. In order to estimate the y_i for x_i, we should solve the following optimization problem

$$\min \quad L = \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \tag{2}$$

$$\text{s.t.} \begin{cases} y_i - w^T \varphi(x_i) - b \le \varepsilon + \xi_i^*, & i = 1, \cdots, N \\ w^T \varphi(x_i) + b - y_i \le \varepsilon + \xi_i, & i = 1, \cdots, N \\ \xi_i, \xi_i^* \ge 0, & i = 1, \cdots, N \end{cases} \tag{3}$$

where C denotes balance term, and ξ_i, ξ_i^* denote the slack variables.

By forming the Lagrange, optimization problem (2), (3) can be translated into a problem of minimizing

$$\begin{aligned} \min L = \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) - \sum_{i=1}^{N} (\gamma_i \xi_i + \gamma_i^* \xi_i^*) - \sum_{i=1}^{N} \alpha_i (y_i - w^T \varphi(x_i) - b + \varepsilon + \xi_i) \\ - \sum_{i=1}^{N} \alpha_i^* (w^T \varphi(x_i) + b - y_i + \varepsilon + \xi_i^*) \end{aligned} \tag{4}$$

with respect to w, ξ_i, ξ_i^* and b. Setting the derivatives of L to zeros gives [10]:

$$\begin{cases} w = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) \varphi(x_i), & i = 1, \cdots, N \\ C = \alpha_i + \gamma_i, & i = 1, \cdots, N \\ C = \alpha_i^* + \gamma_i^*, & i = 1, \cdots, N \\ \sum_{i=1}^{N} \alpha_i = \sum_{i=1}^{N} \alpha_i^* \end{cases} \tag{5}$$

Where α_i, α_i^*, γ_i, γ_i^* are the Lagrange multipliers. There are two Lagrange multipliers for each training sample.

When a new data sample arrives, the output of SVM is given by:

$$\begin{aligned} y_i &= \sum_{j=1}^{N} (\alpha_j^* - \alpha_j) \varphi(x_j)^T \varphi(x_i) + b \\ &= \sum_{j=1}^{N} (\alpha_j^* - \alpha_j) K(x_i, x_j) + b, \ i = 1, \cdots, N \end{aligned} \tag{6}$$

where $K(x_i, x_j)$ denotes the kernel function which may be linear, polynomial, radial basis function and others that satisfy Mercer's condition [1, 2]. We will discuss the sequential SVM paradigm for regression estimation. We re-express equation (6) in terms of a moving window over the data and compute the output of SVM each time new data are received:

$$y_{k+1} = \sum_{j=0}^{L} (\alpha_{k+1,k-L+j}^* - \alpha_{k+1,k-L+j}) K(x_{k+1}, x_{k-L+j}) + b_{k+1} \tag{7}$$

We re-write equation (7) in an equivalent form:

$$y_{k+1} = [1\ K(x_{k+1}, x_{k-L})\ K(x_{k+1}, x_{k-L+1}) \cdots K(x_{k+1}, x_k)] \bullet \begin{bmatrix} b_{k+1} \\ \alpha^*_{k+1,0} - \alpha_{k+1,0} \\ \alpha^*_{k+1,1} - \alpha_{k+1,1} \\ \vdots \\ \alpha^*_{k+1,L} - \alpha_{k+1,L} \end{bmatrix} \tag{8}$$

In order to estimate b , $\alpha^* - \alpha$ sequential, we adopt following state-space Markovian representation:

$$\theta_{k+1} = \theta_k + \eta_k \tag{9}$$

$$y_{k+1} = C_{k+1}\theta_{k+1} + \delta_{k+1} \tag{10}$$

where $C_{k+1} = [1\ K(x_{k+1}, x_{k-L})\ K(x_{k+1}, x_{k-L+1}) \cdots K(x_{k+1}, x_k)]$,

$\theta_k = \left[b_k\ \alpha^*_{k,0} - \alpha_{k,0}\ \alpha^*_{k,1} - \alpha_{k,1} \cdots \alpha^*_{k,L} - \alpha_{k,L}\right]^T$, η_k and δ_{k+1} denote the process and measurement noise respectively.

3 The Particle Filter Realization of SVM

We assume η_k to be non-Gaussian distributed with covariance R and δ_{k+1} to be non-Gaussian distributed with covariance Q . For this type distribution of noise, the parameter θ may be estimated recursively by the particle filter.

The method of particle filter [11] provides an approximative solution for the problem of estimating recursively the posterior density function $p(\theta_k \mid y_{k-L:k})$, for the discrete time system on the form (9), (10). In other words it provides an approximative solution to the optimal recursive Bayesian filter given by

$$p(\theta_{k+1} \mid y_{k-L:k}) = \int p(\theta_{k+1} \mid \theta_k) p(\theta_k \mid y_{k-L:k}) d\theta_k \tag{11}$$

$$p(\theta_k \mid y_{k-L:k}) = \frac{p(y_k \mid \theta_k) p(\theta_k \mid y_{k-L:k-1})}{p(y_k \mid y_{k-L:k-1})} \tag{12}$$

For expression on $p(\theta_{k+1} \mid \theta_k)$ and $p(y_k \mid \theta_k)$ in (11) and (12) we use the known probability densities p_{η_k} and p_{δ_k} . The main idea is to approximate the (12) with a sum of delta-Dirac functions located in the samples, $\theta^{(i)}_k$. Using the importance weights the posterior can be written as

$$p(\theta_k \mid y_{k-L:k}) \approx \sum_{i=1}^{N} \overline{q}^{(i)}_k \delta(\theta^{(i)}_k - \theta_k) \tag{13}$$

where $\delta(\cdot)$ denotes the delta-Dirac function. A straightforward way to recursively update the particles $\theta_k^{(i)}$ and weights $\overline{q}_k^{(i)}$ is given by

$$\theta_{k+1}^{(i)} \sim p(\theta_{k+1} \mid \theta_k^{(i)}),\ i=1,\cdots,M\ , \tag{14}$$

$$\overline{q}_{k+1}^{(i)} = \frac{p(y_{k+1} \mid \theta_{k+1}^{(i)})\overline{q}_k^{(i)}}{\sum_{j=1}^{M} p(y_{k+1} \mid \theta_{k+1}^{(i)})\overline{q}_k^{(j)}},\ i=1,\cdots,M\ , \tag{15}$$

initiated at time $k=0$ with

$$\theta_0^{(i)} \sim p(\theta_0),\ \overline{q}_0^{(i)} = \frac{1}{N},\ i=1,\cdots,M \tag{16}$$

The discussion above is formalized in algorithm below. All the operations in the algorithm have to be done for all particles $\theta_k^{(i)}$.

Algorithm: The particle filter

(1) Sample $\theta_0 \mid_{-1} \sim p(\theta_0)$;

(2) Calculate the weights $q_k = p(y_k \mid \theta_k)$ and normalize, i. e., $\overline{q}_k^{(i)} = \frac{q_k^{(i)}}{\sum_{j=1}^{M} q_k^{(j)}}$, $i=1,\cdots,M$;

(3) Generate a new set $\{\theta_k^{(i*)}\}_{i=1}^{M}$ by resampling with replacement M times from $\{\theta_k^{(i)}\}_{i=1}^{M}$, with probability $P\{\theta_k^{(i*)} = \theta_k^j\} = \overline{q}_k^j$;

(4) Predict (simulate) new particles, i.e., $\theta_{k+1|k}^{(i)} = \theta_{k|k}^{(i)} + \eta_k^{(i)}$, $i=1,\cdots,M$;

(5) Increase k and iterate from step 2.

4 Sequential SVM Control of Nonlinear Systems

We consider an affine nonlinear system [12]

$$\dot{x} = f(x) + G(x)u \tag{17}$$

where $u \in R^m$ is the control input, and the state $x \in R^n$ is assumed completely measurable, and $f(x)$, $G(x)$ are a continuous, locally Lipschitz vector fields. The objective is to enable the output x to follow a desired trajectory x_r. For the x_r, we also assume that $\dot{x}_r$ is bounded. We define the tracking error as

$$e = x - x_r\ . \tag{18}$$

According to (17) and (18), we may obtain

$$\dot{e} = f(x) + G(x)u - \dot{x}_r \tag{19}$$

Assumption 1[12]: *the solution of system (19) is uniformly ultimately bounded with respect to an arbitrarily small neighborhood of* $e = 0$.

From *Assumption 1,* there exits an arbitrarily unbounded Lyapunov function $V(e)$ and a control input such that [12]

$$\dot{V}(e) = \frac{\partial V(e)}{\partial e}(f(x) + G(x)u - \dot{x}_r) \le 0 \tag{20}$$

Sontag [13] has provided an explicit formula for stabilizing systems of the form (17) whenever V, f, G are known.

In most applications, f, G are unknown and the construction of V is a hard problem, which has been solved for special classes of systems. To overcome the highly uncertainty about (17) and provide a valid solution for our problem. The approximation capability of SVM is adopted. Without losing generality, we substitute the unknown term in (20) with

$$a(x,e) = \frac{\partial V(e)}{\partial e} f(x) \tag{21}$$

$$b(x,e) = \frac{\partial V(e)}{\partial e} G(x) \tag{22}$$

$$c(\dot{x}_r, e) = \frac{\partial V(e)}{\partial e} \dot{x}_r \tag{23}$$

In what follows, the SVM is used to approximate the unknown functions $a(x,e)$, $b(x,e)$, $c(\dot{x}_r, e)$. So the following approximation holds,

$$a(x,e) = w_f{}^T \varphi_f(z) + d_f + \delta_f \tag{24}$$

$$b(x,e) = w_g{}^T \varphi_g(z) + d_g + \delta_g \tag{25}$$

$$c(\dot{x}_r, e) = w_r{}^T \varphi_r(v) + d_r + \delta_r \tag{26}$$

where $z = [x^T \quad e]^T$, $v = [\dot{x}_r{}^T \quad e]^T$; δ_f, δ_g, δ_r denote the approximation error.

We have the following lemma to present an ideal control, u^*, such that under the ideal control, the output x can follow the objective trajectory x_r asymptotically.

Lemma 1. *For the system (17) satisfying Assumption 1, if the ideal control is designed as*

$$u^* = -(a(x,e) - c(\dot{x}_r, e) + k(t)|e|)/b(x,e) \tag{27}$$

where $k(t) > 0$, *for all* t, *is a design parameter, then the filtered tracking error converges to zeros.*

Let us claim for a moment that

$$b(x,e)=\begin{cases} b(x,e) & |b(x,e)|>\zeta \\ \zeta & 0\le b(x,e)\le\zeta \\ -\zeta & -\zeta\le b(x,e)<0 \end{cases} \tag{28}$$

where ζ denotes a small enough positive design constant.

The lemma can be proven easily by substituting (21)-(23), (27) into (20).

5 Numerical Simulation

In this section we give a simple example to illustrate the effectiveness of the proposed SVM adaptive control of nonlinear systems by filtered training algorithm of SVM. Consider the following scalar system [12]

$$\dot{x}=2x-x^3+x^2u \tag{29}$$

The problem is to control (29) to track the bounded desired trajectory $x_r(t)=1-e^{-3t}$. The initial condition is $x(0)=0.2$.

The unknown functions $a(x,e)$, $b(x,e)$, $c(\dot{x}_r,e)$ are estimated by sequential SVM. The kernel functions are all the Gaussian functions. The SVM is trained online by particle filter with $L=40$, $k(t)=10$, and the number of particles is chosen as $M=200$ in simulation. Fig. 1 shows that the output x can track the desired trajectory x_r effectively. Fig. 2 is the tracking error curve. The dot line denotes desired trajectory x_r, while the solid line denotes the output curve x in Fig. 2, Fig. 3, Fig. 4 are simulation results of [12]. Comparing Fig. 1, Fig 2 with Fig. 3, Fig. 4, we may know that the convergence rate of tracking error of adaptive SVM control is bigger than that of adaptive neuro-control.

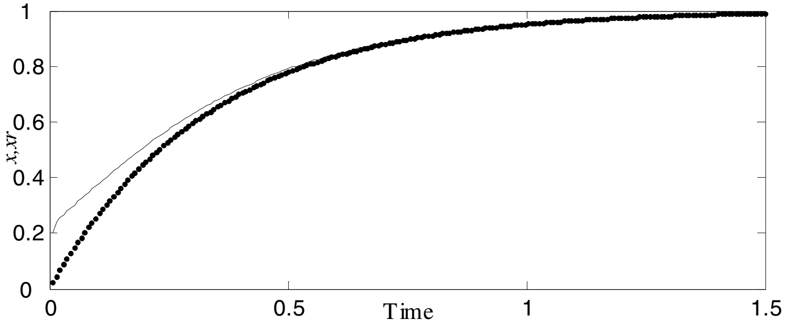


Fig. 1. Tracking performance of SVM adaptive control design

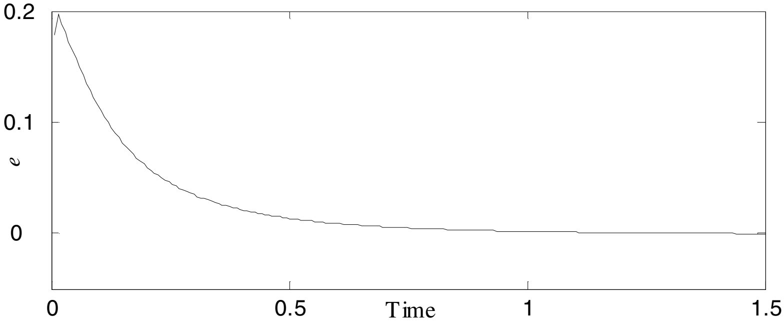


Fig. 2. Tracking error of SVM adaptive control design

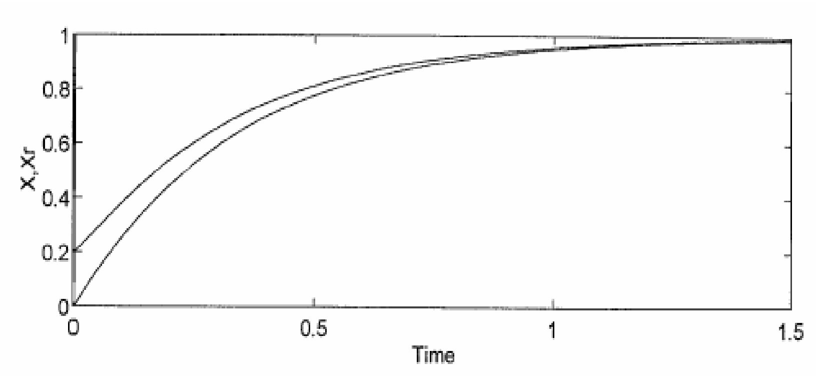


Fig. 3. Closed loop system performance of adaptive neural control design

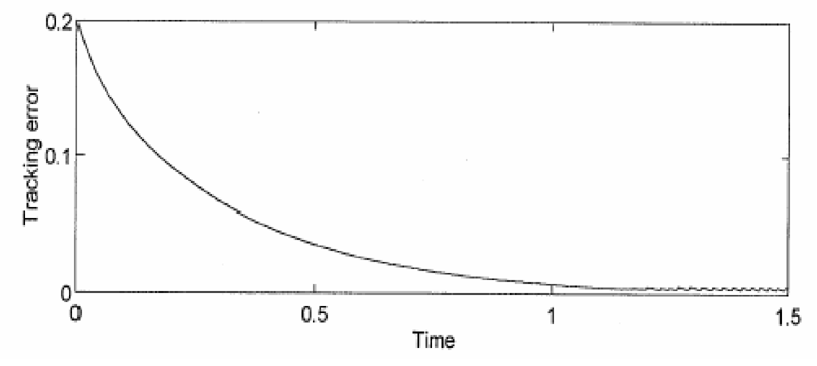


Fig. 4. Tracking error of adaptive neural control design

.

The numerical simulation result above shows good transient performance and the tracking error is small. This demonstrates that the sequential SVM controller can achieve a satisfactory performance.

6 Discussion and Conclusion

In this paper we discussed sequential SVM for regression estimation and sequential SVM control of nonlinear systems via Lyapunov function derivative estimation. Firstly we provided the representation of state-space for SVM. The sequential SVM is realized by particle filter. The advantage of the proposed method is that the computational complexity decreases and thus the on-line training and control become feasible. Finally we apply the sequential SVM to estimate the Lyapunov function derivative for nonlinear adaptive control. The simulation result demonstrates that this method is feasible for Lyapunov function derivative estimation. Interesting questions for future research include: (1) the method of training SVM in this paper decreases the generalization ability, then how to train the SVM on-line without decreasing the generalization ability should be studied; (2) in this paper we discussed sequential SVM control of nonlinear systems via Lyapunov function derivative estimation. In what follows, the approximation property of SVM needs to be studied.

Acknowledgment

China 973 Plan supported this work under grant NO. 2002CB312200.

References

1. Vapnik, V. N.: Statistical learning theory, John Wiley and Sons. New York (1998)
2. Vapnik, V. N.: The nature of statistical learning theory. Springer-Verlag, New York (1995)
3. De Freitas, J. F. G., Niranjan, M., Gee A. H., Doucet A.: Sequential monte carlo methods to train neural network models. Neural computation, 4(2000)955-993
4. Kalman, R. E, Bucy, R. S.: New results of in linear filtering and prediction theory. Transaction of the ASME (Journal of Basic Engineering), 83D(1961)95-108
5. Storvik, G.: Particle filters in state space models with the presence of unknown static parameters. IEEE Transactions on Signal Processing, 50(2002)281-289
6. Arnaud, D., De Freitas, N., Murphy, K., Russell, S.: Rao-Blackwellised particle filtering for dynamic Bayesian networks. In Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence. Stanford (2000)176-183
7. Liu, S., Chen, R.: Sequential monte carlo methods for dynamic systems. Journal of the American Statistical Association, 443(1998)1032-44
8. Crisan, D., Doucet A.: A survey of convergence results on particle filtering methods for practitioners. IEEE Transactions on Signal Processing, 3(2002)736-746
9. Sun, Z. H.: Study on support vector machine and its application in control. PhD. Thesis. Zhejiang University, Hangzhou, China (2003)

10. Wolfe, P.: A duality theory for nonlinear programming. Quarterly of Applied Mathematics, 19(1961)239-244
11. Arulampalam, S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for on-line nonlinear /non- Gaussian Bayesian tracking. IEEE Trans. Signal Process, 2 (2002)174-189
12. George, A. R.: Stable adaptive neuro-control design via Lyapunov function derivative estimation. Automatica, 37(2001)1213-1221
13. Sontag, E. D. A.: “universal” construction of Artstein’s theorm on nonlinear stabilization. Systems Control of Letters, 131(1989)117-123

Improved GLR Parsing Algorithm

Miao Li, ZhiGuo Wei, Jian Zhang, and ZeLin Hu

Institute of Intelligent Machines, Chinese Academy of Sciences,
P.O.Box 1130, Hefei Anhui 230031, P.R.China
wzhg810923@eyou.com

Abstract. Tomita devised a method of generalized LR(GLR) parsing to parse ambiguous grammars efficiently. A GLR parser uses linear-time LR parsing techniques as long as possible, falling back on more expensive general techniques when necessary. In this paper, the motivation of adopting the GLR parsing algorithm to construct parsers for programming languages is presented. We create a multi-level scheme to fasten the GLR parser. We introduce runtime control mechanisms to the GLR parser to invoke semantic actions attached to grammar rules. The algorithm has been implemented in Development Expert Tools (DET), a compiler which is designed by Institute of Intelligent Machines, Chinese Academy of Sciences, at Hefei. Experiments show that the speed of our GLR parser is comparable to LALR(1) parsers when parsing deterministic programming languages.

1 Introduction

Tomita observed that grammars for natural languages were mostly LR, with occasional ambiguities [1]. Not surprisingly, parsers which deal with unambiguous grammars can run much faster than parsers for ambiguous grammars. This is crucial when one considers that the speed of input recognition is often highly visible to users. As a result, most artificial languages have unambiguous grammars. However, applications such as natural language understanding are rarely able to choose a convenient grammar, so there is still a need for fast parsers for ambiguous grammars.

And with the development and diversifying of programming languages, LALR(1) parsing algorithm begins to reveal its weakness on designing a parser for a language. When designing and describing the grammar of a programming language, the most important thing is to make users understand the grammar. Hence, grammars that we could see on the reference book for programming languages are usually not LALR(1) grammars, but ambiguous grammars. For C++ and some other domain-specific programming languages, the experience has told us that attempting to adapt their grammars to LALR(1) is infructuous or very difficult. The reason is that once the grammar is redesigned, the language that described by it maybe also changed. In order to parsing these ambiguous grammars, Tomita developed the Generalized LR(GLR)parsing. As known, a GLR parser uses linear-time LR parsing techniques as long as possible, falling back on more expensive general techniques when necessary.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 169 – 178, 2005.

2 GLR Parsing Algorithm

A configuration of a LR parser is a pair whose first component is the stack content, and whose second component is the unexpended input, where the bold-faced part represents the stack and the portion after the comma represents the input stream:

$$(S_0 X_1 S_1 X_2 S_2 \ldots X_m S_m ,a_i\ a_{i+1} \ldots a_n \#)$$

During the process of parsing, the action of the parser follows the underlying algorithm in Fig1 :

```
BEGIN:
      $a_i$ = currentInputSymbol( );
      IF Action($S_m$ , $a_i$ ) = "Shift S" THEN the parser executes a
            move entering the configuration :
               ( $S_0 X_1 S_1 X_2 S_2 \ldots X_m S_m\ a_i S,\ a_{i+1} \ldots a_n \#$ )
            with $a_{i+1}$ as the new current symbol;
      IF Action($S_m$ ,$a_i$ )= "Reduce $A \rightarrow \beta$" THEN the parser executes
            a reduce by entering the configuration :
               ( $S_0 X_1 S_1 X_2 S_2 \ldots X_{m-r} S_{m-r} A\ S ,a_i\ a_{i+1} \ldots a_n \#$ )
         where Goto( $S_{m-r}$ ,A)=S ,and r is the length of $\beta$, the
         right-hand side of the production;
      IF Action($S_m$ , $a_i$ ) = Accept THEN the parse is complete;
      IF Action($S_m$ , $a_i$ ) = Error THEN report the error;
END.
```

Fig. 1. Algorithm that the action of the parser follows during the process of parsing

Though many grammars could be parsed by using this algorithm, LR(0)parsers are not all that useful because in many cases, multiple items can appear in a Action-table entry .It is possible to have shift-shift ambiguity ,reduce-reduce ambiguity, or shift-reduce ambiguity. LR(0) parsers can be improved by incorporating look-ahead information into the sets of states in order to split conflict states into deterministic states. Compiler theory uses SLR(1) and LR(1) parsers to get additional power. Unfortunately, using arbitrary look-ahead information causes the DFA to grow exponentially in size.

How to deal with the conflicts that are stated above? Tomita created the GLR parsing. The idea is to handle multiple entries non-deterministically, using pseudo-parallelism (breadth-first search) and maintaining a list of stacks called a Stack List. When a process encounters a multiple entry it splits into several processes, replicating it stack. When a process encounters an error entry, its process is terminated and removed from the stack list. All processes are synchronized so they shift a word at the same time. Thus if a shift action occurs for a process before the others that are ready to shift, it must wait.

This cycle of a parser process starting other yields a wholly impractical algorithm. The time spent on making copies of parser stacks could enormous, not to mention the potentially exponential growth of the number of processes [2].To address this, two important optimizations are made.

First, a new process need not have a copy of its parent's attack. All sub-processes can share a common prefix of a stack. From an implementation perspective, elements of the stack can all contain pointers to point to the previous element of the stack. Then, multiple stack elements can point to a common prefix.

Second, there are a finite number of automaton states the parser can be in. Several processes may be in the same state, albeit they may have different stack contents. A set of processes that are in the same can merge their stacks together, leaving one resulting process. In a LR parser, its current state is the topmost state on the stack. So to merge N stacks, one would remove the top node from each -- they must all have the same state number s -- and create one node with states s that points to the remainder of the N stacks.

The result of these optimizations is called a Graph-Structured Stack.Table1 summaries the difference between a GLR parser and a LR parser.

Table 1. The difference between a Generalized LR parser and a traditional LR parser

Features	GLR parser	LR parser
Parse table	LR(0) tables allowing conflicts	LALR(1) table free of conflicts
Parse stack	Graph-Structure Stack	Liner
Semantic value	No semantic value	In stack node
Parsing algorithm	GLR algorithm	LR algorithm
Typical output	Parse forest	Invoking actions or parse tree
Look ahead	Infinite	Limited ,usually one
Disambiguation	Parser generation, parsing time	Parser time

3 Investigating Into the GLR Parsing Algorithm

We give a grammar G and input string #abaca#. Then let us discuss the GLR Parsing process in detail. Grammar G=(N,∑,P,S),in which N={ A }, ∑={ a,b,c },S=A, the productions of set P are:

$$r_1: S \rightarrow A\#$$

$$\begin{aligned} r_2 &: A \rightarrow a \\ r_3 &: A \rightarrow AbA \\ r_4 &: A \rightarrow AcA \end{aligned} \qquad (1)$$

The LR(0) parse table for grammar G is shown in Table 2, in which S_i means shift the symbol i, R_i means using the production i to reduce the handle.When String #abaca# is shifted in by the parser symbol by symbol, the actions of the parser is shown in Table3.

For further discussion, we use the expression $LAR(G)=<K,V,P,S_0\,,Goto,Action>$ to represent the LR(0) automaton of the Grammar $G=(N,\sum,P,S)$ [2], use the state numbers to represent the state nodes and the symbol $\omega=(\ a_1\,,\ldots,a_i\,,\ldots,a_n\ )$ represents the input string, the $S_{top}=\{\ s_1\,,\ \ldots\ ,\ s_j,\ldots\ ,s_m\}$ represents the top node set of the active stack .

Table 2. LR(0) Parse Table for Grammar G

State	Action				Goto	
	a	b	c	#	S	A
0	S_2					1
1		S_5	S_4	S_3		
2		r_2	r_2	r_2		
3	---------- Accept --------					
4	S_2					6
5	S_2					7
6		$S_5\,r_4$	$S_4\,r_4$	r_4		
7		$S_5\,r_3$	$S_4\,r_3$	R_3		

When using GLR algorithm to parsing input strings, we summaries the properties of the algorithm.

No1, for a active stack top node sj ∈Stop and the current input symbol ai ,if the action Reduce(p:A→a) ∈Action(sj , ai),then the algorithm will search and save all reduction paths which start from node sj and contain k+1 state nodes. Here k is the length of production p. For every reduction path, the algorithm executes an reduction.

No2,reductions of these active stacks are synchronized so they shift a word at the same time. Thus if a shift action occurs for a process before the others that are ready to shift, it must wait.

No3,for elements in the active stack top node set Stop ={ s1 , … , sj,… ,sm} ,if i≠j, then si≠sj .

No4, for a active stack top node sj ∈Stop ,if s∈Stop and there is no directed arc between node s and s_ before the algorithm executes an action Reduce(p:A→a) by the reduction path (sj ,tj ,…, sj-k+1 , tj-k+1 ,s_), then the newly added directed arc between them maybe a portion of some reduction path starting from node s. Here |a|=k and s=Goto(s_,A).

No5, for a active stack top node sj ∈Stop ,if s∈Stop and there exists a directed arc between node s and s_ before the algorithm executes an action Reduce(p:A→a) by

Table 3. Actions of the Parser

State	Symbol Stack	Input	Action
(1) 0	#	abaca#	Shift(2)
(2) 02	#a	baca#	Reduce(r_2);Goto(0,A)=1
(3) 01	#A	baca#	Shift(5)
(4) 015	#Ab	aca#	Shift(2)
(5) 0152	#Aba	ca#	Reduce(r_2);Goto(5,A)=7
(6) 0157	#AbA	ca#	Branch1:Reduce(r_3);Goto(0,A)=1 Branch2:Shift(4)
(7) 01 (7) 0157	#A #AbA	ca# ca#	Shift(4) Shift(4)
(8) 01 4 (8) 01574	#Ac #AbAc	a #	Shift(2)
(9) 01 42 (9) 015742	#Aca #AbAca	#	Reduce(r_2);Goto(4,A)=6
(10) 01 46 (10) 015746	#AcA #AbAcA	#	Reduce(r_4);Goto(0,A)=1 Reduce(r_4);Goto(5,A)=7
(11) 01 (11) 0157	#A #AbA	#	Shift(3) Reduce(r_3);Goto(0,A)=1
(12) 01 (12) 01	#A #A	#	Shift(3) Shift(3)
(13)			Accept

the reduction path (sj ,tj ,…, sj-k+1 , tj-k+1 ,s_), we could say that the shifted substring ν∈ω is a prefix of a ambiguous sentence form and the nonterminal A is an ambiguous nonterminal of the Grammar. Hence, there exists two different derivation sequences π1 and π2 from aAβ to νβ (Here αand β are terminals):

S==>* aAβ==>π1νβ and S==>* aAβ==>π2νβ

4 Optimizing GLR Parsing Algorithm

In this paper, we present a 3-level scheme to optimize GLR algorithm: Optimizing the parsing-table, optimizing the operation of the Graph-Structured Stack, and the pooling technique.

4.1 Optimizing the Parsing-Table

Billton and Lang discovered that LR(0) and GLR(0) could parse the natural language[3].In their book, illustrations about the GLR algorithm are mostly based on LR(0) parsing table. The reason is that the ambiguity of natural is global ambiguity, and the conflicts in the LR(0) parsing table could not be resolved by adding the number of look-ahead symbols. Theoretically, GLR parser could use all kind of LR parsing tables. Thanks to Parr TJ's effort, we now know that in all parsing processes that need look-ahead symbols, 98% of them only need one look-ahead symbol to make decision[4].So, When designing GLR parser we take advantage of LALR(1) table, but not LR(0) table, which are widely used in natural language processing.

Through this modification, the parser will encounter less conflicts and the Graph-Structured Stack will be liner, which decreases the cost on maintaining and operating the Graph-Structured Stack.

4.2 Optimizing Operations on the Graph-Structured Stack

In the third section of the paper we have pointed out that the resource spent on splitting the Graph-Structured Stack, searching and maintaining the reduction path is enormous. The operations applied on the Graph-Structured Stack are too complex.

```
conflictResolving(){
/*Interactive conflict resolving*/
if(Action(state,lookahead)>1)
/*Callback user-defined routine */
begin
selected = UsrSolving(state,lookahead);
case selected
begin
/* select a parsing action */
Ri :DoReduce( rule i);
Sj :DoShift(state i);
Default: /*Perform standard GLR algorithm */
DoAll(state,lookahead);
end;
end ;

}
```

In order to decrease the splitting actions of the Graph-Structured Stack, we adapt the GLR parsing algorithm into a interactive algorithm. When encountering conflicts,the interactive algorithm callback the user-defined conflict-resolving routine, using the conflict information as parameter. Conflict-resolving routine could depend on the parsed grammar elements in the symbol stack and the semantic information ,such as the symbol table and the type of the symbols, to choose an parsing action from those that conflict with each other. If there is no Conflict-resolving function, or the information that the Conflict-resolving function needs is out of hand, then we still use the Standard GLR algorithm. The interactive algorithm for resolving conflicts is presented in function conflictResolving().

GLR parsing algorithm could parse the unambiguous input string in linear time. But because the GLR parser has a more complex data structure than LR parser, its constant coefficient is larger than LR parser's. In order to minish the constant coefficient, we class the shift actions into two categories: deterministic shift and standard GLR shift, and class the reduction action into three categories: deterministic reduction, simple GLR reduction and standard GLR reduction.

Before the parsing program executes a reduction action on the parse stack top, we first have to check the Graph-Structure Stack to see if it has the underlying structural characteristics.

Characteristic 1, the Graph-Structure Stack has only one active stack top node;

Characteristic 2, for all nodes except the last node that are on a reduction path which starting from a active stack top node, their outdegree are all 1.

If the two conditions all exist, the parsing program executes deterministic reduction. If only condition 2 is satisfied, the parsing program executes simple GLR reduction. If both conditions are not satisfied, then the parsing program executes standard GLR reduction.

If it is time for shifting and condition 1 exists, then the parsing program executes a deterministic shift. Otherwise, the parsing program executes a standard GLR shift.

When executing deterministic shift, the program only have to add a new node on top of the stack, and not have to search and compare the newly added states in the active stack top nodes. When executing deterministic reduction, the program only need to pop up k state nodes(k is the length of the right-hand side of the production) and add a new active state node on top of the stack. There is no need to search and compare the newly added states in the active stack top nodes, and the program does not have to search and save the reduction paths.

4.3 Pooling Technique

We have presented in section 3 that there are no identical nodes in the active stack top nodes of the Graph-Structural Stack. Hence, the active stack top nodes could be stored in a Hash table, with the number of the node as the key word. The capability of the GLR algorithm will be improved significantly with the using of Hash table.

During the process of maintaining the Graph-Structural Stack, we have to create and destroy a large number of dynamic objects. We have known that to create and to destroy dynamic objects cost too much resource. So we use object pooling technique to manage these objects, aiming at improve the effectiveness of the GLR algorithm. Object pooling allows the sharing of instantiated objects. Since different processes don't need to reinstantiate certain objects, there is no load time. And since the objects are returned to the pool, there is a reduction in garbage collection.

Our object pool design(In Java language) is generic enough to handle storage, tracking, and expiration times, but instantiation, validation, and destruction of specific object types must be handled by subclassing. Now that the basics out of the way, lets jump into the code. This is the skeletal object:

```
Public abstract class ObjectPool {

Private longexpirationTime;
   private Hashtable locked, unlocked;
   abstract Object create();
   abstract boolean validate( Object o );
   abstract void expire( Object o );
   synchronized Object checkOut(){...}
   synchronized void checkIn( Object o ){...}
}
```

Ultimately, the object pool would allow the subclass to specify the initial size of the hashtables along with their growth rate and the expiration time, but I'm trying to keep it simple for the purposes of this article by hard-coding these values in the constructor.

```
ObjectPool ( )  {
            expirationTime = 30000; // 30 seconds
            locked = new Hashtable();
            unlocked = new Hashtable();
         }
```

The checkOut( ) method first checks to see if there are any objects in the unlocked hashtable. If so, it cycles through them and looks for a valid one. Validation depends on two things. First, the object pool checks to see that the object's last-usage time does not exceed the expiration time specified by the subclass. Second, the object pool calls the abstract validate( ) method, which does any class-specific checking or reinitialization that is needed to re-use the object. If the object fails validation, it is freed and the loop continues to the next object in the hashtable. When an object is found that passes validation, it is moved into the locked hashtable and returned to the process that requested it. If the unlocked hashtable is empty, or none of its objects pass validation, a new object is instantiated and returned. The function checkOut( ) is the most complex method in the ObjectPool class, it's all downhill from here.

The checkIn( ) method simply moves the passed-in object from the locked hashtable into the unlocked hashtable. The three remaining methods are abstract and therefore must be implemented by the subclass.

5 Capability of the Optimized GLR Algorithm

We compared the optimized GLR parser with the LALR(1) parser ,which is generated by the Bison. Because the grammar of Java could be adapted into LALR(1) grammar and Java possesses all typical grammar characters of programming language, we use the Java programming language to test the two Parser. The CPU is AMD2000, RAM is 512MKB, the operating system is Windows/XP. The Java1.4 has 280 BNF productions and 317 LALR(1)automaton states. Bison reports 235 conflicts. The adapted Java grammar has 350 BNF productions and 448 LALR(1) automaton states. Bison reports 0 conflict. Though the adapted grammar resolves the conflicts, but the dimension of the grammar and the LALR(1) automaton state number are increased. The data acquired from the experiment is recorded in Table4, in which the unit of

Table 4 . Comparison of Optimized GLR and LALR(1)

No	No.of lines	No.of tokens	GLR(clock)	LALR(1)(clock)	GLR/LALR(1)
1	43	103	309347	426407	0.725
2	321	1155	4847486	1631176	2.972
3	218	1310	5698296	1668174	2.972
4	528	2376	9815220	2772309	3.540
5	1285	5410	24398177	6847486	3.563
6	2088	11288	62278403	9101114	6.843
7	9916	53495	303525993	58468247	5.191
8	16007	88943	461722165	85804274	5.381
9	32928	128397	969437820	165329119	5.864

time is the clock number of CUP. We could see that the two parsers' time complexities are approximately linear. The LALR(1) parser are 2~5 times faster than the optimized GLR. According to document [5], the standard GLR parser are 10~100 times slower than the LALR(1).

6 Summaries and Conclusions

Current researches in China about GLR algorithm focus on processing natural languages. Document[6] takes advantage of the controlling structure of GLR parser to computing probability, and presents how to using the probability to select the best parsing result. Document[7] illustrates how to process the phenomena of extra grammaticality in natural language parsing. Until now in China, there is no paper on applying GLR algorithm on parsing programming languages.

This paper illustrated the GLR parsing algorithm in detail. We analyzed the running mechanism of GLR parsing algorithm and also pointed out its weaknesses. In order to improve to performance of the GLR parsing algorithm, we designed the 3-level scheme: Optimizing the parsing table, Optimizing Operations on the Graph-Structured Stack, and using the pooling technique. We introduced runtime control mechanisms to the GLR parser to invoke semantic actions attached to grammar rules. Our improved generalized LR(GLR) parsing could not only parse ambiguous grammars efficiently, but also performs much faster than the standard GLR parser, which has been provided by experiments.

References

1. M.Tomita: An Efficient Augumented-Contxt-Free Parsing Algorithm. Computational Lingustics, Vol.13. (1987) 31-46
2. Gao ZY, Jin MZ: The Theory and Construction of Compilers. Beijing University of Aeronautics and Astronautics Press, Beijing (1990) 100-127

3. Billot S, Lang B: The structure of shared forest in ambiguous parsing. In: Proc. Of the 6th Int'l Computational Ligustics(ACL-89) (1989) 143-151
4. Parr TJ: Obtaining practical variants of LL(K) and LR(k) for k>1 by splitting the atomic k-tuple[Ph.D.Thesis].West Lafayette, Purdue University (1993)
5. McPeak S,Necula G. Elkhound: A GLR parser generator. In: Proc. Of the 13th Int'l Conf. on Complier Construction. Spring-Verlag (2004) 51-56
6. M. Tomita: Generalized LR Parsing. Kluwer Academic (1991)
7. McPeak S, Necula G. Elkhoud: A GLR parser generator. In: Proc. of the 13th Int'l Conf. on Compiler Construction. Spring-Verlag (2004) 51-56
8. A. Bhamidipaty and T. A.Proebsting: Very Fast YACC-Compatiable Parsers(For Very Little Effort). Technical Report TR 95-09, Department of Computer Science, University of Arizona (1995)
9. P. Eades, X.Lin ,and W.F.Smyth: A fast and effective heuristic for the feedback arc set problem. Information Processing Letters, Vol.47. (1993) 319-323
10. J.R.Kipps: GLR parsing in Time $O(n^2)$. In Tomita (1984) 43-59
11. D.E.Knuth: On the Translation of Languages from Left to Right. Information and Control,Vol.8. (1965) 607-693
12. F.E.J.Kruseman Aretz: On a Recursive Ascent Parser. Information Processing Letters,Vol.29. (1998) 201-206
13. R.Leermakers: Recursive ascent parsing: From Early to Marcus. Theoretical Computer Science,Vol.104. (1992) 299-312
14. A.V.Aho, R.Scthi, and J.D. Ullman: Compiler:Principals ,Techniques,and Tools. Addison-Wesley (1986)
15. Weng FL, Zhou B, Wu LD: Process the phenomena of extra grammaticality in NL parsing. Journal of Chinese Information Processing,Vol.8. (1994) 1-13
16. Parr TJ,Quong RW. ANTLR: A predicated-LL(k) parser generator. Software-Practice and Experience,Vol.25. (1995) 789-810
17. J. Aycock: Faster Tomita Parsing. MSc thesis, University of Victoria (1998)
18. Aycock J, Horspool N, Janousek J, Melichar B: Even faster generalize parsing. Acta Informatica,Vol.37. (2001) 633-651
19. Donelly C, Stallmen R: The Bison Manual: Using the YACC-Compatible Parser Generator, for Bison Version 1.875. GUN Press, Boston (2004)
20. Maydene Fisher , Jon Ellis, and Jonathan Bruce: JDBC API Tutorial and Reference, Third Edition, Sun Microsystem (2004)
21. David Flanagan: Java Example in a Nutshell, 2nd Edition , O'Reilly&Association (2003)
22. Eric Armstrong, etc:The Java Web Services Tutorial, Pearson Education (2003)
23. Scott Oaks,Henry Wong: Java Threads , O'Reilly&Association (1999)
24. M.Tomita: An Effective Parsing for Natural Language, Kluwer Academic (1986)
25. Zhou M, Liu X, Huang CN: An efficient stochastic context-free parsing algorithm. Journal of Software, Vol.9. (1998) 59-87

Locomotion Control of Distributed Self-reconfigurable Robot Based on Cellular Automata

Qiu-xuan Wu[1], Ya-hui Wang[2], Guang-yi Cao[1], and Yan-qiong Fei[3]

[1] Dept of Automation Shanghai Jiaotong University, 200030, Shanghai, P.R.China
[2] Dept of Mechanical & Electronical Engineering, Puyang Vocational and Technical College, 457000, Henan, P.R.China
[3] Institute of Robot Shanghai Jiaotong University, 200030, Shanghai, P.R.China
wuqiuxuan@sjtu.edu.cn

Abstract. Based the character of Modular Self-Reconfigurable (MSR) robots, a homogeneous lattice robot called M-Cubes which owe to the property and structure of agent was build, feature vector matrix of modules describe completely the connection relation among modules. Motion mode and action of modules were analyzed. Emergent control is the most suited system control of MSR by comparing control way. Due to MSR robot is similar with cellular automata, emergent control model based CA was proposed. A two layers NN which has 7 inputs and single output simulate the nonlinear rules function, the feature vector of module is input of CA's rules, action of module is output of CA's rules. The simulation results show that emergent control based CA is great significance to enhance robustness and scale extensibility of MSR.

1 Introduction

Self-Reconfigurable robots is a new research area branched from modular robot in 90's, which consists of many of identical and independent module. Each module is a self-contained module including driver, sensor, processor, communication and power, it can connect and detach autonomously with adjacent modules, the robots owned the capability was called modular self-reconfigurable(MSR) robots. MSR robots have a lot of advantages: (1) versatility, it can adjust its shape to fulfill different tasks; (2) robustness, it are modular and distributed control therefore they have a high degree of robustness, if a module fails it can be replaced by other modules in the system. (3) Adaptability, the robots can change shape during run-time, this means that it can adapt to the environment to a higher degree than conventional robotic systems, because conventional robots have a fixed morphology. (4) Scale extensibility, the sizes of the robots can be increased or decreased during run-time by adding or removing modules. This means that if the initial number of modules is insufficient given the task more modules can be allocated in order to complete the tasks. Such novel characteristics are expected to have various advantages for operations such as space, sea, military scout, search and rescue, transportation and maintenance in

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 179 – 188, 2005.

dangerous and in unstructured environments. Recently, MSR robots have attracted many researchers. Various types of modular systems have been proposed in abroad [1,2,3,4,5,6,7,8,9], in these researches, the hardware design of the module and control algorithms are the central issues.

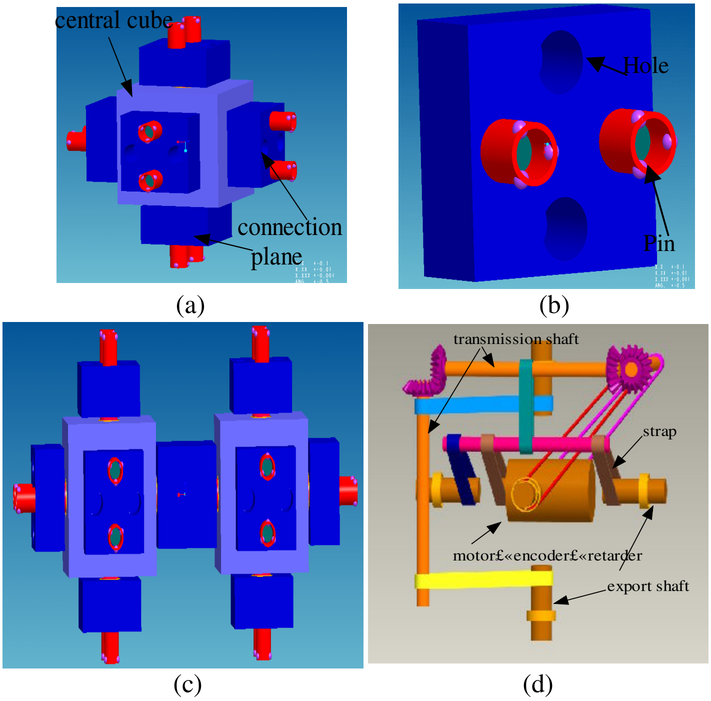
This paper centers the research of MSR distributed locomotion based Cellular Automata. The rest of the paper is organized as followed. Section 2 shows emergent control and its advantages. Section 3 presents homogeneous lattice module and its description. Section 4 presents the model and control based Cellular Automata. Section 5 discusses the experiments on MSR robot.

2 Emergent Control

Control system for self-reconfigurable robots can be classified depending on the relationship between the control system and the modules of the robot, centralized systems and distributed systems[10]. If a monolithic controller in a one-to-many relationship controls modules, the control system is classified as a centralized control system. A distributed control system, on the other hand, is spread out over many controllers resulting in a many to many relationship between controllers and modules. Distributed control systems can be further sub-classified depending on the type of information on which they depend. If a distributed control system depends on global information, we classify the control system as a distributed system based on global information. If the distributed controllers only depend on local information, we classify the control system as a distributed control system based on local information. We also refer to this class of control systems as control systems based on emergence or emergent control system [11]. Centralized control is conventional control ways of robots system, it is not fit for MSR because of robustness weak, scaling difficult and adaptability unskillful etc. On the contrary, the planning process and the motion control of distributed control are distributed in the each module and robustness of system was enhanced. Requirement of reliable communication among modules decrease the robustness and the sizes increased of modules cause big load of communication of each processor for global information distributed control. However, local communication of emergent control among neighbors modules decrease the load of communication, scale extensibility well and support parallelism. It enhanced the robustness of system and are better suited for operation in unstructured environments. In essence, MSR robots are a nonlinear complex system, solving the motion and metamorphose problems are difficult for conventional control algorithms. From the view of complex system, emergence is total complex behavior which happened local reciprocity action among modules, we called the control as emergent control. Emergent is central embodiment of nonlinear system and complex system, Cellular automata are discrete dynamic systems whose behavior is specified in terms of local interactions [12]. CA are models whose space, time and states are discrete, its essence character is emergent compute. The merits of CA are fitting for describing and simulating the emergent control of MSR robots.

3 Self-reconfigurable Robot Module

Recently, a few test beds and specimen were proposed in abroad, robots can be divided into chain-type robots and lattice-type robots according to the function of module, the former has good at locomotion and bad at metamorphic because of having the fixed conjunction. The capability of the latter is on the contrary because of without having the fixed conjunction.



(a) (b)

(c) (d)

Fig. 1. (a) Cell module; (b) Connection plane; (c) Two modules connecting; (d) Inner driver mechanism

3.1 Module Structure

In order to improve the flexible in spatial motion and correct orientation between modules [13], we design an MSR robot called M-Cubes which is 3-dimension homogenous lattice and have12 freedom degree, the simple structure was composed of central cube and connection plane illustrated as Fig.1 (a), connection between two modules illustrated as Fig.1 (c), central cube include power, motor, control circuit and drive mechanism etc, inner drive illustrated as Fig.1(d). Connection plane include connection self-lock mechanism which consist of two holes and two pins, communication interface and sensors etc. Since the robot have the characteristic of agent, it can be designed by mutual way, cell module constitute of meta-connection (meta-sensor), meta-communication, meta-power, meta-control, meta-motion, rules database and knowledge database. Meta-connection accomplish connecting and detaching of modules, communication metal fulfill importance information propagating of modules, meta-power provide energy to motion, meta-control deal with the information transmitted from the sensors and other modules to produce a series motion, meta-motion implement motion between modules, knowledge database store the status and goal task of modules, etc

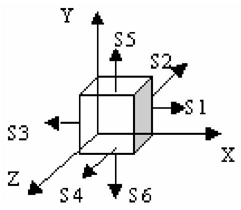


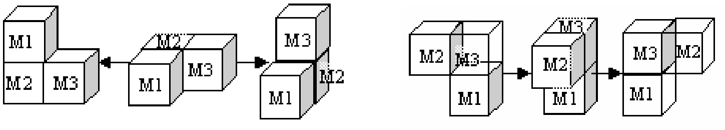
Fig. 2. Single Module feature vector

3.2 Module Description

To study DMAS (distributed multi-agent system) motion control of MSR, the model of single module and multi modules must be described, it can be called configuration expression. The topology structure of robot embodies the relation of orientation and connection among modules, this expression corresponds with the structure of the graph theory, module corresponds with the vertex of diagram, connecting line corresponds with the edge of diagram. If module is simple as a vertex, line is added between modules connecting, the topology structure of MSR robots can be express by graph. To express clearly the position of each module, it must build up a stationary coordinates $\sum_0$, each module code M_i *(i=1,2...n)*, *n* is amount of modules, the orientation of module decides on $\sum_0$, module is simplified a cube, the coordinate $M_i(x, y, z)$of module M_i is the center of cube. Space is divided cube grid units. modules motion in grids. The module only connects six neighbors modules, sensing the information of six neighbors grids are empty/no empty, attach/detach, expressing relation of six connection plane utilizing $(s_1, s_2, s_3, s_4, s_5, s_6)$ illustrated as Fig.2. Consequently the connecting relation of n modules can be described by *n*6* dimension feature vector, the feature vector is digit values. If adjacent module is empty, the value is 0, if neighbor modules is connecting, the feature value is *ID,* if having adjacent module but no connecting, the value is *–ID*, *ID* is the identifier of module from 1 to *n*. Every module can obtain the connection relations of neighbors and *ID* of neighbors modules by searching anticlockwise six direction from right. Since modules have point-to-point communication, only the topology structure of robot keep connection, all modules can be visited according to a rule and ordering, the feature vector of each module can be obtained, a *n**6 dimension matrix can be gotten by arranging each row feature vector according the *ID* order of modules, the matrix can express uniquely the topology structure of robot.

3.3 Basic Motion

Basic motion of M-Cubes is implemented by the rotation of connection plane. It have overturn and level movement, under the same view point, level movement is the motion which module is running on the surface of other modules, overturn is the motion that module rotate to the surface of other module, the kinds of motion need the help of other modules to accomplish, it can be divided into support module, carried module, passenger module according to the function of modules. In Fig.3, $M_{3,} M_{1,} M_2$ are all cell module, M_1 want on the top surface of M_2, then M_1 is passenger, M_2 is carried, M_3 is support. Both the carrier and the passenger must release some of their other connections prior to the carrying motion. M_1 is on the top surface of M_2 by the S_3 face of M_3 rotating 90 degree. Note that each module's role is not fixed but changes dynamically during the reconfiguration process. The same method, M_3 is on the top surface of M_1 too. So is the principle of level motion illustrated as fig.3. But the topology structure of robot must keep connecting during the process of motion. Modules have the action of connecting and detaching, connecting is the action that the pins of two neighbors connection plane plug each other the holes of the others and self-lock mechanism maintain the pins fixed position, the two neighbors connecting plane are coupled as a body. On the contrary, detaching is the inverse action of connecting. The two neighbor connecting planes are decoupled in favor of single module motion. If module motion anticlockwise, all motion of module can be divided 12 actions of rotating 90 degree, Turn_right_to_Top(), Turn_Top_to_left(), Turn_Left_To_Bottom(),Turn_Bottom_to_Right(), Turn_Front_to_Top(), Turn_Top_to_Back(), Turn_Back_to_Bottom(), Turn_Bottom_to_Front(), Level_Right_to_Back(), Level_Back_to_Left(), Turn_Left_to_Front(), Level_Front_to_Right(). If allowing rotation clockwise, the motions have 24 actions.



(a) (b)

Fig. 3. Module Motion. (a) Module overturn; (b) Module level move.

4 MSR Robot Model Based Cellular Automata

Cellular Automata(CAs) are dynamical systems in which space and time are discrete. A cellular automata consists of an array of cells, each of which can be in one of a finite number of possible states, synchronization update the states of all cells in discrete time steps, according to a local, identical interaction rule. The system evolve are implemented by local action among cells. The basic units of CAs have cellular, cellular space, neighbor and transition rules. The central problem of CA is how to get transition rules, since different transition rules decide system configuration[15]. CAs are interesting because even though they are simple, they can show very complex behaviors and patterns. They have been used to study electro-static self-assembly processes.

4.1 Cellular Automata Model of MSR Robot

The characteristics of homogenous lattice MSR robot decide that the robot is a temporal and spatial discrete system. We represent a generic self-reconfiguring robot as a collection of cells, each cell corresponding to one robot module. We view the resulting structure as a particular type of cellular automata. In condition of no considering weight, friction, velocity and inertia, MSR robot based CA is a 3-D CA model, it can be described as followed.

Cellular space is discrete grids that module motion space is divided according to the geometry of module, We represent the basic module of the robot as a cube. Each module is assumed to be able to translate across the surface of other modules as well as transition to other planes of motion. It is also assumed to be able to examine or query its neighbors and determine the presence or absence of a neighbor on all sides.□

Cellular states indicate the next step action, if module motion clockwise, then the state variables total 13, 0 means no action, from 1 to 12 means respectively 12 different action of overturn and level motion. Von Neumann type neighbor was adopted in CA, in 3-D space, up/down, right/left, front/back six cell are neighbors, neighbor radius is 1, utilizing $S_1, S_2, S_3, S_4, S_5, S_6$ indicate the relation of six neighbor cell.

Transition rules, it can be said that the state transition of all cells from step t to step t+1 implement a spatial process. The spatial structure is distribution and assembling of cell module. In MSR robot, spatial process is permutation and combination of cells in space, i.e. configuration, each result of space process is a step update of old configuration. Change of spatial structure can influence the environment of spatial entity, spatial action is a reflection of entity to environment in certain degree. To MSR robot based CAs, rules of CA decide on the combination of neighbors states. Local structure influence personal behaviors in space. V_j^t , V_j^{t+1} mean respectively states of cell in step t and in step t+1. $S_1^t, S_2^t, S_3^t, S_4^t, S_5^t, S_6^t$ are respectively the relation between the *j* cell and neighbors in step t, then transition rules can be denoted equation (1).

$$V_j^{t+1} = f(V_j^t, S_1^t, S_2^t, S_3^t, S_4^t, S_5^t, S_6^t) \ . \tag{1}$$

f is function, in fact, the states of cells are the actions which modules will be performed in he next step. The module to be carried is passenger module, the module have 12 kinds of actions, each step is a action and system is updated. If a module has a action in last step, then the module can not motion in next step, i.e. the module can not act continuously twice. Each S_m^t has four states, i.e. neighbors configurations of each module have $\left(C_4^1\right)^6$ =4096 kinds. A state was added to the center module, there are 4096*13=53248 states. These states are mapped 13 different actions, 0 means no action, the nonlinear transition rules function was trained by a NN which is a 2 layer 7input and single output illustrated as Fig.6. The output of NN can be denoted equation (2).

$$V_j^{t+1} = \sum_{i=1}^{6} S_i^t * W_i + V_j^t * W_7 \ . \qquad (2)$$

$W_1, W_2, \ldots W_7$ are weight of NN, a passenger module was decided by transition rules function. Each module is a agent, communication of modules depend on wireless communication module among neighbors modules. The passenger module will search support module and carrier module to implement action. If a module has no neighbor, then the state of the module is 0.

4.2 Motion Algorithm

Many lattice MSR robot motion like cluster flow, modules migrate from tail to head like pedrail of TANK. Aimed to level movement of M-Cubes, motion rules of having obstacle and having no obstacle are developed. Since the motion of M-Cubes need the help of neighbors, transition rules based metal-module were developed, this is main difference with other paper because of few considering realization of module. Generally, few lattice modules can motion autonomously, the concept of meta-module is provided by Casal and Yim [14]. If modules motion through connecting group instead of single module motion, lattice MSR robot can motion like chain-type robot. First concluding the motion process of MSR from the motion of meta-modules.

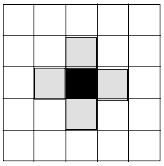


Fig. 4. Von Neumann Neighbors

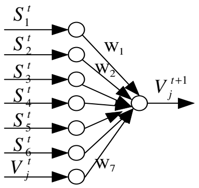


Fig. 5. Two layer NN

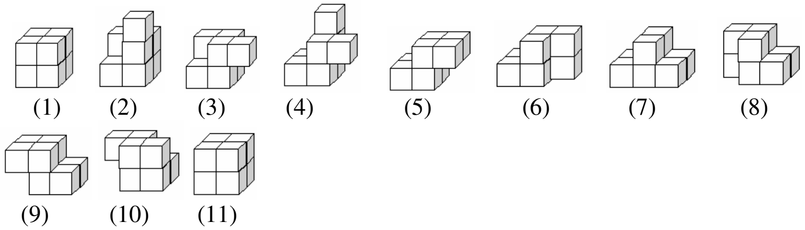


Fig. 6. Process of meta-modules without obstacles

During the process illustrated Fig.6, after steps, meta-module move eastward a cell module distance, metal-modules can migrate from a position to other position by repeating above process. Of course, according the process, MSR robot only motion toward specific direction, it can not swerve. In fact, the swerve rules are similar with

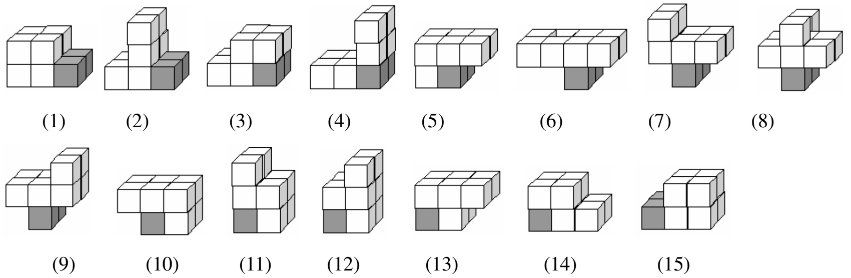


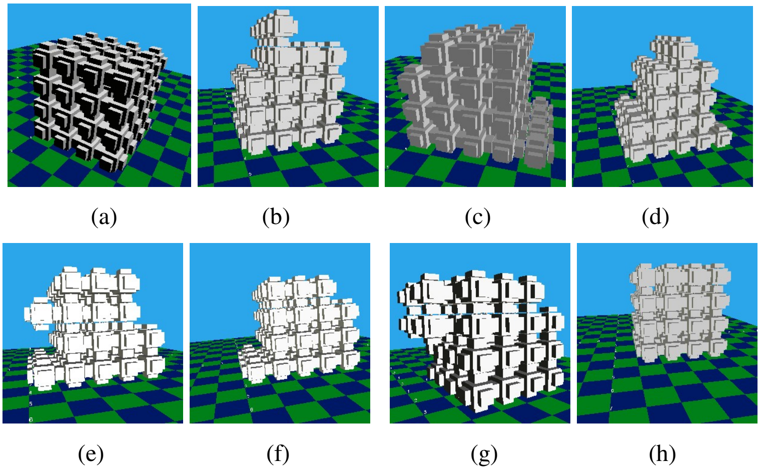
Fig. 7. Process of meta-modules with obstacles

above process, but MSR only rotate 90 degree. we can analyze and obtain the transitions rules of MSR based CA from motion process. To the environment with obstacles, we assume that the obstacles are cubes the same as cell module, it can be denoted utilizing the same module without ability of communication and connection, gray module is obstacle illustrated as Fig.8. The height of obstacles is a module lower than metal-module, the process of meta-module overturn obstacles [16,17] as followed Fig.7.

The states of module and connecting relation of neighbors are seen as inputs of NN and the action of module is seen as output of NN, this NN trained can get motion transition rules function based obstacles and no obstacles, locomotion of MSR can be implemented if the motion transition rules function become transition rules of MSR based CA.

5 Experiment Results

Aimed to the homogenous and lattice MSR robot and above locomotion control algorithms, a simulation environment was designed utilizing Java 3D, 4*4*4 modules migrating toward x-axis direction was simulated illustrated as followed Fig.8, in the figure, blue and green floor denotes substrate, each blue or green grid indicate cell module, the origin of system is the center of floor, Fig.8 (a) means initializing configuration locate zero position in system, Fig.8 (b)(c)(d)(e)(f)(g) is respectively a snapshot of system locomotion, Fig. 8(h) is configuration whose total system move a module distance toward x axis direction. It accomplishes the goal of total system structure locomotion. In the same way, only repeating continuously above locomotion process, MSR robot will motion further distance. MSR robot only based local information and predefined rules synchrony update local information and produce the action of module. It can get properly motion series. control the motion of module. Passenger module, support module and carrier module finish together the overturn or level movement of passenger module through local communication among three modules. Since communication and action only happened in neighbors' modules, the robustness and scale extensity of MSR robot system have great improvement. In theory, it can scale infinite without adding extra resources. Since without considering the balanced of system, it is studying further.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 8. Motion simulation of MSR toward x axis positive direction

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (No.50305021).

References

1. Fukuda ,T., Kawakuchi,Y.: Cellular Robotic System (CEBOT) as One of the Realization of Self-organizing Intelligent Universal Manipulator. In Proc. of IEEE Int'l Conf. on Robotics and Automation (1990) 662-669
2. Pamecha,A., Chiang,C-J., Stein,D., et al.: Design and Implementation of Metamorphic Robots. In Proc. of the 1996 ASME Design Engineering Technical Conference and Computers in Engineering Conference. Irvine, California (1996)
3. Murata,S., Kurokawa,H., Yoshida,E.,et al.: A 3-D Self-reconfigurable Structure. In Proc. of the IEEE Int'l Conf. on Robotics and Automation (1998) 432-441
4. Kotay,K., Rus,D.: Locomotion Versatility through Self-reconfiguration. Robotics and Autonomous Systems. Vol. 26 (1999) 26:217-32
5. Tomita,K., Murata,S., Kurokawa,H., et al.: Self-assembly and Self-repair Method for a Distributed Mechanical System. IEEE Trans. on Robotics and Automation. Vol. 15 (1999) 1035-45
6. Shen,W-M., Will,P. and Castano,a.: Robot Modularity for Self-reconfiguration. In SPIE Conf. On Sensor Fusion and Decentralized Control in Robotic Systems 2. Boston (1999)
7. Yoshida,E., Murata,S., Tomita,K., et al.: An Experimental Study on a Self-repairing Modular Machine. Robotics and Autonomous Systems. Vol. 29 (1999) 79-89
8. C Unsal, P Khosla: Mechatronic Design of a Modular Self-reconfiguring Robotic System. In Proc. of IEEE Int'l Conf. on Robotics and Automation (2000) 1742-1749
9. Rus,D., Vona,M.: Crystalline Robots: Self-reconfiguration with Unit-compressible modules. Autonomous Robots. Vol. 10 (2001) 107-24

10. Tomita,K., Murata,S., Kurokawa,H., et al.: Self-assembly and Self-repair Method for a Distributed Mechanical System. IEEE Transactions on Robotics and Automation. Vol. 15 (1999) 1035-1045
11. Stoy,K.: Emergent Control of Self-Reconfigurable Robots. Odense M · Denmark: The Maersk Mc-Kinney Moller Institute for Production Technology University of Southern Denmark (2004)
12. Chirikjian,G., Pamecha,A., Ebert-Uphoff,I.: Evaluating Efficiency of Self-Reconfiguration in a Class of Modular Robots. Journal of Robotic Systems. Vol. 13 (1996) 317-338
13. Castano,A., Will,P.: Mechanical Design of a Module for Reconfigurable Robots. Proceedings of the 2000 IEEE/RSJ international Conference on intelligent Robots and Systems. Vol. 3 (2000) 2203-2209
14. Cassal. A, Yim. M: Self-reconfiguration Planning for a Class of Modular Robots. SPIE Symposium on Intelligent Systems and Advanced Manufacturing. Boston (1999)
15. Supratid,S., Sadananda,R.: Determinism in Cellular Automata Investigation of Transition Rules, Proceedings of International Conference on Intelligent Sensing and Information Processing (2004) 391-397
16. Butler,Z., Kotay,K., Rus,D., et al.: Cellular Automata for Distributed Control of Self-reconfiguring Robots. In Proceedings of the ICRA 2001 workshop on modular robots. Seoul, Korea (2001)
17. Stoy,K.: Controlling Self-Reconfiguration using Cellular Automata and Gradients. In proceedings of the 8th international conference on intelligent autonomous systems (IAS-8), Amsterdam, The Netherlands (2004) 693-702

Improvements to the Conventional Layer-by-Layer BP Algorithm*

Xu-Qin Li[1], Fei Han[1,2], Tat-Ming Lok[3], Michael R. Lyu[4], and Guang-Bin Huang[5]

[1] Institute of Intelligent Machines, Chinese Academic of Sciences,
PO Box 1130 Hefei Anhui, China 230031
[2] Department of Automation, University of Science and Technology of China
Hefei 230027, China
[3] Information Engineering Dept., The Chinese University of Hong Kong, Shatin,
Hong Kong
[4] Computer Science & Engineering Dept., The Chinese University of Hong Kong, Shatin,
Hong Kong
[5] School of Electrical and Electronic Engineering, Nanyang Technological university,
Singapore
{xqli, hanfei1976}@iim.ac.cn, tmlok@ie.cuhk.edu.hk,
lyu@cse.cuhk.edu.hk, EGBHuang@ntu.edu.sg

Abstract. This paper points out some drawbacks and proposes some modifications to the conventional layer-by-layer BP algorithm. In particular, we present a new perspective to the learning rate, which is to use a heuristic rule to define the learning rate so as to update the weights. Meanwhile, to pull the algorithm out of saturation area and prevent it from converging to a local minimum, a momentum term is introduced to the former algorithm. And finally the effectiveness and efficiency of the proposed method are demonstrated by two benchmark examples.

1 Introduction

The error back propagation algorithm (EBP) was a major breakthrough in neural network research [1][2][3][4][5][6]. However, the basic algorithm is too slow for most practical applications. So researchers have proposed several variations of error back propagation that provide significant speedup and make the algorithm more practical [7][8]. To accelerate the EBP algorithm, some modified error functions, which are different from popular mean-squared errors (MSE's), have been proposed [9].

In 1995, Ergezinger and Thomsen [10] proposed a layer-by-layer algorithm (LBL) which was based on a linearization of the nonlinear processing elements and the optimization of the EBP layer-by-layer. And in order to limit the introduced linearization error, a penalty term was added to the cost function. Commonly the

* This work was supported by the National Natural Science Foundation of China (Nos.60472111 and 60405002), and RGC Project No.CUHK 4170/04E, RGC Project No. CUHK4205/04E and UGC Project No.AoE/E-01/99.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 189 – 198, 2005.

proposed layer-by-layer algorithms were decomposed into two parts: a linear one and a nonlinear one. The linear part of each layer was solved through the least-square errors (LSE's) or mean-squared errors (MSE's). But the nonlinear parts were different in a certain extent, with some of which have to assign the desired input to the hidden targets, while some of them not, or some of which use a heuristic rule to define the learning rate, while some of them use an optimal one to define the learning rate [11][12].

Although these methods have showed a fast convergence through decreasing the possibility to a premature saturation [13][14], sometimes they still result in some inevitable problems. They may not converge to the desired accuracy or involve huge computational complexity due to target assignments at hidden layer. Essentially, these methods were used to define the learning rate so as to adapt the weights [15].

This paper proposes a new prospective to the conventional proposed layer-by-layer method. This method tends to overcome the stalling problem of the layer-by-layer algorithm with a heuristic method. And also the momentum terms are introduced to both the output layer and the hidden layer in order to accelerate convergence when the conjugate gradient is moving in a consistent direction.

This paper is organized as follows. The following Section gives a brief review to the conventional layer-by-layer method. Section III introduces a new prospective to the learning rate, and the momentum method is also integrated into the algorithm. In Section IV, the improvement is demonstrated by two benchmark problems. Finally, Section V concludes this whole paper.

2 Layer-by-Layer BP Algorithm

We consider a single hidden-layer perceptron for the sake of simplicity. The activation functions for output layer and hidden layer are linear function and sigmoid function, respectively. The training patterns are described as $x^{(p)}=[x_1^{(p)},x_2^{(p)},\cdots,x_N^{(p)}]^T$ (p=1,2,…,P) with associated target vectors of output layer $t^{(p)}=[t_1^{(p)},t_2^{(p)},\cdots\ ,\ t_{\mathrm{N}}^{(p)}]^T$ (p=1,2,…,P). So the network can be depicted as:

$$\hat{h}_j^{(p)}=\sum_{i=0}^{N} w_{ji}x_i^{(p)}\ (x_0=0). \tag{1}$$

$$h_j^{(p)}=\tanh(\hat{h}_j^{(p)}/2). \tag{2}$$

$$y_k^{(p)}=\hat{y}_k^{(p)}=\sum_{j=0}^{H} v_{kj}h_j^{(p)}\ (\ h_0=0\ ). \tag{3}$$

The weights should be optimized in order to minimize the MSE at the output layer defined as:

$$E^{out} = \frac{1}{2}\sum_{p=1}^{P}\sum_{k}^{M}(t_k^{(p)} - \sum_{j=0}^{H} v_{kj}h_j^{(p)})^2 \cdot \tag{4}$$

2.1 Optimization of the Output Layer Weights

With a fixed W and the desired output $t^{(p)}$, optimize V for minimizing the cost function E^{out} :

$$\Delta v_{kj} = \eta_k^{out}\alpha_{kj} . \tag{5}$$

$$\alpha_{kj} = -\frac{\partial E^{out}}{\partial v_{kj}} = \sum_{p=1}^{P}\hat{d}_k^{(p)}h_j^{(p)} . \tag{6}$$

$$\eta_k^{out} = \frac{\sum_{j=0}^{H}\alpha_{kj}{}^2}{\sum_{p=1}^{P}(\sum_{j=0}^{H}\alpha_{kj}h_j^{(p)})^2} = \frac{\vec{\alpha}_k^T\vec{\alpha}_k}{\vec{\alpha}_k^T C_h\vec{\alpha}_k} . \tag{7}$$

where $\hat{d}_k^{(p)} = -\partial E^{out} / \partial\hat{y}_k^{(p)} = t_k^{(p)} - \hat{y}_k^{(p)}$, $\vec{\alpha}_k = [\alpha_{k0}, \alpha_{k1} \cdots \alpha_{kH}]^T$ and $C_h = \{\sum_{p=1}^{P} h_j^{(p)}h_j^{(p)'}\}_{(H+1)\times(H+1)}$.

2.2 Assign the Hidden Targets

With the updated V, we assign the hidden targets denoted by $z_j^{(p)}$:

$$z_j^{(p)} = h_j^{(p)} + \varsigma_p\beta_j^{(p)} . \tag{8}$$

$$\beta_j^{(p)} = -\frac{\partial E^{out}}{\partial h_j^{(p)}} = \sum_{k=1}^{M}\hat{d}_k^{(p)}v_{kj} \cdot \tag{9}$$

$$\varsigma_p = \frac{\sum_{j=1}^{H}\beta_j^{(p)2}}{\sum_{k=1}^{M}(\sum_{j=1}^{H}v_{kj}\beta_j^{(p)})^2} = \frac{\vec{\beta}^{(p)T}\vec{\beta}^{(p)}}{\vec{\beta}^{(p)T}C_v\vec{\beta}^{(p)}} \cdot \tag{10}$$

where $\vec{\beta}^{(p)} = [\beta_1^{(p)}\beta_2^{(p)}\cdots\beta_H^{(p)}]^T$ and $C_v = \{\sum_{k=1}^{M} v_{kj}v_{kj}^{'}\}_{H\times H}$, and assign $\hat{z}_j^{(p)} = f^{-1}(z_j^{(p)})$ after truncating $z_j^{(p)}$ to stay in (-1, 1).

2.3 Optimization the Hidden Layer Weights

We use the training patterns $x_i^{(p)}$ and $\hat{z}_j^{(p)}$ to define a new cost function at the hidden layer [9]:

$$E^{hid} = \frac{1}{2}\sum_{p=1}^{P}\sum_{j=1}^{H}(\hat{z}_j^{(p)} - \sum_{i=0}^{N} w_{ji}x_i^{(p)})^2[f^{'}(\hat{z}_j^{(p)})]^2 . \tag{11}$$

and optimize W for minimizing E^{hid} as follows:

$$\Delta w_{ji} = \eta_j^{hid}\gamma_{ji} . \tag{12}$$

$$\gamma_{ji} = -\frac{\partial E^{hid}}{\partial w_{ji}} = \sum_{p=1}^{P} \hat{e}_j^{(p)}[f^{'}(\hat{z}_j^{(p)})]^2 x_i^{(p)} . \tag{13}$$

$$\eta_j^{hid} = \frac{\sum_{i=0}^{N}(\gamma_{ji})^2}{\sum_{p=1}^{P}(\sum_{i=0}^{N}\gamma_{ji}x_i^{(p)})^2[f^{'}(\hat{z}_j^{(p)})]^2} = \frac{\hat{\gamma}_{ji}{}^T\hat{\gamma}_{ji}}{\hat{\gamma}_{ji}{}^T C_x \hat{\gamma}_{ji}} . \tag{14}$$

where $\hat{e}_j^{(p)} = -\partial E^{hid} / \partial \hat{h}_j^{(p)} = \hat{z}_j^{(p)} - \hat{h}_j^{(p)}$ $\hat{\gamma}_{ji} = [\gamma_{j0}\gamma_{j1}\cdots\gamma_{jN}]^T$ and $C_x = \{\sum_{p=1}^{P} x_i^{(p)}x_i^{(p)'}[f^{'}(\hat{z}_j^{(p)})]^2\}_{(N+1)\times(N+1)}$ [9].

3 Modifications to the Conventional Algorithm

We would like to make the learning rate larger at the initial stage, since then we will be taking large steps and would expect to converge faster. However, if we make the learning rate too large, the algorithm will become unstable. It is impossible for us to predict the maximum allowable learning rate for arbitrary functions, but fortunately for quadratic functions we can set an upper limit.

Considering the output layer cost function:

$$E^{out} = \frac{1}{2}\sum_{p=1}^{P}\sum_{k}^{M}(t_k^{(p)} - \sum_{j=0}^{H} v_{kj}h_j^{(p)})^2 \cdot \tag{15}$$

It can be transformed to a quadratic function with respect to v_{kj} :

$$F(v_{kj}) = \frac{1}{2}v_{kj}{}^T A v_{kj} + d^T v_{kj} + c \,. \tag{16}$$

The gradient of this quadratic function is $\nabla F(v_{kj}) = Av_{kj} + d$. Then A is called Hessian matrix of this quadratic function. Using a constant learning rate α , we obtain this expression according to the steepest descent algorithm:

$$v_{kj}(epoch+1) = v_{kj}(epoch) - \alpha(A * v_{kj}(epoch) + d)\,. \tag{17}$$

Or

$$v_{kj}(epoch+1) = [I - \alpha * A]v_{kj}(epoch) - \alpha d\,. \tag{18}$$

This linear dynamic system will be stable if the eigenvalues of the matrix $[I - \alpha * A]$ are less than one in magnitude. We can express the eigenvalues of this matrix in terms of the eigenvectors of the Hessian matrix A. Suppose $\{\lambda_1, \lambda_2, \cdots \lambda_n\}$ and $\{z_1, z_2, \cdots z_n\}$ to be the eigenvalues and eigenvectors of the Hessian matrix. Then $[I - \alpha A]z_i = z_i - \alpha A z_i = z_i - \alpha \lambda_i z_i = (1 - \alpha \lambda_i) z_i$.

So the eigenvectors of $[I - \alpha * A]$ are the same as the eigenvectors of A. Similarly, the eigenvalues of $[I - \alpha * A]$ are $(1 - \alpha\lambda_i)$. Naturally, we can obtain the following expression:

$$\left|(1 - \alpha\lambda_i)\right| < 1\,. \tag{19}$$

The eigenvalues must be positive so that the quadratic function can be guaranteed to converge to a stable minimum point. So Equ. (20) will be reduced to: $\alpha < \frac{2}{\lambda_i}$.

Since it must be true for all the eigenvalues of the Hessian matrix, we have $\alpha < \frac{2}{\lambda_{\max}}$. So the maximum stable learning rate is inversely proportional to the matrix curvature of the quadratic function ($\alpha < \frac{2}{\lambda_{\max}}$) [16][17].

In fact, the optimal learning rate always changes during different applications which makes it difficult to be set optimally. So the so-called optimal learning rate in the algorithm aforementioned does not always perform best. As long as we can find out the Hessian matrix, then calculate the eigenvalues so as to define the maximum stable learning rate, the algorithm tends to converge most quickly in the direction of the eigenvector corresponding to this largest eigenvalue.

It is well known that backpropagation with momentum updating is one of the most popular modifications to the standard algorithm. When a momentum term is added to the EBP algorithm, in which the weights change is a combination of the new steepest decent step and the previous one, the weight trajectory will be much smoother and the convergence will be faster. Intuitively, the momentum term can be also added to the layer-by-layer BP algorithm.

Taking the output layer for example:

When the momentum term is added to the algorithm, the weights are updated according to the description in literature [18].

$$v_{kj}(epoch+1) = v_{kj}(epoch) + (1-\alpha)\Delta v_{kj}(epoch) + \alpha\Delta v_{kj}(epoch-1)\,. \tag{20}$$

Similarly, the hidden layer weights are updated according to the following formula:

$$w_{ji}(epoch+1) = w_{ji}(epoch) + (1-\alpha)\Delta w_{ji}(epoch) + \alpha\Delta w_{ji}(epoch-1)\,. \tag{21}$$

4 Simulation and Results

4.1 Function Approximation

In this section, a function approximation example was trained by a 1-3-1networks with 3 nodes in the hidden layer. The input data is denoted as P= -2,-1.6,-1.2…1.2, 1.6,2, and the desired output data as T=sin (3.14*p/4), which are assigned to the input and output layer of the network respectively [16].

4.1.1 Fixed Learning Rate

Fig. 1 and 2 show the MSE for the training patterns of the two methods for this function approximation example.

It can be seen from Fig. 1 and 2 that our improved method can reduce the MSE dramatically than the previous method, and the previous method tends to reach a certain MSE and remains there for the rest iterations with little or no improvement, while the proposed algorithm can make huge progress to reduce the MSE throughout the entire training process. So our improved method can decrease the MSE to an acceptable level when the training process for the previous method traps into the saturation area.

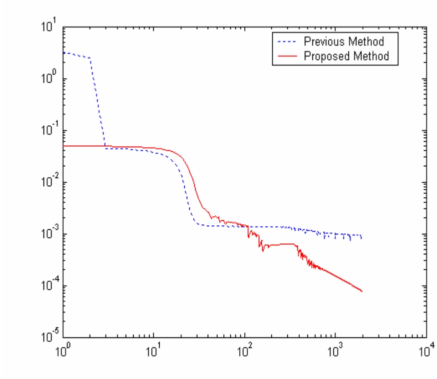


Fig. 1. The MSE cures of previous and our improved methods for a fixed learning rate (0.11) at the output layer

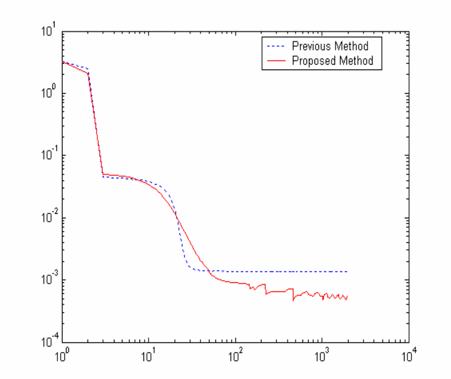


Fig. 2. The MSE cures of previous and our improved methods for a fixed learning rate (0.44) at the hidden layer

4.1.2 Incorporation of the Momentum Term

In this subsection, we further demonstrate the efficiency and effectiveness of another momentum term method. This method is derived by incorporating the momentum term, α into the weight updating formulae. Assume that the coefficient of α for the output layer and the hidden layer were set to 0.88 and 0.70, respectively.

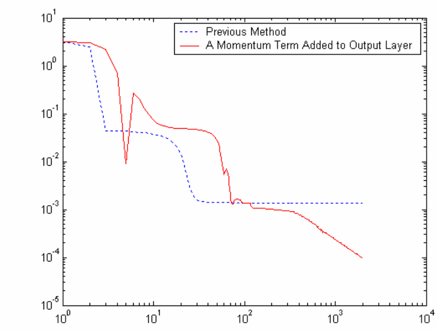


Fig. 3. Learning curves of the MSE for the previous and the output momentum term methods for a function approximation example

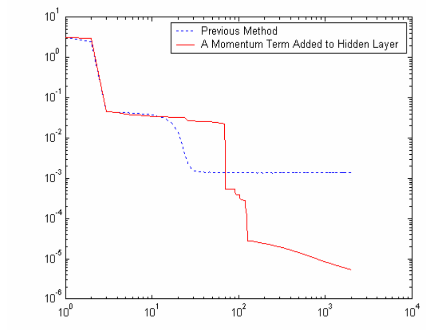


Fig. 4. Learning curves of the MSE for the previous and the hidden momentum term methods for a function approximation example

Fig. 3 and 4 illustrate the improvement of the introduction of momentum term to the original method. Although the previous method converges at the earlier stage, it stops to a local minimum after a certain number of iteration. The proposed algorithm can prevent the training process from falling into the flat regions, and make the MSE decrease dramatically until the MSE converges to a noticeable small level

4.2 XOR Problem

For an XOR problem, the network consists of two input nodes, there hidden nodes, and one output node.

4.2.1 Fixed Learning Rate

In the case of fixed learning rate, we use XOR to conduct some computer simulations. It can be found that the sum-square-errors of the previous algorithm cannot converge to an acceptable level but our proposed method can do. Fig.5 depicts a comparison of the MSEs for the standard EBP algorithm and the proposed algorithm, where the learning rate in the standard EBP method was set to 0.02 and the one for the hidden layer of the proposed method was set to 0.45.

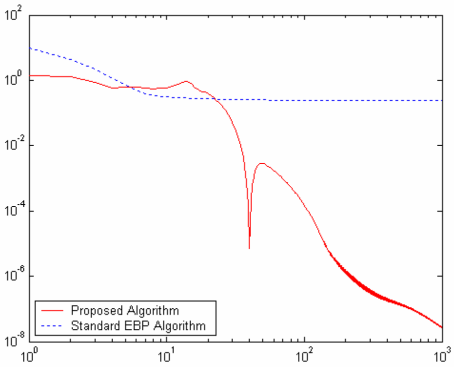


Fig. 5. Learning curves of the MSE for the standard EBP and our proposed methods for the XOR problem

We observe that the training process is easily trapped into saturation area for the standard EBP algorithm, while the proposed method of a fixed learning rate can converge over the whole learning procedure until it reaches to an acceptable level.

4.2.2 Incorporation of the Momentum Term

In addition, we also use XOR problem to verify our proposed momentum term method. Figs. 6 and 7 demonstrate the improvement of the proposed method over the standard EBP algorithm, where the coefficient of α for the output layer (Fig.6) and the hidden layer (Fig.7) were set to 0.80 and 0.60, respectively.

Fig. 6 and 7 show that our proposed method can prevent the training process falling into the flat regions, and make the MSE decrease dramatically until the MSE converges to a noticeable small level, while the standard EBP method stops to a local minimum after a certain number of iterations in spite of convergence at the early stage. So the proposed algorithm with a momentum term can meet a stringent demand in accuracy.

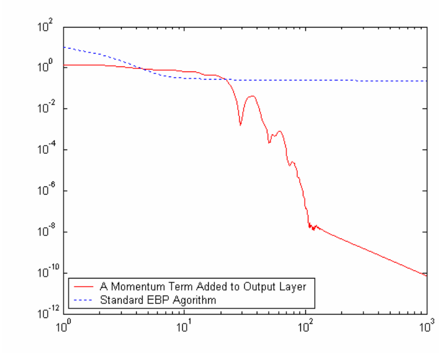


Fig. 6. Learning curves of the MSE for the momentum term and the standard EBP methods for XOR problem

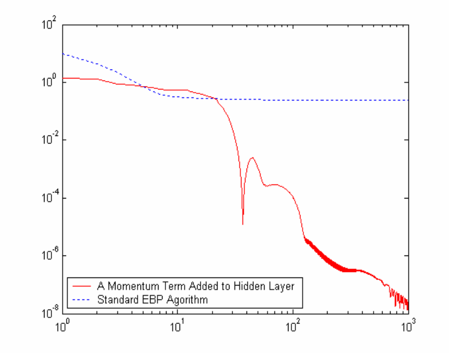


Fig. 7. Learning curves of the MSE for the momentum term and the standard EBP methods for XOR problem

5 Conclusions

This paper proposed some modifications to the conventional layer-by-layer BP algorithm so as to accelerate the learning process and reduce the possibilities to be trapped into the saturation area. To prove the efficiency and effectiveness of the proposed method, we trained a MLP network with a function approximation example and the XOR problem. In all the experiments, the proposed algorithm had demonstrated its improvement with orders of magnitude less than the previous algorithm in MSE. The modified method has also validated that the heuristic rule is sometimes better than the so-called optimal learning rate. So, much more research works about the optimal learning rate for a specific algorithm has to be done. What's more, when the momentum term is added into the algorithm, it can lead to the convergence from a local minimum to a global one, and reduce the MSE significantly. So the momentum term functions are well suitable not only to the standard BP algorithm but also to the layer-by-layer algorithm.

References

1. Huang, D.S., Horace, H.S.Ip and Zheru Chi: A Neural Root Finder of Polynomials Based on Root Moments. Neural Computation, Vol.16, No.8, pp.1721-1762, 2004
2. Huang D.S., Horace H.S.Ip, Law Ken, C. K. and Zheru Ch: Zeroing Polynomials Using Modified Constrained Neural Network Approach. IEEE Trans. On Neural Networks, vol.16, no.3, pp.721-732, 2005
3. Huang, D.S.: A Constructive Approach for Finding Arbitrary Roots of Polynomials by Neural Networks," IEEE Transactions on Neural Networks□Vol.15, No.2, pp.477-491, 2004
4. Huang, D.S.: Systematic Theory of Neural Networks for Pattern Recognition. Publishing House of Electronic Industry of China, Beijing, 1996

5. Rumelhart ,D.E. and McClelland, J.L.: Parallel Distributed Processing. Cambridge, MA: MIT Press, 1986
6. Huang, D.S., Horace, H.S.Ip and Zheru Chi: A Neural Root Finder of Polynomials Based on Root Moments.Neural Computation, Vol.16, No.8, pp.1721-1762, 2004
7. Huang, D.S.: The Local Minima Free Condition of Feedforward Neural Networks for Outer-supervised Learning. IEEE Trans on Systems, Man and Cybernetics, Vol.28B, No.3, 1998,477-480
8. Huang, D.S.: Radial Basis Probabilistic Neural Networks: Model and Application. International Journal of Pattern Recognition and Artificial Intelligence, 13(7), 1083-1101,1999
9. Sang-Hoon Oh and Soo-Young Lee: A New Error Function at Hidden Layers for Fast Training of Multilayer Perceptrons. IEEE Trans. Neural Networks, vol. 10, pp. 960–964, 1999
10. Ergezinger,S., and Thomsen,E.: An Accelerated Learning Algorithm for Multilayer Pereceptrons: Optimization Layer by Layer. IEEE Trans. Neural Networks, vol. 6, pp. 31–42, 1995
11. Wang,G.-J. and Chen,C.-C.: A fast Multilayer Neural Networks Training Algorithm Based on the Layer-by-layer Optimizing Procedures. IEEE Trans. Neural Networks, vol. 7, pp. 768–775, 1996
12. B.Ph.van Milligen, V.Tribaldos,J.A, Jimenez, and C.Santa Cruz: Comments on: An Accelerated Algorithm for Multilayer Perceptrons: Optimization Layer by Layer. IEEE Trans. Neural Networks, vol. 9, pp. 339–341, 1998
13. van Ooyen and Nienhuis, B.: Improving the Convergence of the Backpropagation Algorithm. Neural Networks, vol. 5, pp. 465–471, 1992
14. Huang, D.S.: The Bottleneck Behavior in Linear Feedforward Neural Network Classifiers and Their Breakthrough. Journal of Computer Science and Technology, Vol.14, No.1, 34-43,1999
15. Oh, S.-H.: Improving the Error Backpropagation Algorithm with a Modified Error Function. IEEE Trans. Neural Networks, vol. 8, pp.799–803, 1997
16. Martin T. Hagen and Howard B. Demuth: Neural Network Design. USA, PWS publishing company, 1996
17. Yann LeCun, Leon Botton, Genevieve, B.Orr and Klause-Robert Muller: Efficient Backprop. Neural Networks, LNCS 1524,pp.9-50,1998
18. Fredric M.Ham and Ivica Kostanic: Principles of Neuaocomputing for Science and Engineering,USA, McGraw-Hill Companies, Inc.2001

An Intelligent Assistant for Public Transport Management

Martin Molina

Department of Artificial Intelligence, Universidad Politécnica de Madrid,
Campus de Montegancedo s/n 28660, Boadilla del Monte, Madrid, Spain
mmolina@fi.upm.es

Abstract. This paper describes the architecture of a computer system conceived as an intelligent assistant for public transport management. The goal of the system is to help operators of a control center in making strategic decisions about how to solve problems of a fleet of buses in an urban network. The system uses artificial intelligence techniques to simulate the decision processes. In particular, a complex knowledge model has been designed by using advanced knowledge engineering methods that integrates three main tasks: diagnosis, prediction and planning. Finally, the paper describes two particular applications developed following this architecture for the cities of Torino (Italy) and Vitoria (Spain).

1 Introduction

The problem of management of a fleet of vehicles has been recently facilitated by the current telecommunication and information technology. In this field, a new generation of information systems have been proposed for a wide range of services in public transport and fleet management [1], [2]. In particular, the recent advances of this technology allow operators in a control center to monitor the location and state of each particular vehicle, in order to apply on real time global transport strategies. This is especially useful to quickly react in the presence of incidents produced by unexpected situations such as vehicles with malfunctions, road blocked in the transport network, etc. In this context, advanced computer systems can help operators in improving their answer [3].

The development of this type of systems requires formulating advanced models that capture the different facets of the strategic knowledge for public transport management. In order to provide an integrated solution for this problem, we describe in this paper the architecture of a computer system that follows the idea of *intelligent assistant* [4]. An intelligent assistant is a concept derived from artificial intelligence that identifies a kind of systems whose role is to *assist* the user in decision-making processes. This type of systems emphasizes that the operator is the final responsible of decisions, so the system is not designed to substitute the operator but, on the contrary, its goal is to provide services for assistance such as: information filtering and interpretation to identify significant data, *what if* analysis, justification of conclusions, etc.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 199 – 208, 2005.

According to this, the paper describes first the problem of public transport management corresponding to a fleet of buses of a urban network, identifying the main tasks to be provided: diagnosis, prediction and planning. Then, the paper describes the knowledge model designed to support these tasks, showing the selected knowledge-based methods for this purpose. Finally, the paper describes two of the applications developed following this approach. The first one was developed for the city of Torino (Italy) in the context of an international European Project (funded by the European Comission within the Telematics Applications Programme). The second application was developed in a national project in Spain for the public transport managemen in the city of Vitoria.

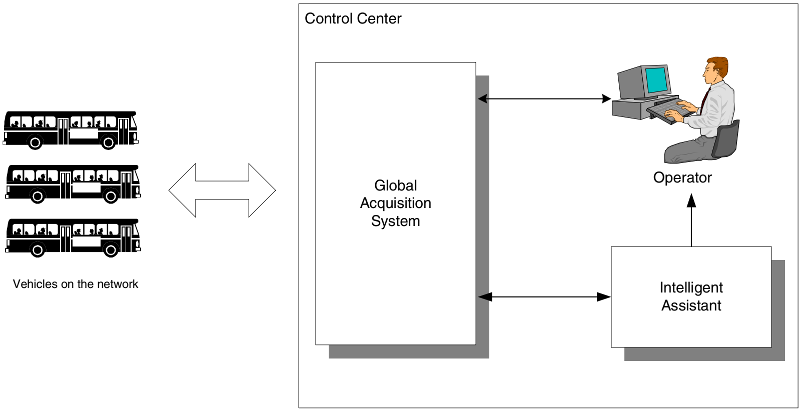


Fig. 1. Basic control scenario in public transport management

2 The Problem of Public Transport Management

A typical public transport management scenario is composed of a fleet of buses that in real time send information about their location and state to a control center (figure 1). The goal of the public transport operators at the control center is to give to the passengers an adequate service in order to: (1) guarantee as much as possible the normal service according to the initial planning, (2) respect the timetable of drivers, (3) guarantee the security of the service, and (3) avoid discomfort of passengers and drivers.

The information available every minute in the control center about every vehicle includes: location, previous and next vehicles, actual headway/delay measure and its trend, etc. Other information are alarms (overloaded vehicle, broken vehicle, vehicle in traffic-jam, lack or relief driver, etc.), in-line drivers, current timetables, depots state, etc. In a typical decision scenario of this type of control centers, operators perform three basic tasks: diagnosis, prediction and planning.

The *diagnosis* task is oriented to identify the presence of abnormal situations based on the received information in real time about the location and state of the vehicles. An abnormal situation in this context is a situation where the planned service is altered due to a set of possible causes. These situations have different degrees of severity that go from a slight delay of a single vehicle to a severe disruption of the service due to a blockade of the line or a malfunction of one vehicle. In general, abnormal situations include different types of problems for each line such as deviations of vehicles from their scheduled services (slight or severe), service disruption, unreachable relief point, slight or severe malfunction of a vehicle, blockade of the line, etc.

The goal of the *prediction* task is to estimate the future impact of the detected problems. For this purpose, operators can estimate the number of passengers affected and the average delay assumed for the next period of time. This is performed using knowledge about the historical demand of the lines and knowledge about the behavior of the vehicles (e.g., average travel times between stops).

Finally, the goal of the *planning* task is to find appropriate actions according to the detected problems that should be taken in order to improve the transport service. These actions in general need to be performed in a particular order and they can cover more than one bus line. In particular, some of the types of basic actions are: (1) *skipping*, a bus does not take passengers during several stops, (2) *detour,* a bus takes a path that goes out of the line in order to avoid an obstacle (for example, traffic incident, constructions, etc.), (3) *limitation*, a vehicle takes the opposite direction of the line, without covering the complete path, (4) *reinforcement,* a reserve vehicle (a bus that is not performing a service in a particular line) is sent to cover an unattended area in the line with the objective of restoring the timetable of delayed vehicles, (5) *shift*, the drivers of two vehicles are exchanged to restore the right order in the duty, and (6) *rotation*, the schedule of buses is shifted one step to recover delays.

3 The Knowledge Model of the Intelligent Assistant

In order to provide an appropriate level of support for public transport management, the intelligent assistant system should simulate the natural thinking process followed by operators using their strategic criteria at the same levels of abstraction. For this purpose, we applied advanced knowledge engineering techniques following a *model-based* approach. This modeling approach considers the existence of a conceptual level, at which the knowledge is first described without considering implementation issues. Some of the recent methodologies for knowledge system development follow this model-based approach (e.g., CommonKADS [5]). These methodologies basically organize the whole knowledge using the following concepts: (1) a *task* that identifies a goal to be achieved (for instance, diagnosis or prediction), (2) a *method* indicates how a task is achieved, by describing the different reasoning steps (subtasks) by which its inputs are transformed into outputs, and (3) a *type of knowledge base* that identifies explicitly the type of domain knowledge that supports a task.

According to this, figure 2 shows a global view of the knowledge model that we designed for the public transport problem. This figure shows a hierarchy of tasks (circles) and methods (rectangles) with types of knowledge bases at the bottom (cylinders). The figure shows that the global task, *public transport management*, is divided into the three main subtasks (*diagnosis*, *prediction* and *planning*) that correspond to the three tasks described in the previous section. The figure also shows how each subtasks is decomposed in simpler subtasks and how they are supported by different types of domain knowledge.

Thus, the goal of the first of the three main tasks, *diagnosis*, is to interpret the current situation in order to detect the presence of problems. This type of reasoning includes a qualitative interpretation of the raw information received in the control center together with a classification of the situation according to a prefixed set of types of problems. For this purpose it is useful to use an adaptation of the heuristic classification method [6], with a data-driven control regime that considers two simpler subtasks: *abstract* and *match*. The abstract subtask is a primary task whose goal is to produce qualitative values from numerical data about the state of the system. For example, this task interprets the delay (in minutes) of each vehicle producing one of three qualitative values (slight, medium or severe). The goal of the *match* subtask is to determine the type of problem that is present at each line, using a set of classes of problems. Two types of knowledge bases support the *diagnosis* task: one that includes abstraction knowledge and another one with problem types.

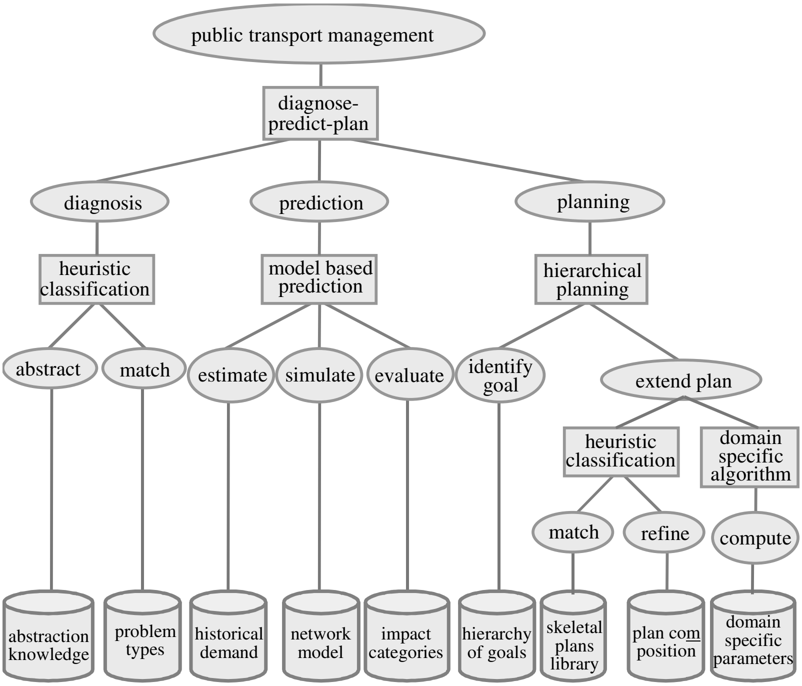


Fig. 2. Knowledge model for public transport management

On the other hand, the goal of the *prediction* task is to estimate the future impact of the problem. This prediction, estimates the total number of passengers affected by the delay together with the average delay for those passengers. This is calculated for the next T minutes (for instance, T=15 minutes). In particular, for example, this method may compute the values for the impact I and total number of affected persons P:

$$I = \sum_{i=1}^{n} p_i \cdot e_i \,. \tag{1}$$

$$P = \sum_{i=1}^{n} p_i \,. \tag{2}$$

where p_i is the estimated number of persons waiting at the stop i, and e_i is the increase of waiting time due to the delay of the bus. The sum is done for the n stops affected in the T minutes. The value I gives a quantitative measure to compare the severity of different problems. Thus, it provides a global criterion to solve conflicts when different problems in different lines need the same resource to improve the service (e.g., a reserve vehicle). The value $E = I/P$ expresses the average increase of waiting time.

The prediction task is divided into three subtasks: *estimate future demand*, *simulate behavior* and *evaluate impact*. The first subtask estimates the future demand by using a local knowledge base with historical demand. This includes the expected number of passengers at each stop each interval of time. The second subtask simulates the movement of delayed buses for the next T minutes by using a model of the bus-line (stops, distances, travel-times, etc.) and determines the affected stops. Finally, the third subtask estimates global metrics (e.g. I and E) and uses domain knowledge to interpret the severity of the situations.

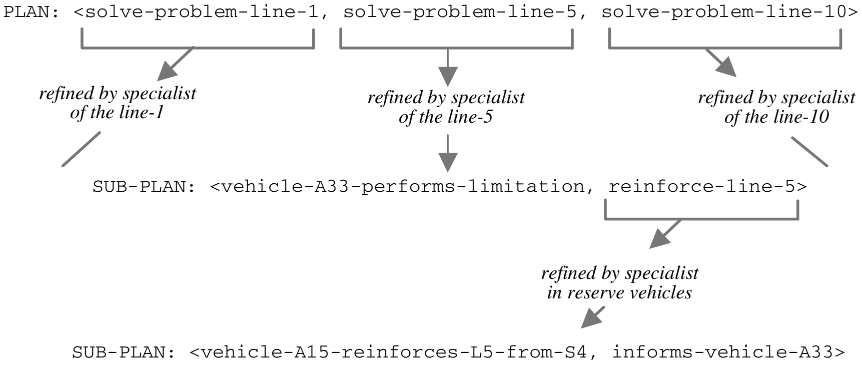

PLAN: <solve-problem-line-1, solve-problem-line-5, solve-problem-line-10>
refined by specialist of the line-1
refined by specialist of the line-5
refined by specialist of the line-10
SUB-PLAN: <vehicle-A33-performs-limitation, reinforce-line-5>
refined by specialist in reserve vehicles
SUB-PLAN: <vehicle-A15-reinforces-L5-from-S4, informs-vehicle-A33>

Fig. 3. Example of development of a plan by the intervention of different specialists that, by turn, refine parts of an abstract plan

The third task, *planning*, recommends a set of control actions that may solve the detected problems. In this type of reasoning, different classes of specialized heuristic knowledge are typically used to dynamically construct the plan. For example, knowledge about alternative paths of bus lines, criteria about reserve vehicle management, specific criteria to exchange drivers, etc. In order to simulate this type of reasoning it is possible to use a method for hierarchical planning that integrates the idea of skeletal planning [7] and the concept of specialists [8]. The method is based on a search in a hierarchy of goals (specialists) that are knowledgeable about partial abstract plans. Each specialist is responsible of producing a partial plan to solve part of the detected problem.

Thus, the knowledge model for this case is represented by a method that divides the *planning* task into two main subtasks: *identify goal* and *extend plan*. The plan is dynamically composed during the reasoning developing a search in such a way that the first subtask identifies the next goal to reach, analyzing the current plan and using a hierarchy of goals (the hierarchy of specialists). Then, the second subtask extends the current plan with partial sub-plans. This process is repeated step by step until the complete final plan is produced. The second subtask is performed by two alternative methods (depending on the level of the hierarchy of goals): (1) one method that selects a plan using heuristic classification knowledge, and (2) another method that constructs the plan by using a domain-specific algorithm with certain parameters (in the bus transport application, there are several instances of this method, for example, the method for reserve vehicles selection).

Figure 3 shows an example about how a plan is developed by the intervention of several specialists. In the example, first, a global plan is produced with three abstract actions that respectively are associated to line-1, line-5 and line-10. The corresponding

Type of KB	**Representation**	**Examples**
Abstraction knowledge	Rules	*IF X is-a vehicle, delay of X = D, D > 10* *THEN level-of-delay of X = severe*

Problem types	Rules
	or
	frames
		IF X is-a vehicle, full-vehicle-alarm of X = on,
		current-stop of X = S
		THEN new problem P, type of P = overflow,
		affected-vehicle of P = X, location of P = S

Historical demand	Table	*Stop Time-Interval Type-of-day Hist-Demand* *---* *S1, 6:30-9:30, weekday, 20,* *S2, 6:30-10:00, weekday, 15, ...*
Network model	Objects	*Object L3 is a Line. Attributes:* *Stops: (S1,S7,S15,S18,S21,S45,S56,S78), ...,*

Impact categories	Rules	*IF X is-a line, affected-passengers of X > 50,*
		average-delay of X > 5
		THEN impact-level of X = severe

Hierarchy of goals	Table	*Specialist Goals Pre-cond. Post-cond.* *---* *Reinforcement, G1, G2, ... A1, ... A2, ...* *Limitation, H1, H2, ... A3, ... A4, ...*
Skeletal plans library	Rules	*IF X is-a problem, type of X single-severe-delay,* *affected-vehicle of X = V* *THEN new plan P, type of P = reinforced-limitation,* *affected-vehicle of P = V*
Plan composition	Frames	*PLAN reinforced-limitation.* *ACTION TYPE concrete: vehicle X performs limitation* *ACTION TYPE abstract: vehicle X is reinforced* *...*
Domain specific parameters	Table	*Parameter Value* *---* *max-number-of-stops-to-regulate, 5,* *min-percentage-of-absorption, 80%, ...*

Fig. 4. Symbolic representation of knowledge bases for the model of public transport management

specialist refines each abstract action. For example, the second action is refined by the specialist of the line-5 and produces a sub-plan with two actions. In addition to that, the second action of this sub-plan needs to be refined by another specialist.

In summary, according to the previous description, the model includes a total of 9 types of knowledge bases. Figure 4 shows the symbolic representation followed by each type of knowledge base. Thus, for instance, the knowledge base that supports the *abstract* task can be represented with *if-then* rules with a forward-chaining inference method. This representation is very flexible and intuitive to formulate a wide range of expressions for qualitative interpretation. The knowledge base for problem types can be represented either with rules or frames. Here, the antecedent of a rule can express the set of conditions for a certain type of problem and the consequent determines the characteristics of the deduced problem. Frames can be also used here, where each frame represents a type of problem with a set of slots that expresses a set of single conditions.

The knowledge base for historical demand can be represented with a table that relates each stop of a line with its historical demand according to the type of day and the time interval. For the knowledge base of the network model, an object-oriented representation can be used with concepts such as stop, line, vehicle, etc. and attributes with values that characterize the concepts and establish relations between concepts. The knowledge base for impact categories can be also represented using *if-then* rules to interpret the quantitative values about the expected delays.

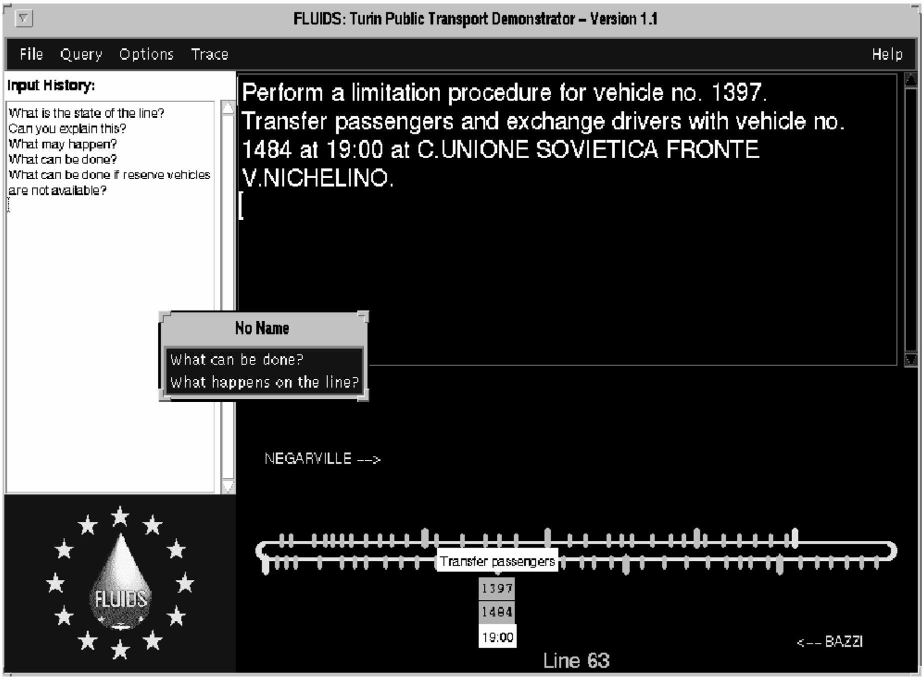
Concerning the knowledge related to the planning task, the knowledge base about the hierarchy of goals can be represented with a table that relates each type of goal with the corresponding specialist and the conditions to be considered. The knowledge base for skeletal plans library includes the existing types of plans. Here, rules can relate conditions of the problem with the corresponding type of plan. The knowledge base for plan composition expresses the set of actions in which a plan is decomposed. For this purpose, frames can be used in such a way that each frame identifies a plan and the set of slots identifies the structure of the plan with abstract or concrete actions. Finally, when a specialist is supported by a set of parameters, the knowledge base with domain specific parameters include the corresponding values.

It is important to note that these 9 types of bases can produce a higher number of specific knowledge bases in a particular model because some of the bases are specified with different contents for each bus line. Thus, in a concrete model for several bus lines, some knowledge bases can be shared for the management of all the lines but other knowledge bases need to be specified for each particular line. Thus, for instance, a particular model with N bus lines could include: 1 abstraction knowledge base, 1 knowledge base for problem types, N knowledge bases for historical demand, N knowledge bases for network model, 1 knowledge base for impact categories, 1 for hierarchy of goals, 1 for skeletal plan library, 1 for plan composition, M for domain specific parameters (where M is the number of specialists that use domain specific algorithms to dynamically construct a partial plan).

4 Applications

The model described in this paper was applied for the development of two different applications. The first application was developed within an international project in

Europe, called Fluids [9], funded by the European Commission within the Telematics Application Programme. One of the main goals of this project was to provide advanced methods for user-system interaction in the context of real-time decision support. The design methods developed in the project were demonstrated in different applications for transport. In particular, a realization was developed for the case of public transport management for the city of Torino (Italy). The analysis of this problem produced a first version of the general approach presented in this paper.



Input History:

What is the state of the line?
Can you explain this?
What may happen?
What can be done?
What can be done if reserve vehicles are not available?

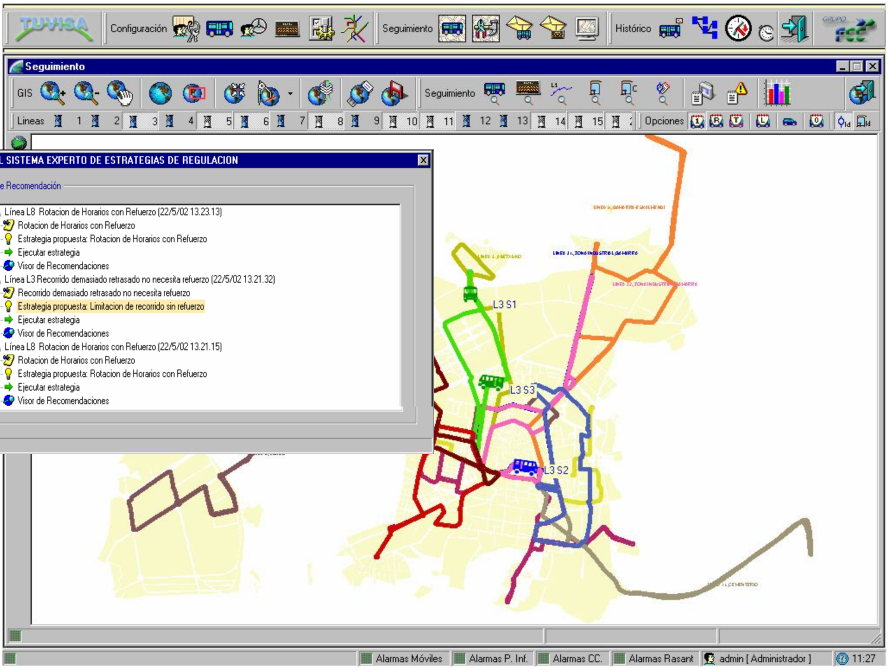
Perform a limitation procedure for vehicle no. 1397.
Transfer passengers and exchange drivers with vehicle no. 1484 at 19:00 at C.UNIONE SOVIETICA FRONTE V.NICHELINO.
[

Fig. 5. User interface of the Fluids application for public transport management

Figure 5 shows an example of the user interface developed for the Fluids project (it was tested on-line in the control center in 1998). The user interface combines a dialogue based on a prefixed set of questions for decision support together with a presentation created dynamically that integrates text and an animation to clarify with illustrations how to carry out specific control actions for vehicle management.

From the point of view of implementation, the development of this type of systems is a complex task because it requires to integrate different types of problem solvers with different types of knowledge bases. The final architecture must be both efficient for real time operation and flexible to accept changes according to the identification of new strategic knowledge. Thus, in order to help in the final implementation of the system, for the case of the Fluids application, we used a software environment called KSM [11] that has been already used for the development of other applications in the field of transport [12]. This environment is an advanced knowledge engineering tool that provides a model-based approach and a set of software components to facilitate the development.

The Fluids application was followed by another application for the city of Vitoria (Spain) which included some improvements compared to the initial version. It was conceived in a general way, according to the model described in this paper, to be reused for different public transport networks. This realization was developed and integrated on real-time in 2001 with the rest of the information system for the management of the bus lines of the control center of the city of Vitoria (Spain). The model includes the management criteria for a total of 15 bus lines. Figure 6 shows the global user interface of this application.



VISOR DEL SISTEMA EXPERTO DE ESTRATEGIAS DE REGULACION

Sistema de Recomendación

- Línea L8 Rotacion de Horarios con Refuerzo (22/5/02 13.23.13)
  - Rotacion de Horarios con Refuerzo
  - Estrategia propuesta: Rotacion de Horarios con Refuerzo
  - Ejecutar estrategia
  - Visor de Recomendaciones
- Línea L3 Recorrido demasiado retrasado no necesita refuerzo (22/5/02 13.21.32)
  - Recorrido demasiado retrasado no necesita refuerzo
  - Estrategia propuesta: Limitacion de recorrido sin refuerzo
  - Ejecutar estrategia
  - Visor de Recomendaciones
- Línea L8 Rotacion de Horarios con Refuerzo (22/5/02 13.21.15)
  - Rotacion de Horarios con Refuerzo
  - Estrategia propuesta: Rotacion de Horarios con Refuerzo
  - Ejecutar estrategia
  - Visor de Recomendaciones

Fig. 6. User interface of the application for public transport management for the city of Vitoria

The implementation of this system was performed following a particular advanced software architecture that integrates the set of inference procedures and knowledge bases using a common working memory together with an explicit control mechanism. The user can access to the knowledge model using a particular user interface developed for this purpose in order to provide the required flexibility for knowledge model maintenance.

5 Conclusions

In summary, the model described in this paper can be an adequate solution to construct an intelligent assistant for public transport management. The model is a global solution that integrates different types of methods and problem-solvers to support three basic tasks: diagnosis, prediction and planning. The model-based approach followed in knowledge engineering provided the appropriate description level to formulate the complete model, according to the current state of the art of knowledge engineering. The model presented in the paper has been validated with the development of two real-world applications for the cities of Torino (Italy) and Vitoria (Spain).

References

1. Witulski, K.: Knowledge Based Route Selection in Public Transport. Operational Experience and Perspectives. Proc. 2nd OECD Workshop on Knowledge-Based Expert Systems in Transportation. Vol.1. Montreal (Canada). (June 1992)
2. Saint Laurent B. de: An Information System for Public Transport: The Cassiope Architecture Example of Passenger Information. Advanced Telematics in Road Transport, Edited by the Commission of the European Communities. Directorate-General Telecommunications. Information Industries and Innovation. Elsevier. (1991)
3. Cepeda M.: New Generation of Vehicle Scheduling and Control Systems. Advanced Telematics in Road Transport, Edited by the Commission of the European Communities. Directorate-General Telecommunications. Information Industries and Innovation. Elsevier. (1991)
4. Boy, G., Gruber, T.R.: Intelligent Assistant Systems: Support for Integrated Human-Machine Systems. Technical Report KSL 90-61. Knowledge Systems Laboratory. Computer Science Department. Stanford University. (1990) And published also in the proceedings of 1990 AAAI Spring Symposium on Knowledge-Based Human-Computer Communication. Stanford University (Mar. 1990)
5. Schreiber G., Akkermans H., Anjewierden A., de Hoog R., Shadbolt N., Van de Velde W., Wielinga, B.: Knowledge Engineering and Management. The CommonKADS Methodology. MIT Press. (2000)
6. Clancey, W.J.: Heuristic Classification. Artificial Intelligence. Vol. 27. (1985) 289 - 350
7. Friedland, P.E.: Knowledge-Based Experiment Design in Molecular Genetics. Proc. Sixth International Joint Conference on Artificial Intelligence. 285-287. Menlo Park. California. También: report STAN-CS-79-771. Stanford University. (1979)
8. Brown, D., Chandrasekaran B.: Design Problem-Solving. Knowledge Structures and Control Strategies. Morgan Kaufman. (1989)
9. Hernández, J., Molina, M., Cuena, J.: Towards an Advanced HCI Through Knowledge Modelling Techniques. Knowledge Engineering and Agent Technology. Cuena, J., Demazeau, Y., García-Serrano A., Treur J. (eds.) IOS Press. (2004)
10. Cuena, J., Molina, M.: The Role of Knowledge Modelling Techniques in Software Development: A General Approach Based on a Knowledge Management Tool. International Journal of Human-Computer Studies. Academic Press. No. 52. (2000) 385 - 421
11. Molina, M., Hernández, J., Cuena, J.: A Structure of Problem-solving Methods for Real-time Decision Support in Traffic Control. Journal of Human and Computer Studies (Academic Press) No.49. (1998) 577 - 600

FPBN: A New Formalism for Evaluating Hybrid Bayesian Networks Using Fuzzy Sets and Partial Least-Squares

Xing-Chen Heng and Zheng Qin

Research Institute of Computer Software of Xi'an Jiaotong University,
710049, Xi'an, Shanxi, PRC
boyhxc@163.com

Abstract. This paper proposes a general formalism for evaluating hybrid Bayesian networks. The formalism approximates a hybrid Bayesian network into the form, called fuzzy partial least-squares Bayesian network (FPBN). The form replaces each continuous variable whose descendants include discrete variables by a partner discrete variable and adding a directed link from that partner discrete variable to the continuous one. The partner discrete variable is acquired by the discretization of the original continuous variable with a fuzzification algorithm based on the structure adaptive-tuning neural network model. In addition, the dependence between the partner discrete variable and the original continuous variable is approximated by fuzzy sets, and the dependence between a continuous variable and its continuous and discrete parents is approximated by a conditional Gaussian regression (CGR) distribution in which partial least-squares (PLS) is proposed as an alternative method for computing the vector of regression parameter. The experimental results are included to demonstrate the performances of the new approach.

1 Introduction

Bayesian networks (BNs) are powerful graph-based framework, combined with a rigorous probabilistic foundation, can be used to model and reason in discrete, continuous and hybrid domains [1]. In discrete domains discrete Bayesian networks provide a general formalism for representing a joint probability distribution of discrete random variables. Exact general-purpose inference algorithms exist and are well developed, such as the junction tree inference algorithm [2]; In continuous domains continuous Bayesian networks are usually represented by Gaussian mixture distributions where sums of weighted Gaussian densities are used to approximate the likelihood functions [3].

The treatment of hybrid Bayesian networks (HBNs) with both discrete and continuous variables today is mainly influenced by the literatures [4], [5], [6], which use so called cg-potentials. The drawback of this approach is that the only three Gaussian parameters (mean, covariance matrix and a regression vector) are used to characterize the continuous densities, and these Gaussian parameters are estimated by a maximum

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 209 – 217, 2005.

likelihood (ML) calculation with lower accuracy and the full training data matrix. Another problem is that these methods can't handle discrete nodes as children of continuous parents. A further extension is to approximate an arbitrary conditional probability distribution by using sigmod-functions [7]. This approach is picked up in [8] to include it into Laurizen's mechanism. However, their accuracy is restricted to the only three Gaussian parameters of the densities again.

This paper takes a step further from the fuzzification and proposes a general formalism of fuzzy partial least-squares Bayesian network (FPBN) for evaluating hybrid Bayesian networks. The form only replaces each continuous variable whose descendants include discrete variables by a partner discrete variable and adding a directed link from that partner discrete variable to the continuous one. A fuzzy partition is used to cover the domain of the continuous variable with overlapping fuzzy sets and each discrete state of the partner discrete variable corresponds to a fuzzy set of the original continuous variable.

The dependence between the partner discrete variable and the original continuous variable is approximated by a fuzzy set instead of using a conditional Gaussian (CG) model, but the dependence between a continuous variable and its continuous and discrete parents is approximated by a conditional Gaussian regression (CGR) distribution in which partial least-squares (PLS) is proposed as an alternative method for computing the vector of regression parameter.

The remainder of this paper is structured as follows. The next section gives a general formalism of FPBN as approximate representation of hybrid Bayesian networks. Section 3 presents a fuzzification method to approximate the dependence between the partner discrete variable and the original continuous variable. The PLS method for calculating the vector of regression parameter is shown in Section 4. Section 5 gives an example of a hybrid Bayesian network and it's evaluation using the new approach.

2 Fuzzy Partial Least-Squares Bayesian Networks as Approximate Representation of Hybrid Bayesian Networks

Definition 1**:** *Hybrid Bayesian Networks*

A general hybrid Bayesian network (HBN) is a directed acyclic graph representing the joint probability distribution of a given set of variables **V**

$$\mathrm{HBN} = (\mathbf{V}\ , \mathbf{P}) = (\mathbf{X}, \mathbf{Y'}, \mathbf{P})\ . \tag{1}$$

Where $\mathbf{X} \subseteq \mathbf{V}$ denotes a set of discrete variables, $\mathbf{Y'} \subseteq \mathbf{V}$ a set of continuous variables, and

$$\mathbf{P} = \{P(V|\Gamma^{+}{}_{V}\ , \psi^{+}{}_{V})\ , V \in \mathbf{V}\}\ . \tag{2}$$

We use upper-case letters such as V for a variable name, and corresponding lower-case letters such as v for a value of a variable. All the continuous variables are denoted by capping with a ' such as Y'. $\mathbf{\Gamma^{+}}_{V}$ and$\boldsymbol{\psi}^{+}{}_{V}$ denote the set of discrete and continuous parents respectively for a variable V. Configurations of $\Gamma^{+}{}_{V}$ and $\psi^{+}{}_{V}$ are denoted by γ^{+}_{V} and ψ^{+}_{V} respectively.

The form of the FPBN for a given HBN is defined as follows:

Definition 2**:** *Fuzzy Partial Least-squares Bayesian Network*
Given a hybrid Bayesian network (HBN), the corresponding form of FPBN is obtained by applying the fuzzification transformation as defined by (4) only to those continuous variables **Z'** $\subseteq$ **Y'** whose descendants in HBN include discrete variables [10], which is defined by

$$\text{FPBN} = (\mathbf{X}, \mathbf{Z}, \mathbf{Z'}, \mathbf{W'}, \mathbf{L}, \mathbf{P}) . \tag{3}$$

Where **Z** is the set of discrete variable defined by discretizing the subset of continuous variables **Z'** $\subseteq$ **Y'**, and **W' = Y' \ Z'** whose descendants don't include discrete variables. **X** and **Z** are discrete variables and they don't have continuous parents, and **Z'** and **W'** are still continuous variables; but each Z'□**Z'** has only one discrete parent **Z** and is leaf node, and each W'$\in$**W'** still keeps its original discrete parents $\boldsymbol{\Gamma}^{+}_{W'}$ and continuous parents$\psi^{+}_{W'}$ and may have continuous descendants. **L** contains the set of directed links of the original HBN and the set of directed links obtained by the following fuzzification transformation of HBN:

$$Z'\in \mathbf{Z'}, \text{ replace } Z' \text{ in HBN by } Z \text{ and create a new directed link } Z \rightarrow Z' . \tag{4}$$

P is the set of conditional probability distributions. P(X|Γ^{+}_{X})and P(Z|Γ^{+}_{Z}) can be assumed to be multinomial which is general for discrete variables. P(Z'|Z) for Z'$\in$**Z'** can be approximated by the method of fuzzy sets that will be described in the section 3. The conditional distribution P(W'|$\boldsymbol{\Gamma}^{+}_{W'}$,$\psi^{+}_{W}$) is approximated by a conditional Gaussian regression (CGR) model.

$$P\,(w' \mid \gamma^{+}_{W} ,\ \psi^{+}_{W'}) = \frac{1}{\sqrt{2\pi\sigma_{\gamma}}} \exp\{ - \frac{(\,w'-u_r - B_r\psi)^2}{2\sigma_r} \} . \tag{5}$$

Where γ and ψ on the right side of this equation abbreviate γ^{+}_{W} and ψ^{+}_{W} respectively and B_r is a vector of regression parameters for the given discrete configuration γ^{+}_{W}. The B_r will be calculated by using the PLS method described in the section 4.

3 Hybrid Likelihood Function Based on Fuzzy Set

The use of continuous variables in Bayesian networks requires a discretization of their domains. In fact it would be infeasible to work with an infinite number of states. A smoother way of discretization proposed in [13] is to use a fuzzy partition, which is to cover the domain of the variable with overlapping fuzzy sets whose memberships sum to 1 for each value of the variable.

A value z' of the continuous variable Z' can be written as a fuzzy set:

$$z' = \sum_{i=1}^{k}(\ v_i / F_i) \ . \tag{6}$$

Where z' takes value F_i with membership v_i ($\sum_{i=1}^{k} v_i = 1$) and F_i is one of the k fuzzy sets which are acquired by using a fuzzy partition on the domain of the continuous variable Z'. The k fuzzy sets can be considered as k discrete states for Z'. The mapping between the continuous variable Z' and the fuzzy set F_i can be approximated by the membership function U_{F_i} (Z') which can be written as follows:

$$U_{F_i}\ (Z' = z') = \frac{1}{\sqrt{2\pi\sigma_{F_i}}} \exp\{-\frac{(z'-u_{F_i})^2}{2\sigma_{F_i}}\} \ . \tag{7}$$

The marginal probability of Z' is also represented as follows:

$$P(Z'=z'\ ) = \sum_F U_{F_i}(Z=z') * p(F_i\ ) = \sum_F \frac{P(F_i)}{\sqrt{2\pi\sigma_{F_i}}} \exp\{-\frac{(z'-u_{F_i})^2}{2\sigma_{F_i}}\} \ . \tag{8}$$

The accuracy of this approximation relative to the original HBN can be made sufficiently high if a sufficient number of discrete states for each Z is used. Meanwhile, the more number of discrete states contribute to the high computational complexity. That is, it is very necessary to finish a high qualitative fuzzy partition for accuracy and complexity. A fuzzification algorithm, which is based on the structure adaptive-tuning neural network model, is proposed in this paper to define the optimal number of discrete states for each continuous variable. In the model, the number of nodes of hidden layer and output layer are tuned automatically with the change of the input of the network. The change of the number of nodes of every layer is finished by nodes' combination and breakup. The whole fuzzification algorithm is described as follows:

Step1: First, initialize a small-scale neural network, next, measure a continuous variable x' and it's parent variables for m times to get m original data-pairs, then make the sampling on x' for n times by equal step length using the interpolation to get the samples vectors X, finally, initialize s=0 which denotes the success times.

Step2: Let (X, Y) be the training sample-pairs, and then learn the neural network by using some learning algorithm with (X, Y) to get the value of weights between nodes in the network. The Y is the output vector that is gotten when the neural network is initialized.

Step3: Let all observed values of the x' in the original data-pairs be the input vectors of the neural network in turn, and compute the every node's samples distribution coefficients R_i and the correlation coefficients $C_{i\,j}$ between nodes.

Step4: Judge if the neural network reaches the demands of the structural stability ($0.3<R_i<0.7$, $C_{i\,j}< 0.8$). If the demands are satisfied, turn to step6, or turn to step 5.

Step5: Decrease s. If the corresponding nodes' deletion or breakup needs to be executed, turn to step3 and continue to learn.

Step6: Increase s. If s overruns a specified value, end up, or turn to step2 and continue to learn.

4 PLS Method

In Bayesian networks it is necessary to compute relationships between continuous nodes. The standard Bayesian network methodology represents this dependency with a multivariate linear regression (MLR) model whose parameters are estimated by a maximum likelihood (ML) calculation.

Referring to Eq. (5), the continuous variable W' 's value w' is dependent on three parameters: the values of the node's parents, an offset term, and a vector of regression parameters. It is then possible to describe W' 's distribution by the following relationship:

$$W' \sim N(\alpha + B_r X^T, \Sigma) . \tag{9}$$

where α is the mean or offset, B_r the regression vector, X the vector of continuous variable parents of W' and Σ the covariance model for the distribution. Based on the assumption above, the continuous variable W' can be modeled as linear combination of the responses X by using the MLR which yields the following equation:

$$W' = \alpha + B_r X^T + \varepsilon , \varepsilon \sim (0, \Sigma) . \tag{10}$$

Eq. (10) is called multivariate linear regression equation of W' regarding X. As once this regression vector B_r has been calculated for W', these values are then used to calculate the covariance Σ. In addition, modeling the offset parameter α separately from the regression vector Bi is not necessary for MLR. Hence we only need to estimate the vector of regression parameter B_r of Eq. (5) with the standard method for estimating the parameters of a Bayesian network, maximum likelihood (ML), which yields the following solution:

$$B_r = ([X]^T [X])^{-1} [X]^T \mathbf{w'} . \tag{11}$$

where [X] is a matrix containing the expected values of the continuous variable parents of W' and **w'** is the property vector of W'.

ML is optimal in a least-squares sense. Unfortunately, a problem arises when inverting $[X]^T [X]$ in the above equation. The matrix inversion step tends to fail in large networks, in systems with small amounts of training data, and with data that is not well defined by a Gaussian distribution [9]. So high co-linearity exists among the response variables, $[X]^T [X]$ is not of full rank and its inversion is ill-defined. For this reason it is necessary to find a few descriptive variables that are linearly independent of one another in order to make inversion possible.

Partial least-squares regression is proposed as one way in this paper to overcome the collinearity problem in multivariate calibration. In PLS, the samples in *X* are mapped to a new axis space in which there are some principal components or latent variables. Each principal component or latent variable is a linear combination of the original measurement variables. These principal components or latent variables which maximize the covariance between *X* and *W'* with each successive principal component describing less variance, are used to estimate B_r through an MLR step [11], [12]. Usually, only a few principal components are required to capture the majority of the variance in the data. Also, as the number of latent variables or principal components used in the regression spans the full rank in *X*, they are linearly independent of one another. Thus, by regressing the property vector **w'** against the newly constructed principal components *U*, a least-squares solution can be obtained:

$$B_r = ([\mathrm{U}]^{\mathrm{T}}[\mathrm{U}])^{-1}[\mathrm{T}]^{\mathrm{T}}\mathbf{w'}\,. \tag{12}$$

Where The PLS solution will approach that of MLR in Eq. (11).

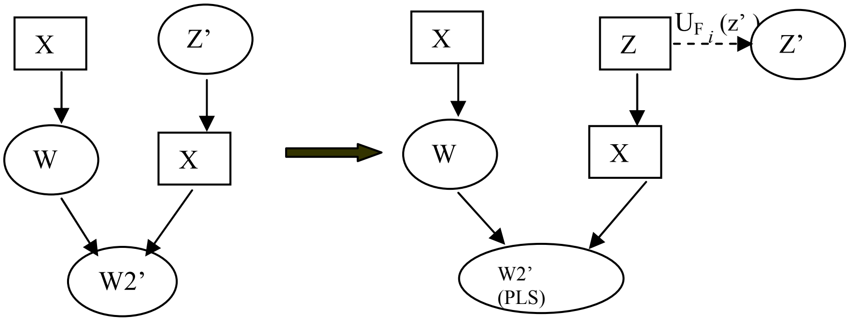
5 Experiments

To evaluate the proposed approach we give an example network for which some evidence is entered. Depending on that evidence the probability distribution over a node is calculated using the approach of Lauritzen [4] and the presented new approach. These results are compared with the true density.

We construct a hybrid Bayesian network as an example shown in figure 1(a). The parameters for this example network are set as following: The two root nodes X1 and Z' have the prior probability distributions P (X1=1) = 0.3, P (X1=2) = 0.7 and Z' ~ N (0,1) respectively.

In figure 1 (b), Z is the fictitious discrete variable (a set of fuzzy sets F) defined by a fuzzification over Z'. Z has two discrete states. The discrete variable X2 has also two discrete states. Hence the conditional probability of X2 given Z is model as following:

$$\mathrm{P}\ (\mathrm{X2}=2 \mid \mathrm{Z}=\mathrm{f}_1) = 0.4 \qquad \mathrm{P}\ (\mathrm{X2}=4 \mid \mathrm{Z}=\mathrm{f}_1) = 0.6\,. \tag{13}$$



(a) (b)

Fig. 1. An example of a hybrid Bayesian network for the evaluation of the new approach. The continuous variables are shown in rounded boxes, the discrete variables are shown in square boxes. The figure 1 (b) is the form of FPBN for a given HBN of (a).

and

$$P\ (X2=2 \mid Z=f_2) = 0.3 \qquad P\ (X2=4 \mid Z=f_2) = 0.7\ . \tag{14}$$

W1' is a continuous variable and has a single probability density for each state of its discrete parent X1. For the sake of simplicity the distribution over W1' is approximated by a conditional Gaussian model with the Gaussian parameters for every state of X1:

$$P\ (W1'=w1' \mid X1=1) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{(w1'-2)^2}{2}\}\ \ (\mu_{w1'}=2, \sigma_{w1'}=1)\ . \tag{15}$$

and

$$P\ (W1'=w1' \mid X1=2) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{(w1'+3)^2}{2}\}\ \ (\mu_{w1'}=-3, \sigma_{w1'}=1)\ . \tag{16}$$

It is possible to assume that the relationship between X2 and W2' is multivariate linear, but different for every state of X2:

$$\text{If } X2 = 2\ ,\ W2' = 5+W1'+\varepsilon\ ,\ \varepsilon \sim N(0\ ,1)\ . \tag{17}$$

and

$$\text{If } X2 = 4\ ,\ W2' = -3+W1'+\varepsilon\ ,\ \varepsilon \sim N(0\ ,1)\ . \tag{18}$$

That yields the following probability distribution over W2':

$$P\ (W2'=w2' \mid W1'=w1'\ , X2=2\ ) = N(w2', 5+w1', 1)\ . \tag{19}$$

and

$$P\ (W2'=w2' \mid W1'=w1'\ , X2=4\ ) = N(w2', -3+w1', 1)\ . \tag{20}$$

For using the PLS method this probability distribution is approximated by conditional Gaussian regression (CGR):

$$P\ (W2'=w2' \mid W1'=w1', X2=2) \approx \sum_{i=-50}^{50} \frac{1}{100} \cdot N(w1', -3+0.2i, 1) N(w2', 7+0.2i, 1)\ . \tag{21}$$

and

$$P(W2'=w2' \mid W1'=w1'\ , X2=4\ ) \approx \sum_{i=-50}^{50} \frac{1}{100} \cdot N(w1', 2+0.2i, 1) N(w2', -6+0.2i, 1)\ . \tag{22}$$

To validate the performances of the proposed approach, we compare the results of the proposed approach with the exact probability density and the approach by Lauritzen. The evidence {X1=1, Z=2} is entered into the network and the probability distribution over W2' is calculated. The result of this experiment is shown in Fig. 2.

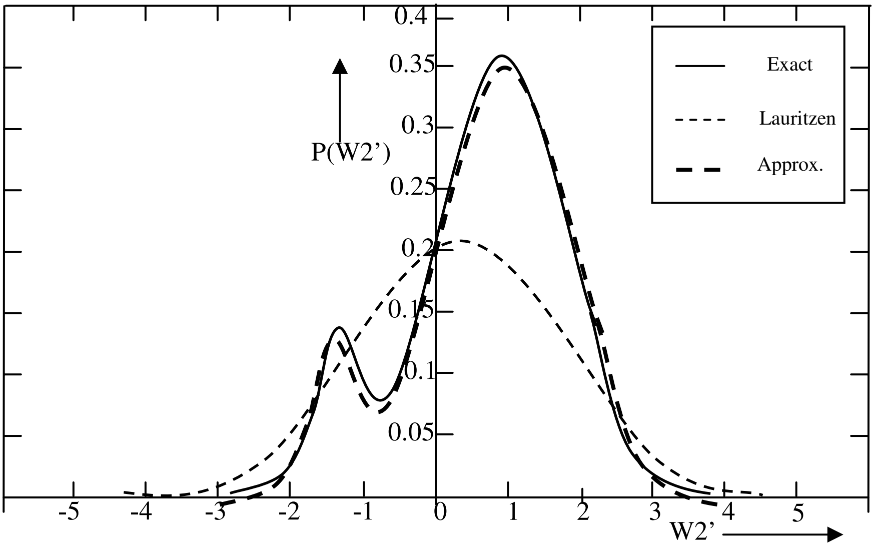


Fig. 2. Given the evidence {X1=1, Z=2}, we compare the probability distribution over W2' yielded by the new approach with the Lauritzen's method

6 Conclusion

Hybrid Bayesian network are the most general form of Bayesian network demanded by practical application. This paper presents a FPBN, a new formalism for evaluating hybrid Bayesian networks using fuzzy sets and partial least-squares. It appears that it is possible to use PLS to estimate only the vector of regression parameters of a continuous node without disturbing others. The FPBN allows arbitrary combination of continuous and discrete variables while making no restrictions on their ancestry. It has been demonstrated to improve the approximation accuracy and network stability.

The distance to an exact solution is only governed by the quality of the approximated likelihood functions. Hence it is possible to make a tradeoff between accuracy and complexity in computation by adjusting the number of fuzzy sets and improving the quality of the parameters estimation method, which are both used to approximate the likelihood functions.

The presented new approach has being used in the context of intention recognition in the battlefield situation assessment domain. Because of its advantage of the flexibility in modeling and the accuracy in the evaluation, a wide field of application is imaginable.

The future important step forwards would be to develop the effective methods for approximating the likelihood functions by a unified notation of the corresponding continuous and discrete likelihood functions, and improve the performance of the present parameter estimation method in the presence of large numbers of continuous variables and missing data.

Acknowledgements

This work is funded by the National ”973” Key Basic Research Development Planning Project (2004CB719401), and also supported by the National Defence”10th–Five-Year” Preliminary Research fund (2002040202).

References

1. Heckerman, D.: Technical Report MSR-TR-95-06. (1995)
2. Jensen, F.V., Lauritzen, S.L., Olesen, K.G.: Bayesian Updating in Causal Probabilistic Networks by Local Computations. Computational Statistics Quarterly. 4 (1990) 269 - 282
3. Driver, E., Morrell, D.: Implementation of Continuous Bayesian Networks Using Sums of Weighted Gaussians. In Besnard and Hanks, (eds). Proc. 11th Conf. Uncertainty in Artificial Intelligence. (1995)
4. Lauritzen, S.L.: Propagation of Probabilities, Means and Variances in Mixed Graphical Association Models. Journal of the American Statistical Association. 87. 1098 - 1108
5. Lauritzen, S.L., Wermuth, N.: Graphical Models for Associations between Variables, Some of which are Qualitative and Some Quantitative. The Annals of Statistics. 17(1) (Mar. 1989) 31 - 57
6. Lauritzen, S.L., Jensen F.: Stable Local Computation with Conditional Gaussian Distributions. Technical Report R-99-2014. Department of Mathematical Sciences. Aalborg University DK
7. Murphy, K.P.: A Variational Approximation for Bayesian Networks with Discrete and Continuous Latent Variables. In: Blackmond-Laskey K. and Prade H. (eds) Proc. of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence. AUAI, Morgan Kaufmann Publishers. (1999) 457 - 466.
8. Lerner, U., Segal, E., Koller, D.: Exact Inference in Networks with Discrete Children of Continuous Parents. In: Proc. of the 17th conference on Uncertainty in Artificial Intelligence. (2001) 319 - 238
9. Lauritzen, S.L., Jensen, F.: Department of Mathematical Sciences. Aalborg University, Research Report R-99-2014. (1996)
10. Heping Pan, Lin Liu: Fuzzy Bayesian Networks- A General Formalism for Representation, Inference and Learning with Hybrid Bayesian Networks. IEEE Transactions. (1999) 401 - 406
11. Nathaniel, A. Woody, Steven, D. Brown: Partial Least-Squares Modeling of Continuous Nodes in Bayesian Networks. Elsevier Science B.V. (2003)
12. Oliver, C. Schrempf, Uwe, D. Hanebeck: A New Approach for Hybrid Bayesian Networks Using Full Densities. In: Budapest and Hungary, (eds), Proc. of the 6th International Workshop on Computer Science and Information Technologies CSIT’. (2004)
13. Jim, F. Baldwin., Enza, Di Tomaso: Inference and Learning in Fuzzy Bayesian Networks. The IEEE International Conference on Fuzzy System. (2003) 630 - 636

π-Net ADL: An Architecture Description Language for Multi-agent Systems

Zhenhua Yu[1,2], Yuanli Cai[2], Ruifeng Wang[2], and Jiuqiang Han[2]

[1] Key Laboratory of Opto-Electronic Technology and Intelligent Control, (Lanzhou Jiaotong University), Ministry of Education, China
[2] School of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an 710049, China
ylicai@mail.xjtu.edu.cn

Abstract. Multi-agent systems (MAS) are studied from the point of view of software architecture. As the existing architecture description languages (ADLs) are difficult to describe the semantics of MAS, a novel architecture description language for MAS (π-net ADL) rooted in BDI model is proposed, which adopts π-calculus and Object-Oriented Petri nets presented in this paper as a formal basis. π-net ADL stresses the description of dynamic MAS architecture, and it is brought directly into the design phase and served as the high-level design for MAS implementation. π-net ADL can visually and intuitively depict a formal framework from the agent level and society level, and analyze, simulate and validate MAS and interactions among agents. Finally, to illustrate the favorable representation and analysis capability of π-net ADL, an example of multi-agent systems in electronic commerce is provided.

1 Introduction

Multi-agent systems (MAS) have been recognized as a main aspect of the distributed artificial intelligence and predicted to be the future mainstream computing paradigm. MAS have been regarded as a new notion and the most promising technology to develop complex software systems, and many effects and attentions have been paid to MAS in complicated, large-scale and distributed industrial and commercial applications [1], [2].

MAS are adaptive and flexible systems where agents may be added or deleted at run-time, and the agent behaviors and interactions among agents may vary dynamically [3], so MAS architecture can evolve during the execution. Hence there exist many difficulties in analyzing the structure and behaviors of MAS. Therefore there is a pressing need for a formal specification to support the design and implementation of MAS, and ensure the developed systems to be robust, reliable, verifiable, and efficient [5]. The formal specification must cope with dynamism and evolution of MAS. Over the past decade, much work has focused on developing practical applications of MAS; however little work has tended to investigate the formal modeling techniques of MAS. So far, there have existed several typical formal specifications for MAS, e.g. dMARS [6], METATEM [7],

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 218–227, 2005.

DESIRE [4], SLABS [8], Gaia [9], MaSE [10], Agent-based G-net [5] and AUML [11]. Despite the important contribution of these formalisms to a solid underlying foundation for MAS, most formal specifications are not oriented for software engineering in terms of providing a modeling notation that directly supports software development and how an implementation can be derived, and less expressive with regard to mental state of agents. It has been recognized that the lack of rigor is one of the major factors that hamper the wide-scale adoption of multi-agent technology [4].

In this paper, to support the development of correct and robust dynamic MAS, a novel architecture description language (ADL) [12] based on π-calculus [16] and Object-Oriented Petri nets (π-net ADL) is proposed. Our proposed formalism studies MAS from the point of view of software engineering. The mainstream software engineering tools, techniques and formal languages (such as software architecture, architecture description language) are adopted to go systematically from system requirements to MAS implementation. The reason π-net ADL adopts π-calculus and Object-Oriented Petri nets (OPN) as its formal basis is that OPN is graphical and mathematical modeling tool which is simplicity and strong expressive power in depicting dynamic system behaviors, and π-calculus is a process algebra specially suited for the description and analysis of concurrent systems with dynamic or evolving topology. π-net ADL supports formal analysis of MAS architecture in a variety of well-established techniques, such as simulation, reachability analysis, model checking and interactive proving. As the Belief-Desire-Intention (BDI) model is well suited for describing an agent's mental state, π-net ADL roots in the BDI formalism. π-net ADL is oriented for software engineering and places emphasis on practical software design methods instead of reasoning theories. The ultimate goal of π-net ADL is to provide a tool that generates executable implementation skeletons from a formal model and enables software engineers to develop reliable and trustworthy MAS.

2 π-Calculus and Object-Oriented Petri Nets (OPN)

2.1 Object-Oriented Petri Nets

Petri nets are a graphical and mathematical modeling tool applicable to many systems that exhibit concurrency and synchronization [13]. Object-Oriented Petri nets combining Petri nets with Object-Oriented methods can tersely and independently represent all kinds of resources in a complex system, increase the flexibility of the model. In the OPN model, a system is composed of mutually objects and their interconnection relations; the formal definition is given as follows.

Definition 1. *A system is a 2-tuple, S=(O, MPR), where O is a finite set of OPN models of physical object in the system, $O = \{O_1, O_2, \ldots, O_i\}$; MPR is a finite set of message passing relations among physical objects (i.e. OPN models).*

The OPN model of a phisical object i, O_i, is defined as follows.

Definition 2. *O_i is a 9-tuple, $O_i = (P, IP, OP, T, F, IIA, OIA, E, C)$, where P is a finite set of physical object places in the system, $P = \{p_1, p_2, \ldots, p_j\}$; IP (Input Place) and OP (Output Place) are sets of input and output message places in OPN, $IP = \{ip_1, ip_2, \ldots, ip_l\}$, $OP = \{op_1, op_2, \ldots, op_m\}$; T is a finite set of physical object transitions in the system, $T = \{t_1, t_2, \ldots, t_k\}$; $F \subseteq (P \times T) \bigcup (T \times P) \bigcup (IP \times T) \bigcup (T \times IP) \bigcup (OP \times T) \bigcup (T \times OP)$ is the input and output relationships between transitions and places; IIA (Input Interface Arc) is a set of the input interface arc from outside to OPN, $IIA = \{iia_1, iia_2, \ldots, iia_n\}$ [14]; OIA (Output Interface Arc) is a set of output interface arc from OPN to outside, $OIA = \{oia_1, oia_2, \ldots, oia_o\}$ [14]; $E : F \rightarrow (ID, CDS)$ is expression functions in the arcs, ID is the identification of the arc and CDS is a complicated data structure; C(P) is a set of color associated with the places P, $C(P) = \{cp_1, cp_2, \ldots, cp_j\}$; C(IP) and C(OP) are sets of color associated with the input and output message places.*

In a system, the interaction among the distinct objects (i.e. OPN) is specified by MPR, which can be defined as follows.

Definition 3. *MPR is defined as MPR=(ILP, C), where ILP is the Intelligent Linking Place denoted by ellipse. The information obtained from the external is saved in the ILP. Each OPN dispatches the information by ILP. C(ILP) is a set of color associated with the ILP.*

In the OPN model, some concepts of CPN are employed and some behavioral semantics does not violate the semantics of CPN formalism. IIA and OIA are interface arcs in OPN that are the interfaces of an object interacting with the external environment or other objects.

2.2 π-Net

π-calculus [16] addresses the description of system with a dynamic or evolving topology, and permits their analysis for bisimilarity and other interesting properties.

OPN are suitable for describing the large-scale, complicated and distributed MAS. However the structure of OPN is static, it is hardly possible to model dynamic MAS architecture. π-calculus is suitable for describing dynamic MAS architecture where agents can be dynamically created and removed, and interactions between agents are also dynamically established and modified, leading to an evolving communication topology. Therefore, the advantages of π-calculus and OPN are combined to propose π-net, in which OPN are used to describe the static and dynamic semantics, analyze the deadlock and reachability of a system, and simulate the system; and π-calculus addresses the description of systems with a dynamic or evolving topology. π-calculus and OPN can complement each other very well.

3 Novel Architecture Description Language for Multi-agent Systems (π-Net ADL)

In order to accurately describe MAS architecture, π-net is adopted as a formal basis of ADL. π-net ADL adopts computing agents and connecting agents as the computation and interaction elements in MAS. π-net ADL studies MAS from the agent level and society level: The agent level to the structure of each agent, and the society level to a formal framework for MAS and interactions among agents.

π-net ADL is a 3-tuple, π-net ADL=(Computing agents, Connecting agents, Configurations). The items in the definition are described in detail as follows.

3.1 Modeling Computing Agents

Computing agents are a finite set of the computing agent in MAS. Computing agents are responsible for interacting with users and environment to provide specific applications.

A computing agent is based upon the Belief-Desire-Intention (BDI) model, which is used to describe its mental states. A computing agent is a 3-tuple, Computing Agent = (ID, AS, EL), where ID is the identifier of a computing agent; AS (Agent Structure) is the OPN model which defines the interfaces and internal implementation of a computing agent. EL addresses the evolvement of computing agent which is described by π-calculus. AS based on the BDI model is composed of the Knowledge-base module, Goal module, Plan module, and Interface module. The modules are described as follows: in practical terms, the Knowledge-base module corresponds to the agent's Beliefs, and it may be represented as simple variables and data structures or, complex systems such as knowledge bases [15]. The Goal module corresponds to the agent's Desires, and it may be associated with a value of a variable, a record structure, or a symbolic expression in some logic so that desires can be prioritized [15]. The Plan module corresponds to the agent's Intentions, which is a list of plans and describes the actions achieving the Goal values of an agent [15]. The interface module allows a computing agent to interact with other agents and the environment, and is used to send and receive messages between agents. A computing agent OPN model is shown in Fig. 1, where Private Utilities represent private method and utilities, such as register and destroy information; the Knowledge-base, Goal and Plan are denoted by ellipses. As Fig. 1 only represents a template of a computing agent, interfaces and internal implementation are added according to the specific system requirements, and the BDI model can be refined further.

For simplicity and clarity of the diagrams, only names of places, transitions and arcs of all agents models are presented in this paper, and inscriptions, colors, guards and marking are left unspecified.

Agents communicate with other ones by message passing, which follows speech act theory and uses complex protocols to negotiate [5], e.g., the FIPA agent communication language(ACL) and KQML. Communication is the basis

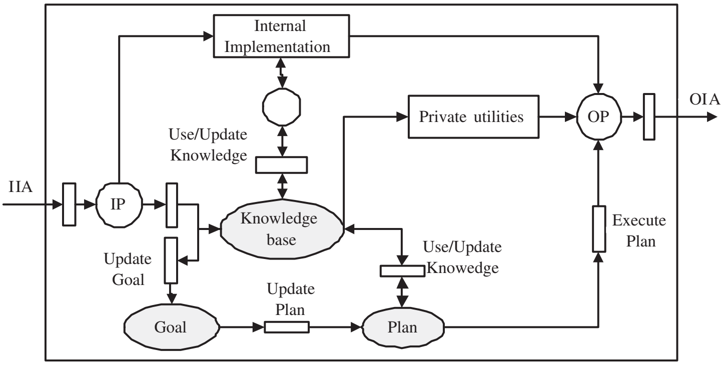


Fig. 1. The OPN model of a computing agent

for interaction and organization without which agents would be unable to cooperate, coordinate, or sense changes in their environment. The agents proposed in this paper speak and understand FIPA ACL.

3.2 Modeling Connecting Agents

Connecting agents are a finite set of the connecting agent which is communication facilitator dealing with the interaction information among agents and defining the rules that govern those interactions.

A connecting agent is defined as Connecting Agent= (MPR, KBP, T, F, Role, EC), where MPR (Message Passing Relations) is the tuple in a system model; KBP represents Knowledge Base Place which is defined to apperceive the external environment, acquire requisite knowledge, and describe services which computing agents provide via interfaces. The roles of a connecting agent identify the logical participants in the interaction represented by the connecting agent, and specify the expected behavior of each participant in the interaction. Role is defined as $Role = \{CID_1, \ldots, CID_n\}$, where CID_i is the identifier of an interface in a computing agent. The services provided by the role are stored in the KBP. There are two types of roles, static and dynamic role respectively. Dynamic role will change with the computing agent deleted or added. EC addresses the evolvement of connecting agent which is described by π-calculus.

In MAS, computing agents first enroll their information (such as name, address, interface and capability) in connecting agents, and the enroll process [17] is

$$EnrollService = (vid)\overline{enroll}(id) \cdot id \cdot s$$

The corresponding enroll process in the connecting agent is

$$EnrollService = (vx)enroll(x) \cdot \overline{x} \cdot s$$

In π-net ADL, computing agents and connecting agents describe agent structure from the agent level, as well as the behaviors and interfaces of the individual agent.

3.3 Modeling Configurations

MAS configurations are connected graphs of computing agents and connecting agents that describe architectural structure. Explicit architectural configurations facilitate communication among a system's many stakeholders, who are likely to have various levels of technical expertise and familiarity with the problem at hand [12].

The multi-agent systems architecture can not only describe individual agent, but also depict the whole system and interaction among agents. The multi-agent systems architectural configuration based on π-net ADL is shown in Fig. 2, and the MAS are studied from the society level, where MAS are conceived as a multitude of interacting agents. In the society level, the key point is the overall behaviors of the MAS, rather than the mere behaviors of individuals. For simplicity and clarity of the diagrams, this model is predigested. The computing agents are represented by IIA, OIA, IP, OP and abstract transitions denoted by shaded rectangles. The abstract transitions can be refined. This architecture consists of four computing agents and one connecting agent.

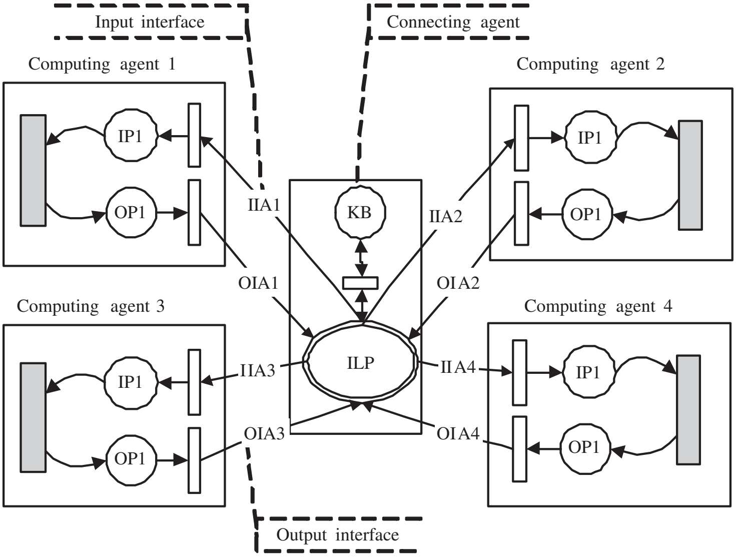


Fig. 2. MAS architectural configurations

The computing agent connects with the connecting agent by the interface; therefore an arborescent topology is formed. The static semantics of the multi-agent systems architecture is described in Fig. 2, and the dynamic semantics of the multi-agent systems architecture is controlled by the firing rule. The firing of the transition makes the Token dispatch, which expresses the message passing and well depicts interactions among agents. The purpose of modeling multi-agent systems in π-net ADL is to make full use of the well-established analysis methods proposed for Petri nets. These methods are commonly used to detect the deadlock, and boundedness properties of systems models. π-net ADL can systematically analyze, verify and validate the properties of the implemented system.

π-net ADL is a visual ADL, which can make users effectively understand and analyze MAS before MAS are implemented, and narrow the gap between agent formalism and practical systems.

4 An Example of Multi-agent Systems in Electronic Commerce

In this section, a multi-agent system in an electronic commerce is considered. The buyer agents and seller agents negotiate price, and finally the buyer agents determine whether to buy or not. The MAS architecture based on π-net ADL is set up, and then the model is analyzed by mathematical methods of Petri nets to ensure a correct design.

4.1 MAS Modelling in Electronic Commerce

The architecture of the price negotiation MAS in electronic marketplace based on π-net ADL is shown in Fig. 3, which represents a pair wise negotiation process. At the beginning of negotiation, the MAS is composed of two functional agents (one buyer agent and one seller agents) bargaining for goods. For simplicity, the buyer agent and seller agent only remain knowledge-base place, which is regarded as abstract place and can be refined according to the specification. Some constraints are omitted in this figure.

The buyer agent model consists of three states: Inactive, Waiting and Thinking. The buyer agent begins the negotiation, which has an extra state (Waiting) and timing machinery not present in seller agent. In this model, the buyer and seller agent are initially waiting in the Inactive place. Each agent must register its basic information to the connecting agent. The connecting agent can accept or reject the registration based on the enrolled agent's reputation or function. The buyer agent starts the negotiation process by sending a call for proposals to the seller agent, and its state changes from Inactive to Waiting. The seller receives the message from ILP through the IIA (IRecMes), and then deals with the message and sends the responses (Accept Proposal or Refuse Proposal). If the seller agent accepts the proposal, the buyer agent's state changes from Waiting to Inactive and the negotiation is finished; if the seller agent refuses the proposal, the buyer agent's state changes from Waiting to Thinking, then it sends a new proposal.

During the execution of the system, its architecture may be dynamically changed, e.g. other buyer agents may want to join the system; there are some faults in the seller agent, so the backup seller agent is added and used; and an additional seller agent is needed to improve the computation speed. The dynamic architecture is represented by π-net ADL as follows.

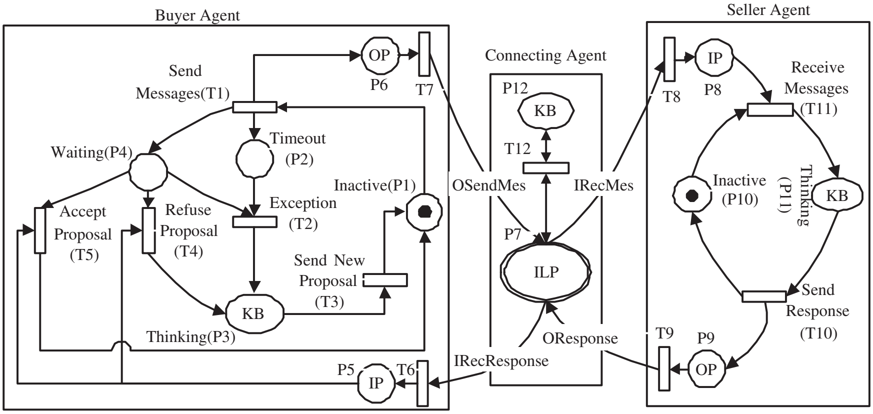


Fig. 3. MAS model in electronic commerce

The creation process of agents is shown as follows.

$$Agent_i ::= \tau\_NEW\_IAgent\_Agent$$

When another seller agent is added to improve the service quality, in order to balance the load of two seller agents, the architecture is needed to reconfigure and the reconfigurable process is shown as follows.

$$\begin{aligned} buyer.request(r) <> \ & \mathbf{if}(seller[1].n \leq seller[2].n) \\ & \mathbf{then} \quad seller[1].serve(r) \\ & \mathbf{else} \quad seller[2].serve(r) \end{aligned}$$

From the modelling process of the dynamic electronic commerce system, π-net ADL effectively describes the dynamic architecture.

4.2 Analysis of MAS Model in Electronic Commerce

A significant advantage provided by π-net ADL based on Petri nets is that the verification and validation of the model can be accomplished before implementation, and help ensure a correct design (such as liveness, deadlock freeness, boundness and concurrency) with respect to the original specification to enable software engineers to develop reliable and trustworthy MAS. In this section, the deadlock of the MAS model is analyzed. It is important that how to handle deadlock situations for development of electronic commerce systems and operating systems, where the communication plays a key role. Due to limited space, other properties will be discussed in our future working paper.

The theory of invariants [13] is employed as the deadlock detection method to analyze the simplified MAS model.

Theorem 1. *Let N is a Petri net model, an n-vector I is a P-invariant (place invariant) of N if and only if* $I^T \cdot [N] = 0^T$. $\|I\| = \{p \in P | I(p) \neq 0\}$ *is called the support of an invariant. If all P-invariants are marked in the initial marking and there are no empty siphons, the N is live[13].*

By analyzing, there are four P-invariants in the MAS model, and their supports are $\|I_1\| = \{P12\}$, $\|I_2\| = \{P1, P2, P3, P5, P6, P7, P8, P9, P11\}$, $\|I_3\| = \{P10, P11\}$, and $\|I_4\| = \{P1, P3, P4\}$ respectively. All P-invariants are marked in the initial marking; moreover there are no empty siphons, so the model is live.

Deadlock analysis can help eliminate human errors in the design process, and verify some key behaviors for the MAS model to perform as expected, and increase confidence in the MAS design process.

5 Conclusions

Multi-agent systems are regarded as the most promising technology to develop complex software systems. In this paper, from the software architecture point of view, a novel architecture description language for MAS (π-net ADL) rooted in π-net is proposed to support the modelling and analysis of multi-agent systems. π-net ADL based on Belief-Desire-Intention (BDI) agent model stresses practical software design methods instead of reasoning theories, and analyze the static and dynamic semantics, and depict the overall and individual characteristics of MAS. π-net ADL can be applied to investigate MAS from the agent level and society level. Finally, an example of an agent society in electronic marketplace is used to illustrate modelling capability of π-net ADL; and moreover, how to detect the deadlock in the MAS model by the theory of invariants is discussed. π-net ADL, as a visual ADL, can promote the intercourse and understand among clients, architecture designers and developers, and provide an effective modelling method for MAS modelling and verifying.

Acknowledgements

This work is supported by the Opening Foundation of the Key Laboratory of Opto-Electronic Technology and Intelligent Control (Lanzhou Jiaotong University), Ministry of Education, China. Grant NO. K04110.

References

1. Wooldridge, M.J., Jennings, N.R.: Agent Theories, Architectures, and Languages: a Survey. Lecture Notes in Artificial Intelligence. Springer-Verlag. Berlin Heidelberg New York. **890** (1995) 1–32
2. Wooldridge, M.J., Jennings, N.R.: Intelligent Agents: Theory and Practice. Knowledge Engineering Review. **10** (1995) 115–152
3. Jiao, W., Zhou, M., Wang, Q.: Formal Framework for Adaptive Multi-Agent Systems. In Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology. Beijing (2003) 442–446

4. Brazier, F.M.T., Dunin-Keplicz, B.M., Jennings, N.R., Treur, J.: Desire: Modelling Multi-Agent Systems in a Compositional Formal Framework. International Journal of Cooperative Information Systems. **1** (1997) 67–94
5. Xu, H., Shatz, S.M.: A Framework for Model-Based Design of Agent-Oriented Software. IEEE Transactions on Software Engineering. **29** (2003) 15–30
6. Luck, M., d'Inverno, M.: A Formal Framework for Agency and Autonomy. In Proceedings of the First International Conference on Multi-Agent Systems. San Francisco. (1995) 254–260
7. Fisher, M., Wooldridge, M.: On the Formal Specification and Verification of Multi-Agent Systems. International Journal of Cooperative Information Systems. **1** (1997) 37–65
8. Zhu, H.: Slabs: A Formal Specification Language for Agent-Based Systems. International Journal Software Engineering and Knowledge Engineering. **11** (2001) 529–558
9. Wooldridge, M., Jennings, N. R., Kinny, D.: The Gaia Methodology for Agent-Oriented Analysis and Design. International Journal of Autonomous Agents and Multi-Agent Systems. **3** (2000) 285–312
10. DeLoach, S.A.: Multiagent Systems Engineering. In Proceedings of Agent Oriented Information Systems. Stockholm. (2000) 45–57
11. Odell, J., Parunak, H.V.D., Bauer, B.: Representing Agent Interaction Protocols in UML. In Proceedings of the First International Workshop on Agent Oriented Software Engineering. Limerick. (2001) 121–140
12. Medvidovic, N., Taylor, R.N.: A Classification and Comparison Framework for Software Architecture Description Languages. IEEE Transactions on Software Engineering. **26** (2000) 70–93
13. Murata, T.: Petri Nets: Properties, Analysis, and Application. Proceedings of the IEEE. **77** (1989) 541–580
14. Saldhana, J.A., Shatz, S.M.: Formalization of Object Behavior and Interactions from UML Models. International Journal of Software Engineering and Knowledge Engineering. **11** (2001) 643–673
15. Kavi, K.M., Aborizka, M., Kung, D.: A Framework for Designing, Modeling and Analyzing Aagent Based Software Systems. In Proceedings of the Fifth International Conference on Algorithms and Architectures for Parallel Processing. Beijing (2002) 196–200
16. Sangiorgi, D., Walker, D.: The Pi-Calculus: a Theory of Mobile Processes. Cambridge University Press (2001)
17. Jiao, W., Shi, Z.: A Dynamic Architecture for Multi-Agent Systems. In Proceedings of Technology of Object-Oriented Languages and Systems. Nanjing (1999) 253–260

Automatic Construction of Bayesian Networks for Conversational Agent

Sungsoo Lim and Sung-Bae Cho

Dept. of Computer Science, Yonsei University,
Shinchon-dong, Seodaemun-ku,
Seoul 120-749, Korea
{lss, sbcho}@cs.yonsei.ac.kr

Abstract. As the information in the internet proliferates, the methods for effectively providing the information have been exploited, especially in conversational agents. Bayesian network is applied to infer the intention of user's query. Since the construction of Bayesian network requires large efforts and much time, an automatic method for it might be useful for applying conversational agents to several applications. In order to improve the scalability of the agent, in this paper, we propose a method of automatically generating Bayesian networks from scripts composing knowledge base of the conversational agent. It constructs the structure of hierarchically composing nodes and learns the conditional probability distribution table using Noisy-OR gate. The experimental results with subjects confirm the usefulness of the proposed method.

Keywords: Conversational agent, Script, Bayesian network, Hierarchical structure, Noisy-OR gate.

1 Introduction

Conversational agent is a system that exchanges information between user and agent using natural language dialogue [1,2]. The goal of conversational agent is to provide the most proper script from the conversation database.

Recently, the interest in discovering knowledge represented in Bayesian networks [3,4,5,6] is increasing because Bayesian networks can handle incomplete data sets and facilitate the combination of domain knowledge and data. And there are also several cases in which Bayesian network is applied to conversational agent.

Although conventional conversational agents use pattern matching techniques to reply to the input query [7,8], Bayesian network is applied to analyze user's queries more precisely and to model dialogues. It shows good performance in inferring the intention of a user [9,10]. However, Bayesian network is not easy to design so even experts need much time for the construction of it. Moreover, the network is dependent on the domain of application. When the domain changes, we should modify or reconstruct the network. These difficulties debase the scalability of conversational agents [10]. In this paper, we propose an automatic Bayesian network constructing

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 228 – 237, 2005.

method from scripts in order to increase the scalability of conversational agent. It might be useful for novices to design the conversational agent, since it is only duty to provide scripts.

This paper is organized as follows: We begin Section 2 with the introduction of traditional conversational agent and the agent which uses Bayesian network in order to infer the intention of user's queries. In section 3, we propose how to construct the structure and the parameters of Bayesian network in the conversational agent with an automation mechanism. In section 4, we present the experimental results of the proposed method. Finally, section 5 describes the summary and future works.

2 Conversational Agent

2.1 Traditional Conversational Agent

As an alternative for the usual interfaces of web sites, conversational agents are recently being developed because of the capability of conversations with users by natural language [11]. Eliza, one of the first conversational agents, was born at Massachusetts Institute of Technology in 1966. Eliza was contrived for the research on natural language processing. It uses a simple pattern matching technique [8]. For example, if user inputs a sentence including the word 'name', Eliza answers 'My name is Eliza.' which matches the word 'name.' However, it can reply only short conversations because Eliza neither models user nor keeps the state of conversation.

ALICE (Artificial Linguistic Internet Computer Entity, http://www.alicebot.org) is written in a language called AIML (Artificial Intelligence Markup Language) that is based on XML. By using sequential pattern matching, it enhances the efficiency of analyzing sentences. However, it has shortcomings of not being able to respond to users reflecting their intentions because of simple sequential pattern matching based on keywords. Tackling this problem requires much time and effort to construct the response database.

There are a few products on sale. Nicole made by Native Minds is working as a virtual agent in cyber space (www.nativeminds.com). This agent provides website information through conversation with users. And it also shows many human facial expressions that react to user's conversation. Nicole is not the only one; others available include SmartBot of Artificial Life Company (www.artificial-life.com), Verbot of Virtual Personalities Company (www.vperson.com) and so on.

2.2 Conversational Agent Using Bayesian Network

Since the pattern matching has limitations to manage the uncertainties such as elliptic words, useless words and duplicated information, it needs to model dialogues [12]. Bayesian network is one of the modeling tools, and applied well to conversational agents [8, 13].

Bayesian probabilistic inference is a famous model for inference and representation of the environment lacking in information. Nodes of Bayesian network represent random variables, while an arc represents the dependency between variables. For the inference of the network, the structure and the probability distribution need to be specified in advance. Usually the structure and the probability distribution are calculated by experts or collected data from the domain. In conversational agents, when a query appears and the agent observes some evidence, Bayesian inference algorithm computes the probabilities of nodes based on the conditional probability table and independence assumption. Representing and storing dialogue information on the network enhances the functionality of the conversational agent.

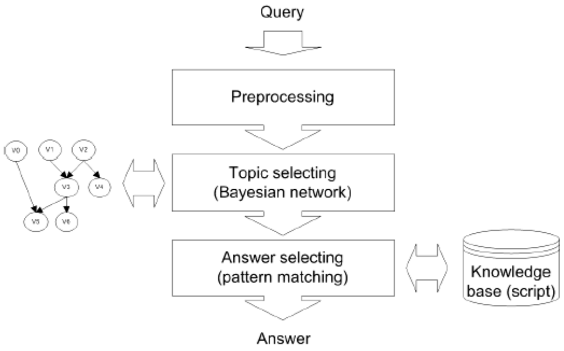


Fig. 1. Structure of the conversational agent using Bayesian network

We design the conversational agent as shown in Fig. 1. The Bayesian network is used to infer the topic of a user's query and model the context of dialogue. This leads to a definition of the scope of the dialogue. Since Bayesian networks are based on graph theory, they are effective in inference and representation. In this paper, the Bayesian network is hierarchically constructed with three levels based on function: keyword, topic and sub-topic [14]. First level is for the keywords used in the domain, and the individuals and attributes of the domain from topics. Sub-topic represents the individual that has determined its attributes. This hierarchical modeling helps to hold a conversation to understand a detailed user's intention. Fig. 2 presents how Bayesian network works in conversational agents.

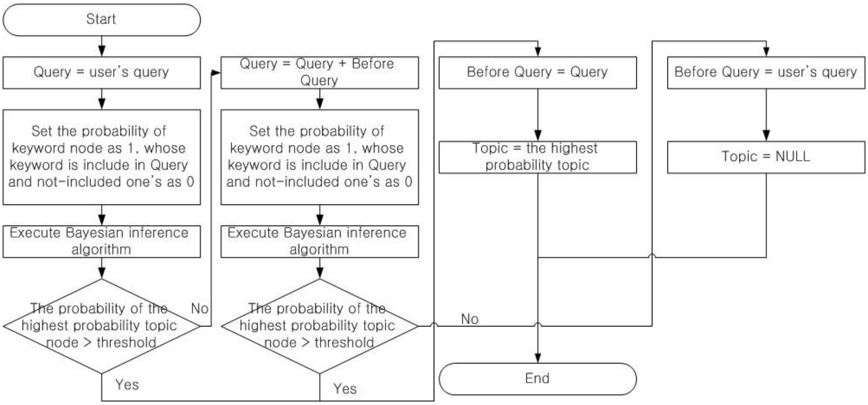

Start
Query = user's query
Set the probability of keyword node as 1, whose keyword is include in Query and not-included one's as 0
Execute Bayesian inference algorithm
The probability of the highest probability topic node > threshold
No
Yes
Query = Query + Before Query
Set the probability of keyword node as 1, whose keyword is include in Query and not-included one's as 0
Execute Bayesian inference algorithm
The probability of the highest probability topic node > threshold
No
Yes
Before Query = Query
Topic = the highest probability topic
Before Query = user's query
Topic = NULL
End

Fig. 2. Inference procedure of Bayesian network in conversational agent

Usually people do not give just one query to express their intention, and they produce a query based on previous conversation [15,16]. Therefore, if the topic selector cannot infer the topic of query, it infers the topic referring to the previous query.

After selecting the topic, the agent compares script by matching patterns with the keywords in scripts and the keywords in user's queries and chooses the answer whose topic is same as the topic selected. The script is basic knowledge of the conversational agent for answering user's queries. When there are many scripts, performance declines because of the redundancy of information. In this paper, we divide the scripts into several groups based on their topics. This reduces the number of scripts to be compared. A script is stored as an XML form, and Fig. 3 shows the definition of a script.

```
[script] := "<SCRIPT>" [topic] [keyword]+ [answer]+ "</SCRIPT>"
[topic] := "<TOPIC>" [word]+ "</TOPIC>"
[keyword]:= "<KEYWORD>" [word]+ "</KEYWORD>"
[answer] := "<ANSWER>" [word]+ "</ANSWER>"
```

Fig. 3. Script BNF

In script, [topic] that is the topic of user's queries represents top topic in Bayesian network, [keyword] that is the keywords in user's queries represents the pattern of user's question, and [answer] means reply which the agent gives to user.

3 Automatic Construction of Bayesian Network

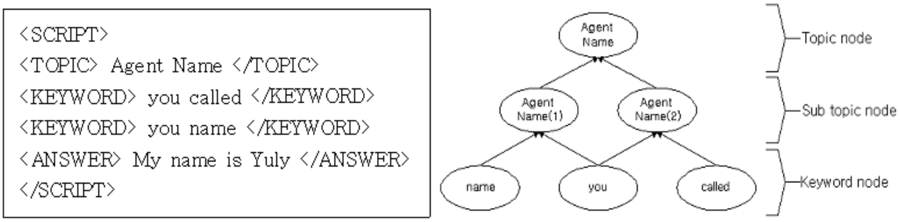
Although the Bayesian network is useful, as we discussed above, it takes much time and efforts to construct a Bayesian network. Therefore, we propose a method of constructing Bayesian network automatically.

A Bayesian network is represented by $BN = < N, A, \Theta >$, where $< N, A >$ is a directed acyclic graph (DAG): each node $n \in N$ represents a domain variable, and each arc $\alpha \in A$ between nodes represents a probabilistic dependency between the associated nodes. Associated with each node $n_i \in N$ is a conditional probability distribution (CPT), collectively represented by $\Theta = \{\theta_i\}$, which quantifies how much a node depends on its parents.

The construction of Bayesian network consists of two parts: one is generating the network structure depending on domains and the other is parameter learning (CPT). Because the conversational agent depends on domains, we cannot obtain general log data, so that we use scripts for the data of constructing Bayesian network. The structure is generated by the rules made from analyzed scripts, and the parameters are obtained by using Noisy-OR gate which is a good method for small data sets.

3.1 Structure Generation

The causality of Bayesian network used in conversational agent is relatively clear, and we approach the structure generation based on grammars defined on the scripts. We propose a method of automatically generating the structure of Bayesian network from the script. Moreover, it constructs the network composed as the hierarchical node structure to model the detailed intention of users.



```
<SCRIPT>
<TOPIC> Agent Name </TOPIC>
<KEYWORD> you called </KEYWORD>
<KEYWORD> you name </KEYWORD>
<ANSWER> My name is Yuly </ANSWER>
</SCRIPT>
```

Fig. 4. Generation of Bayesian networks based on the script

A *topic node* is obtained from the <TOPIC> of a script, which is an objective variable of the inference. *Keyword nodes* are generated from the <KEYWORD> of the script while a <KEYWORD> constructs one *sub topic node*. And a *mid topic node* is made by bifurcating *sub topic nodes*. Fig. 4 shows that an example script can be interpreted to generate the network.

The calculation of parents' probabilities requires exponential computation as the number of children. Therefore, the bifurcation of nodes might be useful to avoid much computation. Fig. 5 shows a simple process of the bifurcation. The maximum number of the children of a parent is limited as three in this paper, so that it divides nodes into two sub trees when a parent node has four children.

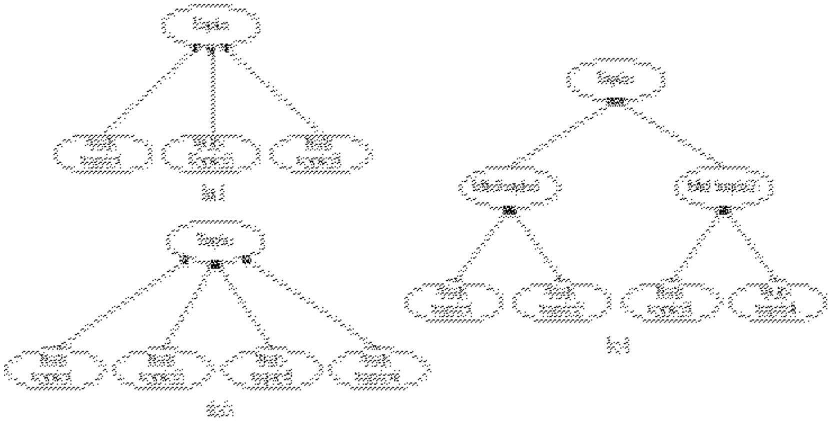


Fig. 5. Bifurcation of node (a) Initial state, (b) Sub topic addition, (c) Nodes bifurcation

When it does not divide nodes, 2^k calculations are necessary to obtain the probability of a parent, which has *k sub topic nodes*. On the other hand, when it divides nodes, only $(k–1)/(n–1)×2^n$ calculation is needed as shown in Fig. 6. By the bifurcation, the computation is reduced from $O(2^k)$ to $O(k)$.

ASSUMPTION : All nodes have at least $n-1$ children except keyword nodes

DEFINITION :

n : the number of maximum children

h : the height of Bayesian network from root to parent of sub topic

l : the number of nodes which need to calculate the probability

k : the number of sub nodes

CALCULATION of T :

If $height = 1$, then $l = 1$

If $height = 2$, then $l = 1 + n$

If $height = 3$, then $l = 1 + \mathrm{n} + \mathrm{n}^2$

$\cdots$

If $height = h$, then $l = 1 + n + n^2 + \cdots + n^h = \dfrac{1 \times \{n^h - 1\}}{n-1} = \dfrac{k-1}{n-1} (\because n^h = k)$

The total computation T :

$$T = l \times 2^n = \frac{k-1}{n-1} \times 2^n = \mathrm{O}(k)$$

Fig. 6. Calculation of the total computation T

3.2 Parameter Learning

In many cases there are not plenty of samples to generate CPT. In conversation agent, when the domain of agent changed, we must regather the conversations for learning data, but it takes much time and effort. So, we adopt a leaky Noisy-OR gate [17] to learn parameters of Bayesian network, which is a popular parameter learning technique. The leaky Noisy-OR gate defines x_i as the cause (children), and y as the result (parent). p_o, called leaky probability, is the probability of y when no evidence is present, and p_i is the probability of y when only x_i is used as an evidence.

$$p_0 = \Pr(y \mid \bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n)$$

$$p_i = \Pr(y \mid \bar{x}_1, \bar{x}_2, \cdots, \bar{x}_i, \cdots, \bar{x}_n)$$

The probability of the result about the subset X_p composed of the x_i is given in the following formula by the leaky Noisy-OR gate:

$$\Pr(y \mid X_p) = 1 - (1 - p_0) \prod_{i:x_i \in X_p} \frac{1 - p_i}{1 - p_0}$$

We only set the probability of p_i. Here we assign the value of 0.001 to the leaky probability p_0, and the probability of *topic nodes* and *mid topic nodes* are calculated as follows. Where n is the number of child nodes and α indicates the weight.

$$p_i = \alpha + \frac{1-\alpha}{n} \qquad (1)$$

The probability of a *sub topic node*, where child node is *keyword node*, can be obtained by the following formula, while m is the number of the parents of x_i.

$$p_i = \alpha + \frac{1-\alpha}{mn} \qquad (2)$$

Because a node closer to the *topic nodes* should be more effective, α is rather higher for nodes closer to *topic nodes*. In this paper, we set α as 0.5 for *sub topic nodes*, and we increase it by a linear function. Since children affect the same amount of effects to parents, we use divide-operation. So in formula (1), we divide by n, and in formula (2), we divide by mn, it is because the keyword node can has more than one parent.

4 Experiments

4.1 Illustration

In this paper, a query is divided into two types: one is a sufficient information query and the other is a use-previous-information query. That is because in common conversation an ellipsis frequently happens, and speaker usually uses background knowledge. The proposed conversational agent can deal with these types of queries.

The sufficient information query is the query that contains all the information to estimate user's intention. In this case, the agent gives a proper answer obtained by the inference of the system. And the use-previous-information query is the query that has ellipsis (or use major terms). In this case, the agent needs additional information about query. Here the agent uses previous query for additional information.

Fig. 7 shows examples of conversation that the conversational agent can process. The upper is the case of the sufficient information query and the lower is the case of the use-previous-information query.

User: What is your name?
    keyword (your, name)
    topic selector (Agent name)
    answer selector (My name is Yuly)
Yuly: My name is Yuly.

User: It is pretty.
  keyword (pretty)
  topic selector (Agent name)
  answer selector (Thank you very much)
Yuly: Thank you very much.

Fig. 7. Conversation Example of the two types

For the dialogue of sufficient information, the preprocessor extracts the words "your" and "name" and the words as the inputs of topic selector. The topic selector selects "Agent name" as its topic, and then the answer selector selects "My name is Yuly" as a response.

For the dialogue of use-previous information, on the other hand, the preprocessor extracts a word "pretty," so the topic selector cannot select a topic whose probability is above threshold. Hence the topic selector uses additional information from the previous query "What is your name?". It reprocesses with keywords "pretty," "your" and "name" so it selects "Agent name" as its topic.

4.2 Experiment Result

We adopt the measure as shown in Fig. 8 to demonstrate the proposed method. Since Bayesian network decides which topic a query is categorized into, the result of the inference is a set of the probabilities of topic nodes. It can be sure of the classification, if only the targeting topic node (T_o) has the highest probability among them. The more difference occurs against others (T_k), the more correctly a query is classified. We use the classification performance as a primitive evaluation measure.

DEFINITION:

N : the number of topics

T_0 : the probability of the topic node corresponding to user's query

T_k : the probability of the k th topic node

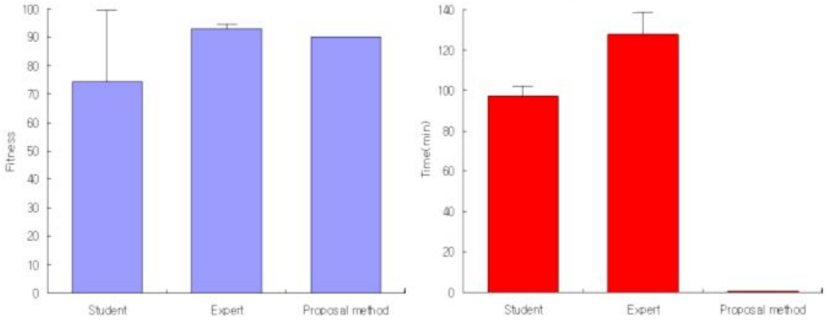
Fitniss measure:

$$M = \begin{Bmatrix} \text{if } \max(T_k) = T_0, & \left(T_0 - \sum_{k=1}^{n} D_k\right) \times 100(\%) \\ \text{otherwise,} & 0 \end{Bmatrix}$$

$$D_k = \begin{Bmatrix} \text{if } T_0 - T_k < a, & \dfrac{a-(T_0-T_k)}{a \times N} \\ \text{otherwise,} & 0 \end{Bmatrix}$$

Fig. 8. Evaluation measure

The experiments have been performed by 10 graduate students majoring in computer science and 5 conversational agent experts. They generated Bayesian network based on eleven topics. The way to construct Bayesian network and use the GENIE tool [18], which is one of the popular tools for designing Bayesian network, has been explained to each subject. Fig. 9 shows the result of the experiments representing comparisons between the proposed method and the manual design.



(a) (b)

Fig. 9. Comparative results with the manual design. (a) Bayesian Network Fitness (a = 0.1) (b) Generating Time.

As the result shows, the students' group constructed Bayesian network that has 74.4% fitness in average and it took 97.1 minutes to design the network, and the experts' group constructed Bayesian network that has 92.8% fitness in average and it took 127.6 minutes. On the other hand, the proposed method constructs the result that had 90% fitness within a few seconds. The experts' group spent more time than students' group. It's because the experts made Bayesian Network more carefully and they analyzed the script with more detail.

5 Concluding Remarks

It is necessary to design the network first in building the conversational agent using Bayesian network. However, it is difficult for the beginners, and even experts need much time and cost for the analysis and design of the target application. In this paper, we have proposed an automatic construction method for conversational agents using Bayesian network. A usability test showed that the proposed method is efficient in time and accuracy.

For the more accurate inference, not only automatic construction but also learning mechanism is necessary for conversational agents using Bayesian network. It also requires convenient interface to design Bayesian network for novices in the construction of conversational agents.

Acknowledgement

This paper was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Commerce, Industry and Energy.

References

1. Lee, S. –I., Sung, C. and Cho, S. –B.: An effective conversational agent with user modeling based on Bayesian network. Lecture Note in Computer Science 2198 (2001) 428-432
2. Macskassy, S. and Stevenson, S.: A conversational agent. Master Essay, Rutgers University (1996)
3. Heckerman, D.: A tutorial on learning Bayesian networks. Tech. rep., Microsoft Research, Advanced Technology Division (1995)
4. Cooper, G. F.: A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. Data Mining and Knowledge Discovery 1 (1997) 203-224
5. Xiang, Y. and Chu, T.: Parallel learning of belief networks in large and difficult domains. Data Mining and Knowledge Discovery 3 (1999) 315-339
6. Silverstein, C., Brin, S., Motwani, R. and Ullman, J.: Scalable techniques for mining causal structures. Data Mining and Knowledge Discovery 4 (2000) 163-192
7. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A.: Towards conversational human-computer interaction. AI Magazine, 22(4) (2001) 27-38
8. Weizenbaun, J.: ELIZA - A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1) (1965) 36-45

9. Horvitz, E., Breese, J., Heckerman, D., Hovel, D. and Rommelse, K.: The lumiere project: Bayesian user modeling for inferring the goals and needs of software users. Proc. of the 14th Conf. Uncertainty in Artificial Intelligence (1998) 256-265
10. Hong, J.-H. and Cho, S.-B.: A two-stage Bayesian network for effective development of conversational agent. Lecture Note in Computer Science 2690 (2003) 1-8
11. Chai, J., Horvath, V., Nicolov, N., Budzikowska, M., Kambhatla, N. and Zadrozny, W.: Natural language sales assistant: A web-based dialog system for online sales. Proc. of the 13th Annual Conf. on Innovative Applications of Artificial Intelligence (2001) 19-26
12. Paek, T. and Horvitz, E.: Conversation as action under uncertainty. Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence (2000) 455-464
13. Heckerman, D.: A tutorial on learning with Bayesian networks. Microsoft Research, Technical Report MSR-TR-95-06 (1995)
14. Horvitz, E. and Paek, T.: A computational architecture for conversation. Proc. of the 7th Int. Conf. on User Modeling (1999) 201-210
15. Allen, J.: Mixed initiative interaction. IEEE Intelligent Systems, 14(6) (1999) 14-23
16. Wu, X., Zheng, F. and Xu, M.: TOPIC Forest: A plan-based dialog management structure. Int. Conf. on Acoustics, Speech and Signal Processing (2001) 617-620
17. Onisko, A., Druzdzel, M. J. and Wasyluk, H.: Learning Bayesian network parameters from small data sets: Application of noisy-OR gates. Int. Journal of Approximate Reasoning, 27(2) (2001) 165-182
18. Genie2 & Smile, http://www.sis.pitt.edu/~genie/

Stochastic Lotka-Volterra Competitive Systems with Variable Delay

Yi Shen[1], Guoying Zhao[1], Minghui Jiang[1], and Xuerong Mao[2]

[1] Department of Control Science and Engineering, Huazhong University,
of Science and Technology, Wuhan, Hubei, 430074, China
[2] Department of Statistics and Modelling Science, University of Strathclyde,
Glasgow G1 1XH, U.K.

Abstract. In this paper we reveal that the environmental noise will not only suppress a potential population explosion in the stochastic Lotka-Volterra competitive systems with variable delay, but also make the solutions to be stochastically ultimately bounded. To reveal these interesting facts, we stochastically perturb the Lotka-Volterra competitive systems with variable delay $\dot{x}(t) = \mathrm{diag}(x_1(t), \ldots, x_n(t))[b + Ax(t - \delta(t))]$ into the Itô form $\mathrm{d}x(t) = \mathrm{diag}(x_1(t), \ldots, x_n(t))[b + Ax(t - \delta(t))]\mathrm{d}t + \sigma x(t)\mathrm{d}w(t)$, and show that although the solution to the original delay systems may explode to infinity in a finite time, with probability one that of the associated stochastic delay systems do not. We also show that the stochastic systems will be stochastically ultimately bounded without any additional conditions on the matrix A.

1 Introduction

Deterministic subclasses of the Lotka-Volterra competitive systems are well-known and have been extensively investigated in the literature concerning ecological population modelling. One particularly interesting subclass describes the facultative mutualism of two species, where each one enhances the growth of the other, represented through the deterministic systems

$$\begin{aligned} \dot{x}_1(t) &= x_1(t)(b_1 - a_{11}x_1(t) + a_{12}x_2(t)), \\ \dot{x}_2(t) &= x_2(t)(b_2 - a_{22}x_2(t) + a_{21}x_1(t)), \end{aligned} \tag{1}$$

for a_{12} and a_{21} positive constants. The associated dynamics have been developed by many authors [1-9]. Now in order to avoid having a solution that explodes at a finite time, $a_{12}a_{21}$ is required to be smaller than $a_{11}a_{22}$. To illustrate what happens when the latter condition does not hold, suppose that $a_{11} = a_{22} = \alpha$ and $a_{12} = a_{21} = \beta$ (i.e., we have a symmetric system) and $\alpha^2 < \beta^2$. Moreover, let us assume that $b_1 = b_2 = b \geq 1$ and that both species have the same initial value $x_1(0) = x_2(0) = x_0 > 0$. Then the resulting symmetry reduces system (1) to the single deterministic system

$$\dot{x}(t) = x(t)[b + (-\alpha + \beta)x(t)]$$

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 238–247, 2005.

whose solution is given by

$$x(t) = \frac{b}{-(-\alpha+\beta) + \frac{b+(-\alpha+\beta)x_0}{x_0}e^{-bt}}.$$

Now the assumption that $\alpha^2 < \beta^2$ causes $x(t)$ to explode at the finite time $t = \frac{1}{b}\{\ln[b+(-\alpha+\beta)x_0] - \ln[(-\alpha+\beta)x_0]\}$. So are the delay Lotka-Volterra competitive systems [5]. Nevertheless, this can be avoided, even when the condition $a_{12}a_{21} < a_{11}a_{22}$ does not hold, by introducing (stochastic) environmental noise.

Throughout this paper, unless otherwise specified, we let $(\Omega, F, \{F_t\}_{t\geq 0}, P)$ be a complete probability space with a filtration $\{F_t\}_{t\geq 0}$ satisfying the usual conditions (i.e., it is increasing and right continuous while F_0 contains all P-null sets). Moreover, let $w(t)$ be a one-dimensional Brownian motion defined on the filtered space and $R^n_+ = \{x \in R^n : x_i > 0 \text{ for all } 1 \leq i \leq n\}$. Finally, denote the trace norm of a matrix A by $|A| = \sqrt{\text{trace}(A^TA)}$ (where A^T denotes the transpose of a vector or matrix A) and its operator norm by $\|A\| = \sup\{|Ax| : |x| = 1\}$. Let $\tau > 0$ and $C([-\tau, 0]; R^n)$ denote the family of all continuous functions from $[-\tau, 0]$ to R^n_+.

For the generalization of the question, we consider a time-varying Lotka-Volterra competitive systems with n interaction components, which corresponds to the case of facultative mutualism, namely

$$\dot{x}_i(t) = x_i(t)(b_i + \sum_{j=1}^{n} a_{ij}x_j(t-\delta(t))), \quad 1 \leq i \leq n.$$

This system can be rewritten in the matrix form

$$\dot{x}(t) = \text{diag}(x_1(t), \ldots, x_n(t))(b + Ax(t-\delta(t))), \quad t \geq 0, \tag{2}$$

where $x(t) = (x_1(t), \ldots, x_n(t))^T, b = (b_i)_{1\times n}$ and $A = (a_{ij})_{n\times n}$, $\delta : R_+ \to [0, \tau]$ and $\delta'(t) \leq 0$. Stochastically perturbing each parameter

$$a_{ij} \to a_{ij} + \sigma_{ij}\dot{w}(t)$$

results in the new stochastic form

$$\mathrm{d}x(t) = \text{diag}(x_1(t), \ldots, x_n(t))((b + Ax(t-\delta(t)))\mathrm{d}t + \sigma x(t)\mathrm{d}w(t)), \quad t \geq 0, \tag{3}$$

Here $\sigma = (\sigma_{ij})_{n\times n}$, and we impose the condition

(H) $\sigma_{ii} \neq 0, \quad \sigma_{ij}\sigma_{ik} \geq 0, \quad i, j, k = 1, \ldots, n.$

For a stochastic system to have a unique global solution (i.e., no explosion in a finite time) for any given initial value $\{x(t) : -\tau \leq t \leq 0\} \in C([-\tau, 0]; R^n_+)$ the coefficients of system (3) are generally required to satisfy both the linear growth condition and the local Lipschitz condition [10]. However, the coefficients of the system (3)do not satisfy the linear growth condition, though they are locally Lipschitz continuous, so the solution of the system (3) may explode at a finite time. Under the simple hypothesis (H), in this paper we show that this solution is positive and global.

In a population dynamical system, the non-explosion property is often not good enough but the property of ultimate boundedness is more desired. The conditions for the ultimate boundedness are much more complicated than the conditions for the nonexplosion [3-5,7,9] and lots of research are still going on. Naturally, when we study the stochastic Lotka-Volterra competitive system (3) with variable delays we would like to find out under what conditions the solutions will be stochastically ultimately bounded. In the first instance, one may feel that we will need some additional conditions on the matrices A and σ . However, in this paper we shall show that the simple hypothesis (H) on the noise is enough to guarantee the stochastically ultimate boundedness of the solutions of the stochastic Lotka-Volterra competitive system (3) with variable delay .

2 Positive and Global Solutions

Theorem 2.1. Under hypothesis (H), for any system parameters $b \in R^n$ and $A \in R^{n\times n}$, and any given initial data $\{x(t) : -\tau \le t \le 0\} \in C([-\tau,0]; R^n_+)$, there is a unique solution $x(t)$ to the system (3) on $t \ge -\tau$ and the solution will remain in R^n_+ with probability 1, namely $x(t) \in R^n_+$ for all $t \ge -\tau$ almost surely.

Proof. Since the coefficients of the system are locally Lipschitz continuous, for any given initial data $\{x(t) : -\tau \le t \le 0\} \in C([-\tau,0]; R^n_+)$ there is a unique maximal local solution $x(t)$ on $t \in [-\tau, \tau_e)$, where τ_e is the explosion time [10]. To show this solution is global, we need to show that $\tau_e = \infty$ a.s. Let $k_0 > 0$ be sufficiently large for

$$\frac{1}{k_0} < \min_{-\tau\le t\le 0} |x(t)| \le \max_{-\tau\le t\le 0} |x(t)| < k_0.$$

For each integer $k \ge k_0$, define the stopping time

$$\tau_k = \inf\{t \in [0,\tau_e) : x_i(t) \notin (1/k, k) \text{ for some } \quad i = 1,\dots,n\}$$

where throughout this paper we set $\inf \emptyset = \infty$ (as ussual $\emptyset$ denotes the empty set). Clearly, τ_k is increasing as $k \to \infty$. Set $\tau_\infty = \lim_{k\to\infty} \tau_k$, whence $\tau_\infty \le \tau_e$ a.s. If we can show that $\tau_\infty = \infty$ a.s., then $\tau_e = \infty$ a.s. and $x(t) \in R^n_+$ a.s. for all $t \ge 0$. In other words, to complete the proof all we need to show is that $\tau_\infty = \infty$ a.s.. To show this statement, let us define a C^2-function $V : R^n_+ \to R_+$ by

$$V(x) = \sum_{i=1}^{n} (x_i^\theta - \theta \lg(x_i)),$$

where $0 < \theta < 1$.

The nonnegativity of this function can be seen from

$$u^\theta - \theta \lg u > 0 \quad \text{on} \quad u > 0.$$

Let $k \geq k_0$ and $T > 0$ be arbitrary. For $0 \leq t \leq \tau_k \wedge T$, we can apply the Itô formula to $\int_{t-\delta(t)}^t |x(s)|^2 \mathrm{d}s + V(x(t))$ to obtain that

$$\begin{aligned}
&\mathrm{d}[\int_{t-\delta(t)}^t |x(s)|^2 \mathrm{d}s + V(x(t))] \\
&\leq [|x(t)|^2 - |x(t-\delta(t))|^2]\mathrm{d}t + \sum_{i=1}^n \{\theta(x_i^{\theta-1}(t) - x_i^{-1}(t))x_i(t) \\
&\quad \times[(b_i + \sum_{j=1}^n a_{ij}x_j(t-\delta(t)))\mathrm{d}t + \sum_{j=1}^n \sigma_{ij}x_j(t)\mathrm{d}w(t)] \\
&\quad +0.5(\theta(\theta-1)x_i^{\theta-2}(t) + \theta x_i^{-2}(t))x_i^2(t)(\sum_{j=1}^n \sigma_{ij}x_j(t))^2\mathrm{d}t\} \\
&= \{|x(t)|^2 - |x(t-\delta(t))|^2 + \sum_{i=1}^n \theta(x_i^\theta(t) - 1) \\
&\quad \times(b_i + \sum_{j=1}^n a_{ij}x_j(t-\delta(t))) + 0.5\sum_{i=1}^n(\theta + \theta(\theta-1)x_i^\theta(t))(\sum_{j=1}^n \sigma_{ij}x_j(t))^2\}\mathrm{d}t \\
&\quad +\sum_{i=1}^n\sum_{j=1}^n \theta(x_i^\theta(t) - 1)\sigma_{ij}x_j(t)\mathrm{d}w(t). \qquad (4)
\end{aligned}$$

Compute

$$\begin{aligned}
&\sum_{i=1}^n \theta(x_i^\theta(t) - 1)(b_i + \sum_{j=1}^n a_{ij}x_j(t-\delta(t))) \\
&\leq \sum_{i=1}^n \theta b_i(x_i^\theta(t) - 1) + \sum_{i=1}^n\sum_{j=1}^n[0.25n\theta^2 a_{ij}^2(x_i^\theta(t) - 1)^2 \\
&\quad +n^{-1}x_j^2(t-\delta(t))] \\
&= \sum_{i=1}^n \theta b_i(x_i^\theta(t) - 1) + \sum_{i=1}^n\sum_{j=1}^n 0.25n\theta^2 a_{ij}^2(x_i^\theta(t) - 1)^2 \\
&\quad +|x(t-\delta(t))|^2
\end{aligned}$$

and

$$\sum_{i=1}^n(\sum_{j=1}^n \sigma_{ij}x_j(t))^2 \leq \sum_{i=1}^n\sum_{j=1}^n \sigma_{ij}^2 \sum_{j=1}^n x_j^2(t) = |\sigma|^2|x(t)|^2.$$

Moreover, by hypothesis (H),

$$\sum_{i=1}^n x_i^\theta(t)(\sum_{j=1}^n \sigma_{ij}x_j(t))^2 \geq \sum_{i=1}^n \sigma_{ii}^2 x_i^{2+\theta}(t).$$

Substituting these into (4) yields

$$\begin{aligned} &\mathrm{d}[\int_{t-\delta(t)}^{t} |x(s)|^2 \mathrm{d}s + V(x(t))] \\ &\leq F(x(t))\mathrm{d}t + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta(x_i^{\theta}(t) - 1)\sigma_{ij}x_j(t)\mathrm{d}w(t), \end{aligned} \tag{5}$$

where

$$\begin{aligned} F(x) = &(1 + 0.5\theta|\sigma|^2)|x|^2 + \sum_{i=1}^{n} \theta b_i(x_i^{\theta} - 1) + 0.25n\theta^2 \\ &\times \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2(x_i^{\theta} - 1)^2 - 0.5\theta(1-\theta)\sum_{i=1}^{n}\sigma_{ii}^2 x_i^{2+\theta} \end{aligned} \tag{6}$$

It is straightforward to see that $F(x)$ is bounded, say by K, in R_+^n. We therefore obtain that

$$\mathrm{d}[\int_{t-\delta(t)}^{t} |x(s)|^2 \mathrm{d}s + V(x(t))] \leq K\mathrm{d}t + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta(x_i^{\theta}(t) - 1)\sigma_{ij}x_j(t)\mathrm{d}w(t).$$

Integrating both sides from 0 to $\tau_k \wedge T$, and then taking expectations, yields

$$E\{\int_{\tau_k\wedge T - \delta(\tau_k \wedge T)}^{\tau_k \wedge T} |x(s)|^2 \mathrm{d}s + V(x(\tau_k \wedge T))\} \leq \int_{-\tau}^{0} |x(s)|^2 \mathrm{d}s + V(x(0)) + KE(\tau_k \wedge T).$$

Consequently,

$$EV(x(\tau_k \wedge T)) \leq \int_{-\tau}^{0} |x(s)|^2 \mathrm{d}s + V(x(0)) + KT. \tag{7}$$

Note that for every $\omega \in \{\tau_k \leq T\}$, there is some i such that $x_i(\tau_k, \omega)$ equals either k or $1/k$, and hence $V(x(\tau_k, \omega))$ is no less than either

$$k^{\theta} - \theta \lg(k)$$

or

$$k^{-\theta} - \theta \lg(k^{-1}) = k^{-\theta} + \theta \lg(k).$$

Consequently,

$$V(x(\tau_k, \omega)) \geq (k^{\theta} - \theta \lg(k)) \wedge (k^{-\theta} + \theta \lg(k)).$$

It then follows from (7) that

$$\begin{aligned} &\int_{-\tau}^{0} |x(s)|^2 \mathrm{d}s + V(x(0)) + KT \geq E[1_{\{\tau_k \leq T\}}(\omega)V(x(\tau_k, \omega))] \\ &\geq P\{\tau_k \leq T\}[(k^{\theta} - \theta \lg(k)) \wedge (k^{-\theta} + \theta \lg(k))], \end{aligned}$$

where $1_{\{\tau_k \leq T\}}$ is the indicator function of $\{\tau_k \leq T\}$. Letting $k \to \infty$ gives $\lim_{k\to\infty} P\{\tau_k \leq T\} = 0$ and hence $P\{\tau_\infty \leq T\} = 0$. Since $T > 0$ is arbitrary, we must have $P\{\tau_\infty < \infty\} = 0$, so $P\{\tau_\infty = \infty\} = 1$ as required.

Remark 1. It is well known that the system (1) and (2) may explode to infinity at a finite time for some system parameters $b \in R^n$ and $A \in R^{n\times n}$. However, the explosion will no longer happen as long as there is a noise. In other words, this result reveals the important property that the environmental noise suppresses the explosion for the time-varying delay system.

3 Stochastically Ultimate Boundedness

Now let us give the definition of stochastically ultimate boundedness.

Definition 3.1. System (3) is said to be stochastically ultimately bounded if for any $\varepsilon \in (0, 1)$, there is a positive constant $\gamma = \gamma(\varepsilon)$ such that for any initial data $\{x(t) : -\tau \leq t \leq 0\} \in C([-\tau, 0]; R^n_+)$, the solution $x(t)$ of system (3) has the property that

$$\limsup_{T\to\infty} P\{|x(t)| \leq \gamma\} \geq 1 - \varepsilon. \tag{8}$$

Theorem 3.2. Under hypothesis (H), system (3) is stochastically ultimately bounded.

To prove Theorem 3.2, we present a useful lemma from which the stochastically ultimate boundedness will follow directly.

Lemma 3.3. Let hypothesis (H) hold and $\theta \in (0, 1)$. Then there is a positive constant $K = K(\theta)$, which is independent of the initial data $\{x(t) : -\tau \leq t \leq 0\} \in C([-\tau, 0]; R^n_+)$, such that the solution $x(t)$ of system (3) has the property that

$$\limsup_{t\to\infty} E|x(t)|^\theta \leq K. \tag{9}$$

Proof. Define

$$V(x) = \sum_{i=1}^n x_i^\theta \quad \text{for} \quad x \in R^n_+.$$

By the Itô formula, we have

$$\mathrm{d}V(x(t)) = LV(x(t), x(t-\delta(t)))\mathrm{d}t + (\sum_{i=1}^n \theta x_i^\theta(t) \sum_{j=1}^n \sigma_{ij} x_j(t))\mathrm{d}w(t), \tag{10}$$

where $LV : R^n_+ \times R^n_+ \to R$ is defined by

$$LV(x, y) = \sum_{i=1}^n \theta x_i^\theta (b_i + \sum_{j=1}^n a_{ij} y_j) - \frac{\theta(1-\theta)}{2} \sum_{i=1}^n x_i^\theta (\sum_{j=1}^n \sigma_{ij} x_j)^2.$$

Compute

$$\begin{aligned} LV(x,y) &\le \sum_{i=1}^n \theta b_i x_i^\theta + \sum_{i=1}^n \sum_{j=1}^n (\frac{n}{4}\theta^2 a_{ij}^2 x_i^{2\theta} + \frac{1}{n} y_j^2) - \frac{\theta(1-\theta)}{2} \sum_{i=1}^n \sigma_{ii}^2 x_i^{2+\theta} \\ &= \sum_{i=1}^n \theta b_i x_i^\theta + \frac{n}{4}\theta^2 \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 x_i^{2\theta} - \frac{\theta(1-\theta)}{2} \sum_{i=1}^n \sigma_{ii}^2 x_i^{2+\theta} + |y|^2 \\ &= F(x) - V(x) - e^\tau |x|^2 + |y|^2, \end{aligned} \tag{11}$$

where

$$\begin{aligned} F(x) = e^\tau |x|^2 + \sum_{i=1}^n (\theta b_i + 1) x_i^\theta + \frac{n}{4}\theta^2 \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 x_i^{2\theta} \\ - \frac{\theta(1-\theta)}{2} \sum_{i=1}^n \sigma_{ii}^2 x_i^{2+\theta}. \end{aligned} \tag{12}$$

Note that $F(x)$ is bounded in R_+^n, namely

$$K_1 := \sup_{x \in R_+^n} F(x) < \infty.$$

We therefore have

$$LV(x,y) \le K_1 - V(x) - e^\tau |x|^2 + |y|^2, \tag{13}$$

Substituting this into (10) gives

$$\begin{aligned} \mathrm{d}V(x(t)) \le (K_1 - V(x(t)) - e^\tau |x(t)|^2 + |x(t-\delta(t))|^2)\mathrm{d}t \\ + (\sum_{i=1}^n \theta x_i^\theta(t) \sum_{j=1}^n \sigma_{ij} x_j(t)) \mathrm{d}w(t). \end{aligned}$$

Once again by the Itô formula we have

$$\begin{aligned} \mathrm{d}[e^t V(x(t))] &= e^t (V(x(t))\mathrm{d}t + \mathrm{d}V(x(t))) \\ &\le e^t (K_1 - e^\tau |x(t)|^2 + |x(t-\delta(t))|^2)\mathrm{d}t \\ &\quad + e^t (\sum_{i=1}^n \theta x_i^\theta(t) \sum_{j=1}^n \sigma_{ij} x_j(t)) \mathrm{d}w(t). \end{aligned}$$

We hence derive that

$$\begin{aligned} e^t EV(x(t)) &\le V(x(0)) + K_1(e^t - 1) - E\int_0^t e^{s+\tau} |x(s)|^2 \mathrm{d}s \\ &\quad + E\int_0^t e^s |x(s-\delta(s))|^2 \mathrm{d}s \\ &\le V(x(0)) + K_1(e^t - 1) - E\int_0^t e^{s+\tau} |x(s)|^2 \mathrm{d}s \end{aligned}$$

$$+E\int_{-\tau}^{t} e^{s+\tau}|x(s)|^2\mathrm{d}s$$
$$= V(x(0)) + K_1(e^t - 1) + \int_{-\tau}^{0} |x(s)|^2\mathrm{d}s$$

This implies immediately that

$$\limsup_{t\to\infty} EV(x(t)) \le K_1.$$

On the other hand, we have

$$|x|^2 \le n \max_{1\le i\le n} x_i^2$$

so

$$|x|^\theta \le n^{\theta/2} \max_{1\le i\le n} x_i^\theta \le n^{\theta/2} V(x).$$

We therefore finally have

$$\limsup_{t\to\infty} E|x(t)|^\theta \le n^{\frac{\theta}{2}} K_1.$$

and the assertion (9) follows by setting $K = n^{\theta/2}K_1$.

The proof of Theorem 3.2.

Proof. By Lemma 3.3, there is $K > 0$ such that

$$\limsup_{t\to\infty} E(\sqrt{|x(t)|}) \le K.$$

Now, for any $\varepsilon > 0$, let $\gamma > K^2/\varepsilon^2$. Then by Chebyshev's inequality,

$$P\{|x(t)| > \gamma\} \le \frac{E(\sqrt{|x(t)|})}{\sqrt{\gamma}}.$$

Hence

$$\limsup_{t\to\infty} P\{|x(t)| > \gamma\} \le \frac{K}{\sqrt{\gamma}} < \varepsilon.$$

This implies

$$\limsup_{t\to\infty} P\{|x(t)| \le \gamma\} \ge 1 - \varepsilon.$$

as required.

4 Moment Average in Time

The result in the previous section shows that the solutions of system (3) will be stochastically ultimately bounded. That is, the solutions will be ultimately bounded with large probability. The following result shows that the average in time of the second moment of the solutions will be bounded.

Theorem 4.1. Under hypothesis (H), there is a positive constant K, which is independent of the initial data $\{x(t) : -\tau \le t \le 0\} \in C([-\tau, 0]; R^n_+)$, such that the solution $x(t)$ of system (3) has the property that

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T E|x(t)|^2 \mathrm{d}t \le K. \tag{14}$$

Proof. We use the same notations as in the proof of Theorem 3.2. Write (12) as

$$F(x) = F_1(x) - 2|x|^2 + V(x) + e^\tau |x|^2 \tag{15}$$

with

$$F_1(x) = 2|x|^2 + \sum_{i=1}^n \theta b_i x_i^\theta + \frac{n}{4}\theta^2 \sum_{i=1}^n\sum_{j=1}^n a_{ij}^2 x_i^{2\theta} - \frac{\theta(1-\theta)}{2}\sum_{i=1}^n \sigma_{ii}^2 x_i^{2+\theta}.$$

Clearly, F_1 is bounded in R^n_+, namely $K = \max_{x\in R^n_+} F_1(x) < \infty$. So

$$F(x) \le K - 2|x|^2 + V(x) + e^\tau |x|^2. \tag{16}$$

By (10), (11) and (16), we have

$$\begin{aligned}\mathrm{d}V(x(t)) \le{}& (K - 2|x(t)|^2 + |x(t-\delta(t))|^2)\mathrm{d}t \\ &+(\sum_{i=1}^n \theta x_i^\theta(t)\sum_{j=1}^n \sigma_{ij}x_j(t))\mathrm{d}w(t).\end{aligned} \tag{17}$$

By (17) and $\delta'(t) \le 0$, we have

$$\begin{aligned}&\mathrm{d}[\int_{t-\delta(t)}^t |x(s)|^2\mathrm{d}s + V(x(t))] \\ &\le (K - |x(t)|^2)\mathrm{d}t + (\sum_{i=1}^n \theta x_i^\theta(t)\sum_{j=1}^n \sigma_{ij}x_j(t))\mathrm{d}w(t).\end{aligned} \tag{18}$$

Using this estimation, integrating both sides of (18) from 0 to any T, and then taking expectations, we obtain that

$$0 \le \int_{-\tau}^0 |x(s)|^2\mathrm{d}s + V(x(0)) + KT - E\int_0^T |x(t)|^2\mathrm{d}t,$$

namely,

$$E\int_0^T |x(t)|^2\mathrm{d}t \le \int_{-\tau}^0 |x(s)|^2\mathrm{d}s + V(x(0)) + KT.$$

Dividing both sides by T and then letting $T \to \infty$ we get

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T E|x(t)|^2\mathrm{d}t \le K.$$

as required.

5 Conclusion

In this paper we investigate that the environmental noise play a key role in suppressing a potential population explosion on the stochastic delay-varying Lotka-Volterra competitive systems with variable delay . And we show that although the solution to the original delay systems may explode to infinity in a finite time, with probability one that of the associated stochastic time-varying delay systems do not.

Acknowledgments

The work was supported by Natural Science Foundation of Hubei (2004ABA055) and National Natural Science Foundation of China (60274007, 60074008).

References

1. Ahmad, A., Rao, M. R. M.: Asymptotically Periodic Solutions of N-competing Species Problem with Time Delay. J. Math. Anal. Appl., **186** (1994) 557-571
2. Bereketoglu, H., Gyori, I.: Global Asymptotic Stability in A Nonautonomous Lotka-Volterra Type System with Infinite Delay. J. Math. Anal. Appl., **210** (1997)279-291
3. Freedman, H. I., Ruan, S.: Uniform Persistence in Functional Differential Equations. J. Differential Equations, **115** (1995) 173-192
4. Gopalsamy, K.: Stability and Oscillations in Delay Differential Equations of Population Dynamics. Kluwer Academic, Dordrecht, (1992)
5. He, X., Gopalsamy, K.: Persistence, Attractivity, and Delay in Facultative Mutualism. J. Math. Anal. Appl., **215** (1997) 154-173
6. Kolmanovskii, V., Myshkis, A.: Applied Theory of Functional Differential Equations. Kluwer Academic, (1992)
7. Kuang, Y.: Delay Differential Equations with Applications in Population Dynamics. Academic Press, Boston, (1993)
8. Kuang, Y., Smith, H. L.: Global Stability for Infinite Delay Lotka-Volterra Type Systems. J. Differential Equations, **103** (1993) 221-246
9. Teng, Z., Yu, Y.: Some New Results of Nonautonomous Lotka-Volterra Competitive Systems with Delays. J. Math. Anal. Appl., **241** (2000) 254-275
10. Mao, X.: Exponential Stability of Stochastic Differential Equations. Dekker, New York, (1994)

Human Face Recognition Using Modified Hausdorff ARTMAP

Arit Thammano and Songpol Ruensuk

Faculty of Information Technology,
King Mongkut's Institute of Technology Ladkrabang,
Bangkok, 10520 Thailand
arit@it.kmitl.ac.th, r.songpol@gmail.com

Abstract. This paper proposes a new neural network approach specifically designed for solving two dimensional binary image recognition problems. The proposed neural network is an extension of the Hausdorff ARTMAP introduced by Thammano and Rungruang [1]. The objectives of this research are to improve the accuracy and correct the drawbacks of the original network. The performance of this proposed model has been compared with that of the original Hausdorff ARTMAP. The experimental results on two benchmark databases, the ORL and Yale face databases, show that the proposed network surpasses the original Hausdorff ARTMAP in both performance and processing time.

1 Introduction

Person identification has received increasing attention in recent years. In general, there are three ways to identify an individual: the person knows something (e.g., a PIN, a password); the person possesses something (e.g., an ID card, a passport); or by measuring something about the person's body [2]. The later encompasses the biometric identification. Among all of the biometric identification methods, face recognition is the most natural, non-intrusive, and user-friendly biometric measure because it requires no disturbance to the person being identified. While more intrusive biometric recognition systems (e.g., palm, fingerprint) are presently more accurate, face recognition still has a critical role in certain domains since the person being identified may be at a distance from the sensor, the person does not have to be compliant, and recognition can be performed continuously.

A variety of techniques have been applied to deal with the face recognition problems. The reader should pay attention to Chellappa et al. [3] and the references therein for a more complete survey of previous research works on face recognition. In the early years, many researchers used the structure parameters of faces as the features in the facial image recognition. Kelly [4] used various kinds of facial features, including width of the head and distances between eyes, top of head to eyes, between eyes and nose and the distance from eyes to mouth. During the past decade, researchers have paid much attention to the statistical approaches -- such as eigenfaces [5], KL transform [6], and SVD [7] -- and the neural network approaches [8, 9]. Neural network is very suitable for face recognition systems. It has the ability to automatically learn the rules from the given collection of representative examples,

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 248 – 256, 2005.

instead of following a set of human-designed rules [10]. Moreover, it is well-known that the neural network is more robust to noise than other methods. Thammano and Rungruang [1] proposed the Hausdorff ARTMAP neural network, which employs the concept of the Hausdorff distance to measure the likeness or similarity between the incoming input pattern and the reference patterns of each subject. The results show that the Hausdorff ARTMAP is very effective in dealing with the face recognition problems. It outperforms many different techniques studied in the past. The research described in this paper concerns a modification of the Hausdorff ARTMAP neural network in order to further improve the accuracy and correct its drawbacks. The ORL and Yale face databases are used in this study to evaluate the performance of the proposed neural network.

Following this introduction, section 2 briefly describes the concept of the Hausdorff distance. The original Hausdorff ARTMAP is introduced in section 3. Section 4 presents the proposed model. In section 5, the experimental results are demonstrated and discussed.

2 Hausdorff Distance

The Hausdorff distance, when used as a measure of similarity between two two-dimensional binary patterns, has shown to agree closely with human performance [11]. The Hausdorff distance measures the extent to which each point of an input pattern lies near some point of a reference pattern. Given two finite sets A = $\{a_1, a_2, \ldots, a_p\}$ and B = $\{b_1, b_2, \ldots, b_q\}$, the Hausdorff distance between sets A and B is defined as:

$$H(A,B)=\max\{h(A,B),h(B,A)\}. \tag{1}$$

where the function h(A, B) is called the directed Hausdorff distance from set A to set B, which can be computed as follows:

$$h(A,B)=\max_{a\in A}\{\min_{b\in B}(\|a-b\|)\}. \tag{2}$$

where ||a - b|| is the Euclidean distance between point a and point b. The Hausdorff distance exhibits many desirable properties for pattern recognition. However, some modifications of the directed Hausdorff distance are made in this study in order to increase the noise immunity of the measurement.

$$h(A,B)=\frac{\sum_{a\in A} h(a,B)}{|A|}. \tag{3}$$

where |A| is the number of points in set A. h(a, B) is the pointwise Hausdorff distance for point a. The pointwise Hausdorff distance is computed as follows:

$$h(a,B)=\min_{b\in B}(\|a-b\|). \tag{4}$$

3 Hausdorff ARTMAP

The architecture of the Hausdorff ARTMAP is a three-layer neural network as shown in Figure 1. The first layer is the input layer, which consists of X×Y nodes. Each node represents a pixel in the input pattern. The second layer is the cluster layer. The nodes in this second layer are constructed during the training phase. The third layer is the output layer. Each node in the output layer represents a class that the Hausdorff ARTMAP has to learn to recognize. During the supervised learning, the binary input pattern I^m is presented to the model, together with its respective target output vector. The input pattern is denoted by

$$I^m_{x,y} = \{1,0\} \quad : x = 1,2,\ldots,X; y = 1,2,\ldots,Y\,. \tag{5}$$

where m is the m^th input pattern. X and Y are the dimensions of the input pattern.

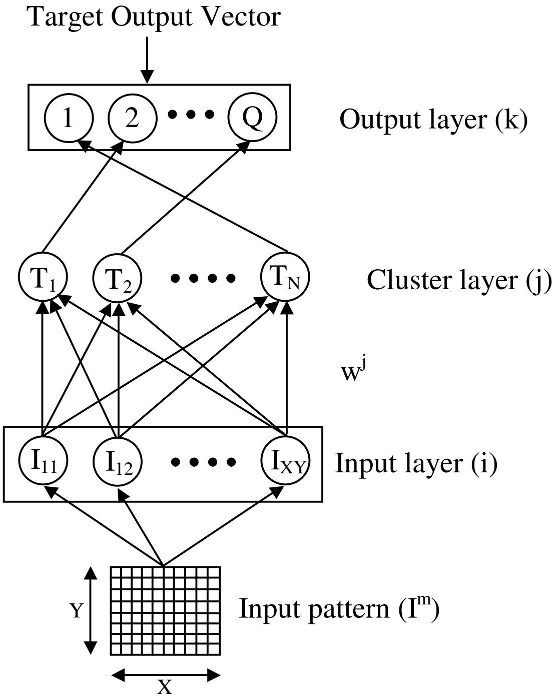


Fig. 1. Architecture of the Hausdorff ARTMAP

Each node in the cluster layer is fully connected to the nodes in the input layer via the connections w^j. The weight w^j, which has the same dimension as the input pattern, represents the reference pattern of the j^th node in the cluster layer. Once the input pattern is transmitted to the cluster layer, the choice function of each j^th node in the cluster layer is evaluated as follows:

$$T_j(I^m)=H(I^m,w^j)\ . \tag{6}$$

where $H(I^m, w^j)$ is the Hausdorff distance between the input pattern I^m and the reference pattern of the j^{th} node (w^j). The system then makes a cluster choice by selecting the winning node J with minimum choice function value, among all the nodes j in the cluster layer. The cluster choice is indexed by J, where

$$T_J(I^m)=\min\{T_j(I^m)\}\quad :j=1,2,\ldots,N\,. \tag{7}$$

where N is the number of nodes in the cluster layer. In case of a tie, the node with the smallest index is chosen. Next, the vigilance criterion is evaluated to check whether the degree of mismatch between the input pattern and the reference pattern of the chosen cluster is within an acceptable level.

$$T_J(I^m)\le\rho\,. \tag{8}$$

where ρ is the vigilance parameter, which has the value between 0 and the length of the diagonal line. Resonance will occur if the chosen cluster meets the above criterion. However, if the condition in (8) is not satisfied, a new cluster node J is recruited to code the input pattern. The weight of this new node is initialized to be equal to the input pattern and this new node will automatically satisfy (8).

$$w^J=I^m\,. \tag{9}$$

Next, the system associates the winning node J in the cluster layer with the target output vector. If the winning node J does not belong to the correct class defined by the target output vector, a new cluster node J is recruited and its weight is initialized using equation (9). Then the connection between a new cluster node and the target output is created. However, if the winning node represents the class to which I^m belongs, the weight vector w^J is then updated according to

$$b^J_{x,y}=\sum_{u=-S}^{S}\ \sum_{v=-S}^{S} I^m_{x+u,\,y+v}\ . \tag{10}$$

$$w^{J^{new}}_{x,y}=\begin{cases} w^{J^{old}}_{x,y}, & \text{if } b^J_{x,y}>0\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

where S is a positive integer value; it tells the model how many pixels surrounding the current x, y position should be considered during the weight adjustment.

During testing, each test pattern is applied in turn and its class is predicted. The class whose cluster node returns the minimum output value is the result of the prediction.

Results of the experiment conducted by Thammano and Rungruang [1] show that the Hausdorff ARTMAP is very effective in dealing with the face recognition problems. It outperforms many different techniques studied in the past. However, the Hausdorff ARTMAP suffers from the following problems. First, the performance of the Hausdorff ARTMAP depends directly on the order in which the training images are examined. Second, the time used for recognizing the subject is long due to its large reference pattern size. The above drawbacks motivate the development of the proposed model presented in the next section.

4 Modified Hausdorff ARTMAP

The architecture of the modified Hausdorff ARTMAP is exactly the same as the Hausdorff ARTMAP. The modifications are done to the training algorithm. The details of the training algorithm are presented as follows:

1. The closest similarity between the input pattern m and other input patterns within the same output class K is located.

$$d_m = \min_{\forall n, n \in K, n \neq m} \left[\min_{o=1}^{c} \left[H(I^n, w^m) \right] \right]. \tag{12}$$

   where I^n represents the input pattern n which belongs to the same class as the input pattern m. w^m is the small area on the input pattern m; it is used to represent the whole I^m image. c is the number of locations on the input pattern n, which w^m is compared to.

   Next, d_m is compared to the similarity threshold ($\rho_{similar}$). $\rho_{similar}$ is a predetermined value between 0 and the length of the diagonal line. If d_m is less than or equal to $\rho_{similar}$, the process will continue to the next step. However, if d_m exceeds $\rho_{similar}$, w^m size will be increased by predefined pixels and this step will be repeated.

2. After w^m is identified, its similarities with other input patterns outside the output class K are determined. The minimum of the above similarities is compared to the dissimilarity threshold ($\rho_{dissimilar}$) based on equation 14.

$$do_m = \min_{\forall p, p \notin K} \left[\min_{o=1}^{c} \left[H(I^p, w^m) \right] \right]. \tag{13}$$

$$do_m > \rho_{dissimilar} \, . \tag{14}$$

   where I^p represents the input pattern p which does not belong to the same class as the input pattern m. $\rho_{dissimilar}$ is a predetermined value between 0 and the length of the diagonal line. However, it must be greater than or equal to $\rho_{similar}$. If the condition in (14) is satisfied, the process will continue to the next step. If not, w^m size will be increased by predefined pixels and then go back to step 1.

3. This third step determines the capability of w^m in representing the input patterns within the same output class.

$$R_{mn} = \begin{cases} 1, & \text{if } d_{mn} \leq \rho_{similar} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$$d_{mn} = \min_{o=1}^{c} \left[H(I^n, w^m) \right]. \tag{16}$$

where m = 1, 2, 3, …, e and n = 1, 2, 3, …, e. e is the number of input patterns in class K. "R_{mn} = 1" means that w^m is capable of representing the input pattern n. On the contrary, "R_{mn} = 0" means that w^m is incapable of representing the input pattern n.

4. In this step, the nodes in the cluster layer are created and a select group of w^m is used to be their reference weights. w^m which is capable of representing the maximum number of input patterns in the class is the first to be chosen. The next most capable w^m are subsequently picked until all input patterns in the class are represented. In case of a tie, the averages of the similarities between each w^m in question and the rest of the input patterns in the same class are calculated; the one with the smallest average is selected.

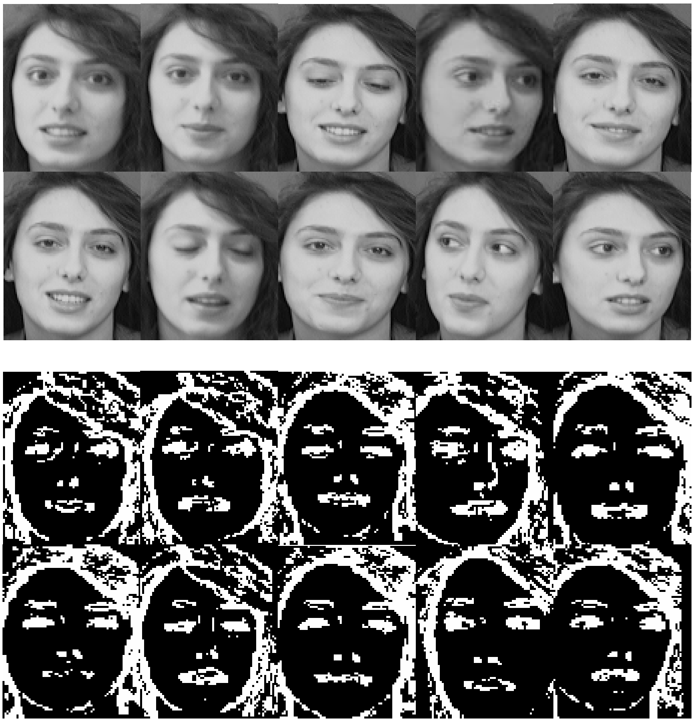
5 Experimental Results

To test the performance of the modified Hausdorff ARTMAP for face recognition, the experiments have been conducted on 2 databases: the ORL face database [12][13] and the Yale face database [14][15]. The results of the experiments are then compared to those of the original Hausdorff ARTMAP. In order to be comparable, the preprocessing step of this study replicates that of Thammano and Rungruang's study. First, the gray-level edge image E(x, y) is obtained by applying morphological operations [16] on the original face image f(x, y). Then, the gray-level edge image is converted to a binary edge image using the adaptive threshold method.

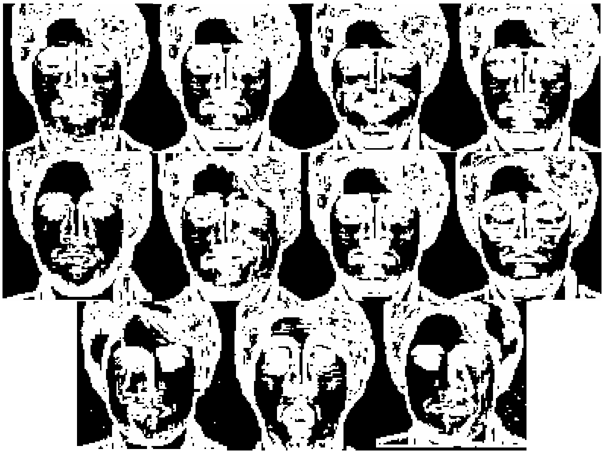
$$n(x,y)=\frac{E(x,y)}{f(x,y)}\,. \tag{17}$$

The values of the function n(x, y) are then sorted in descending order, and the threshold is set so that 30% of the points with the largest magnitudes in n(x, y) are selected.

For the ORL database, there are 10 different images of each of 40 distinct subjects; therefore, the total of 400 different face images are used in this experiment. The images of each subject were taken at different times with different lighting, facial expressions (open/closed eyes, smiling/non-smiling), and facial details (glasses/no-glasses) as shown in Figure 2(a). Four images of each subject are randomly chosen for training and the remaining six images are used for testing. Previously, Lin et al. [17] carried out a comparison study of many different techniques – the principal component analysis (PCA), the conventional Hausdorff distance (HD), the doubly modified Hausdorff distance (M2HD), the spatially weighted Hausdorff distance (SWHD), the spatially weighted doubly Hausdorff distance (SW2HD), the spatially eigen-weighted Hausdorff distance (SEWHD), and the spatially eigen-weighted doubly Hausdorff distance (SEW2HD) – on this ORL face database. The reported recognition rates varied from 46 – 91%. The best recognition rate (91%) was achieved from the SEW2HD. Table 1 shows the recognition results of both the Hausdorff ARTMAP and the modified Hausdorff ARTMAP on the ORL database. The best performance of the Hausdorff ARTMAP is obtained when the vigilance parameter (ρ)



(a)



(b)

Fig. 2. Examples of the original face images and the binary edge images of the ORL (a) and Yale (b) face databases

is set at 0.8 and S is 2. However, the performance of the model might be lower if the sequence of the training images is changed. For the modified Hausdorff ARTMAP, the best recognition rate of 95.83% is achieved when $\rho_{similar}$ and $\rho_{dissimilar}$ are 0.35 and 0.9 respectively.

The Yale database (Figure 2(b)) contains 165 different face images of 15 distinct subjects. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. In this experiment, randomly picked 90 images (6 images per subject) are included in the training set. The remaining 75 images (5 images per subject) are included in the testing set. The recognition results of the Hausdorff ARTMAP vary from 89.33 – 96.00% depending on the order in which the training images are examined. On the other hand, the recognition results of the modified Hausdorff ARTMAP are always 96.00%, whatever the order of the training images is.

Table 1. Experimental results on ORL database

	Number of Cluster Nodes	Size of Reference Patterns	% Correct
Hausdorff ARTMAP	152-153	100×100	94.58-95.42
Modified Hausdorff ARTMAP	159	85×85, 59×59	95.83

Table 2. Experimental results on Yale database

	Number of Cluster Nodes	Size of Reference Patterns	% Correct
Hausdorff ARTMAP	83-88	100×100	89.33-96
Modified Hausdorff ARTMAP	63	93×93, 61×61	96

References

1. Thammano, A., Rungruang, C.: Hausdorff ARTMAP for Human Face Recognition. WSEAS Transactions on Computers, Issue 3, Vol. 3. (2004) 667-672
2. Martínez, A. M., Yang, M., Kriegman, D. J.: Special Issue on Face Recognition. Computer Vision and Image Understanding, Vol. 91, No. 1-2. Academic Press (2003) 1-5
3. Chellappa, R., Wilson, C. L., Sirohey, S.: Human and Machine Recognition of Faces: A Survey. Proceedings of the IEEE, Vol. 83, No. 5. (1995) 705-740
4. Kelly, M. D.: Visual Identification of People by Computer. Technical Report AI-130. Stanford AI Project, Stanford CA (1970)
5. Turk, M. A., Pentland, A. P.: Face Recognition Using Eigenfaces. Proceedings of the International Conference on Pattern Recognition. (1991) 586-591
6. Akamatsu, S., Sasaki, T., Fukamachi, H., Suenaga, Y.: A Robust Face Identification Scheme – KL Expansion of an Invariant Feature Space. SPIE Proc.: Intelligent Robots and Computer Vision X: Algorithms and Techniques, Vol. 1607. (1991) 71-84
7. Cheng, Y., Liu, K., Yang, J., Wang, H.: A Robust Algebraic Method for Human Face Recognition. Proceedings of 11th International Conference on Pattern Recognition. (1992) 221-224
8. Kung, S. Y., Lin, S. H., Fang, M.: A Neural Network Approach to Face/Palm Recognition. Proceedings of the IEEE Workshop on Neural Networks for Signal Processing. (1995) 323-332
9. El-Bakry, H. M., Abo Elsoud, M. A.: Human Face Recognition Using Neural Networks. Proceedings of 16th National Radio Science Conference (NRSC'99). (1999)
10. Lin, S., Kung, S., Lin, L.: Face Recognition/Detection by Probabilistic Decision-Based Neural Network. IEEE Transactions on Neural Networks, Vol. 8, No. 1. (1997) 114-132
11. Rosandich, R. G.: HAVNET: A New Neural Network Architecture for Pattern Recognition. Neural Networks, Vol. 10, No. 1. Pergamon Press (1997) 139-151
12. Samaria, F., Harter, A.: Parameterisation of a Stochastic Model for Human Face Identification. Proceedings of 2nd IEEE Workshop on Applications of Computer Vision (1994)
13. ORL Face Database: Retrieved from http://www.uk.research.att.com/facedatabase.html
14. Bellhumer, P. N., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, Special Issue on Face Recognition, Vol. 17, No. 7. (1997) 711-720
15. Yale Face Database: Retrieved from http://cvc.yale.edu/projects/yalefaces/yalefaces.html
16. Gonzalez, R. C., Woods, R. E.: Digital Image Processing. Prentice-Hall, Englewood Cliffs NJ (2002)
17. Lin, K., Lam, K., Siu, W.: Spatially Eigen-weighted Hausdorff Distances for Human Face Recognition. Pattern Recognition, Vol. 36, No. 8. Pergamon Press (2003) 1827-1834

A Dynamic Decision Approach for Long-Term Vendor Selection Based on AHP and BSC

Ziping Chiang

Department of Logistics Management, Leader University,
709 Tainan, Taiwan, R.O.C.
ziping@mail.leader.edu.tw

Abstract. For solving the dynamic condition problem of vendor selection, the analytic hierarchy process method is modified to a dynamic approach in the period of analytic cycle. The balanced scorecard is used to define the 4 major frameworks of supplier selection including customers, financial, internal business processes, and innovation and learning. The 16 attributes are extended from major frameworks. The main character of proposed method is the scores of attributes and alternatives from the estimation of commander's trade-off can be changed in time axis under the changeable and conjecturable business environments. In case study, the advantage and limit of the model are illustrated.

1 Introduction

The vendor selection problems are proposed in many literatures [1],[2],[3],[4]. In one of the famous literatures, Dickson (1966) issued 23 criteria for measuring the most suitable supplier [1]. In some recently papers, the analytic structures of vendor selection including methodologies, criteria and weights are studied [4]-[11]. But few literatures propose how decision maker measure the analytic structure in different time sections. In the trend of global supply chain management, enterprises try to select a suitable vendor and cooperate over a long period of time. The attributes and weights may change in time axis under the changeable business environments. Traditional multiple attributes decision making methods are hard to solve the long-term performance measurement problems. Thus, this paper proposed a dynamic decision approach for long-term vendor selection problem. Section 2 describes the proposed model including notations, problem statement, the advantages of integrating analytic hierarchy process (AHP) and balanced scorecard (BSC), hypotheses, proposed dynamic method, and attributes. A case study is used to illustrate the calculating steps and the advantage and limit of the model in section 3. Finally, conclusion and future work are in section 4.

2 Proposed Model

Before describing the model, the following notations are defined:

A_j possible alternatives (non-dominated solutions) which decision makers have to choose, j=1, 2, …, m

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 257 – 265, 2005.

x_{ijk} the estimative scores of *i*th and *j*th pare-wise comparison of *k*th time, k=1, 2, … , p
k analytic time sections
n number of criteria
m number of possible alternatives
p number of analytic time sections.

2.1 Problem Statement

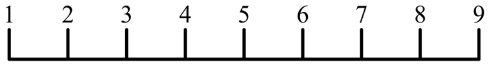
This paper proposes an approach to solve two problems as the following:

1. why we integrate AHP and BSC?
2. how we build the dynamic model for integrating AHP and BSC?

2.2 Why We Integrate AHP and BSC

AHP method is developed by Prof. Saaty and based on three important components 1.) the hierarchy articulation of the elements of the decision problem, 2) the identification of the priority, 3) checking the logic consistency of the priority [12].

The procedure is articulated in different steps. The first step consists of the definition of the problem and of the identification of the criteria in a hierarchy of some levels as goal, criteria, sub-criteria, indicators, indices, and alternatives. After defining the hierarchy articulation of the elements, the second step consists of assessing the value of the weights related to each criterion through the pair-wise comparison between the elements. The scale for the assignment of priority values can be shown by Figure 1 [12]. By this method, different criteria, sub-criteria, indicators, indices, and alternatives in the same level can be weighted with a homogeneous measurement scale. Finally, this method is able to check the consistency of the matrix through the calculation of the eigenvalue. The limitation of consistency of the pair-wise comparison matrix tolerates inaccuracy from decision maker (DM).



1=Equal, 3=Moderate, 5=Strong, 7=Very Strong
9=Extreme, and 2, 4, 6, 8 are the intermediate values
between the two adjacent judgments.

Fig. 1. The scale of AHP [12]

Actually, the characters of AHP are simplicity, ease of use, flexibility and its ability to handle complex and ill-structured problems. Three features of the AHP differentiate it from other decision making approaches [13],[14]:

1. its ability to handle both tangible and intangible attributes;
2. its ability to structure the problems, in a hierarchical manner, to gain insights into the decision-making process;
3. its ability to monitor the consistency with which a decision maker makes a judgment.

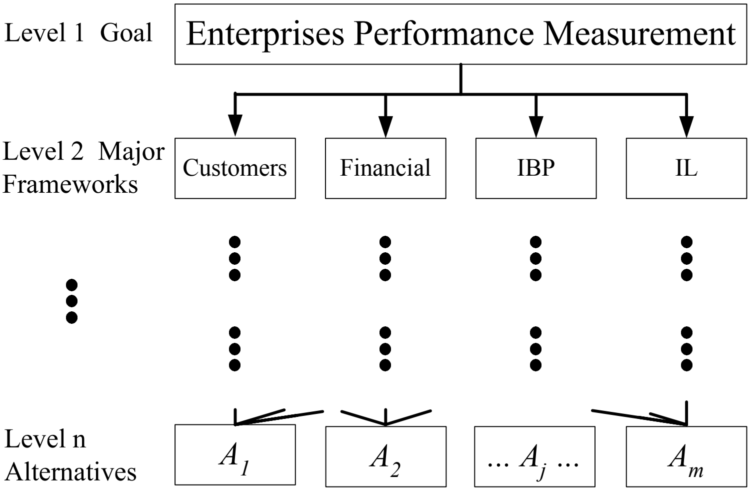


Fig. 2. The hierarchy base on AHP and BSC

Many researchers have concluded that AHP is a useful, practical and systematic method for vendor rating and has been applied successfully [15]-[20].

But one of the AHP's limits is DM should structure the complete hierarchy which reflect all frameworks of goal. In some cases, DM uses the incomplete hierarchy, and make unfitting conclusion. For completely estimating the performance of enterprises, balanced scorecard (BSC) is useful and suggested by Kaplan and Norton [21]. The BSC promotes the balanced pursuit of objectives in four key areas: customers, financial, internal business processes (IBP), and innovation and learning (IL). The balanced scorecard provides summary-level data emphasizing the most critical to the company in each of the four areas [22], [23].

Clinton, Webber, and Hassell (2002) proposed a hierarchy include: customers (revenue, market share, and quality function deployment score), financial (cash value-added, residual income, and cash flow), Return on investment (number of good units produced, minimizing variable costs per unit, and number of on-time deliveries), and IL (market share, number of new products, and revenue from products and services) [24]. In this hierarchy, the attribute "market share" is included in both customers and IL.

Besides, Searcy (2004) also suggests integrating the AHP and BSC for estimating the performance of enterprises and uses quality, safety, financial, customer, operating, and employee to structure the analytic frameworks [25].

Thus, we issue the hierarchy based on 4 major frameworks of BSC. The diagram can be shown as Figure 2.

2.3 Proposed Dynamic Method

Before illustrate the proposed methodology, we assume:

1. Adjacent x_{ijk} and $x_{ij(k+1)}$ are linear and continuous,
2. $p \geqq 3$,
3. DM can evaluate the value of attributes in analytic period.

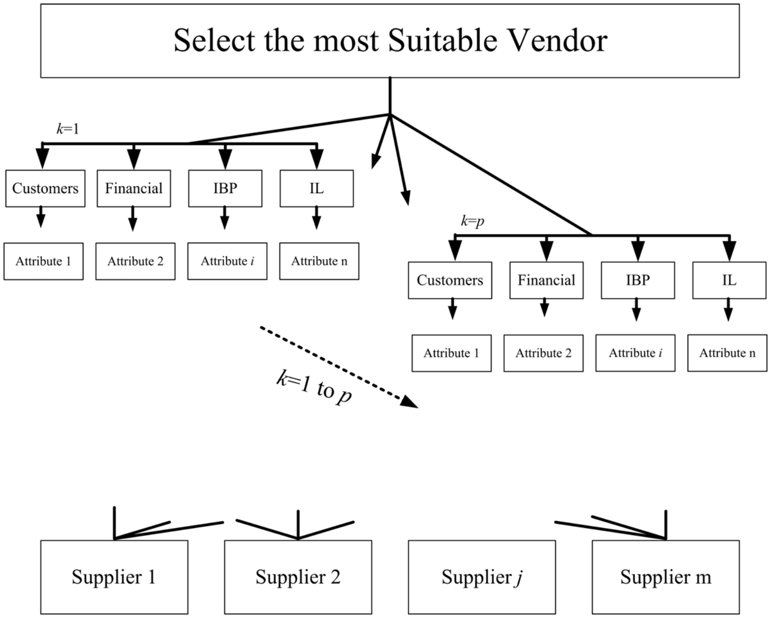


Fig. 3. The diagram of proposed dynamic approach

Table 1. The relations of 4 major frameworks and attributes in vendor selection problem hierarchy

Major frameworks	Attributes
Customers	Number of customers' complaint (NCC)
	Response speed to customers (RSC)
	Market share (MS)
	Products average price (PAP)
Financial	Cash value-added (CVA)
	Residual income (RI)
	Return on investment (ROI)
	Revenue
IBP	Number of good units produced (NGUP)
	Minimizing variable costs per unit (MVCPU)
	Number of on-time deliveries (NOTD)
	Electronic business capability (EBC)
IL	Number of new products (NNP)
	Revenue from maintenance (RM)
	Number of patent (NP)
	Employees' training hours (ETH)

As the Figure 3, we use AHP for each k=1 to p and integrate the analytic period (k=1 to p) to a synthetic index.

2.4 Attributes

For describing the 4 major frameworks as the figure 2 in vendor selection problem, the attributes can be shown as the Table 1.

3 Case Study

We use company A for illustrating the proposed model as the following 6 steps:

Step 1: For k=1 to p, estimating the scores of attributes' pare-wise comparison in hierarchy. For vendor selection problem, the project of company A has 16 members, 12 are the managers or the key man in relational departments and others are the external consultants. The setting parameters are p=3, n=16, and m=3. The project uses Delphi method to unify the commands of the members. The results of expert judgments can be shown as Table 2.

Table 2. The attributes' pare-wise comparison matrix at k=1, 2, 3

x_{ijk} (Goal)	Customers	Financial	IBP	IL
Customers	(1,1,1)	(5,5,5)	(5,5,5)	(3,2,1)
Financial	-	(1,1,1)	(1/2,1,1)	(2,1,1/4)
IBP	-	-	(1,1,1)	(2,1, 1/4)
IL	-	-	-	(1,1,1)
x_{ijk} (Customers)	NCC	RSC	MS	PAP
NCC	(1,1,1)	(1,2,3)	(1/2,1/2,1/3)	(1/7, 1/7, 1/3)
RSC	-	(1,1,1)	(1,1, 1/3)	(1/6,1/7, 1/3)
MS	-	-	(1,1,1)	(1/5,1/5,1)
PAP	-	-	-	(1,1,1)
x_{ijk} (Financial)	CVA	RI	ROI	Revenue
CVA	(1,1,1)	(2,2,2)	(1,1/3,2)	(1,1/2,2)
RI	-	(1,1,1)	(1/3, 1/3,1/3)	(1,1,1)
ROI	-	-	(1,1,1)	(2,1,1)
Revenue	-	-	-	(1,1,1)
x_{ijk} (IBP)	NGUP	MVCPU	NOTD	EBC
NGUP	(1,1,1)	(2,1,2)	(2, 2, 2)	(1/3, 1/3,1/3)
MVCPU	-	(1,1,1)	(1,1,1)	(1/5,1/4,1/4)
NOTD	-	-	(1,1,1)	(1/4,1/5,1/5)
EBC	-	-	-	(1,1,1)
x_{ijk} (IL)	NNP	RM	NP	ETH
NNP	(1,1,1)	(2,2,2)	(3,3,3)	(2,2,2)
RM	-	(1,1,1)	(2,2,2)	(1,1,1)
NP	-	-	(1,1,1)	(1,1,1)
ETH	-	-	-	(1,1,1)

Step 2: For *k*=1 to *p*, estimating the scores of alternatives' pare-wise comparison in hierarchy. In case study, 3 companies are made pare-wise comparison based on expert judgments. The scores can be shown as Table3.

Table 3. The scores of alternatives' pare-wise comparison matrix at *k*=1, 2, 3

	A_1	A_2	A_3
NCC	(0.443,0.443,0.200)	(0.169,0.169,0.400)	(0.388,0.388,0.400)
RSC	(0.413,0.413,0.333)	(0.327,0.327,0.333)	(0.260,0.260,0.333)
MS	(0.540,0.540,0.540)	(0.297,0.297,0.297)	(0.163,0.163,0.163)
PAP	(0.250,0.163,0.210)	(0.250,0.297,0.240)	(0.500,0.540,0.550)
CVA	(0.327,0.327,0.327)	(0.260,0.260,0.260)	(0.413,0.413,0.413)
RI	(0.550,0.550,0.550)	(0.210,0.210,0.255)	(0.240,0.240,0.255)
ROI	(0.413,0.311,0.250)	(0.327,0.493,0.500)	(0.260,0.196,0.250)
Revenue	(0.400,0.297,0.290)	(0.400,0.540,0.540)	(0.200,0.163,0.163)
NGUP	(0.443,0.443,0.443)	(0.169,0.169,0.169)	(0.388,0.388,0.388)
MVCPU	(0.413,0.413,0.413)	(0.327,0.327,0.327)	(0.260,0.260,0.260)
NOTD	(0.388,0.388,0.388)	(0.443,0.443,0.443)	(0.169,0.169,0.169)
EBC	(0.443,0.443,0.493)	(0.387,0.387,0.311)	(0.169,0.169,0.196)
NNP	(0.413,0.210,0.250)	(0.327,0.550,0.500)	(0.260,0.240,0.250)
RM	(0.327,0.327,0.327)	(0.413,0.413,0.413)	(0.260,0.260,0.260)
NP	(0.387,0.387,0.387)	(0.443,0.443,0.443)	(0.169,0.169,0.169)
ETH	(0.142,0.142,0.142)	(0.429,0.429,0.429)	(0.429,0.429,0.429)

Step 3: For *k*=1 to *p*, analyzing consistency. For determining consistency, the project team use Saaty's consistency ratio (CR). If the CR is less than 0.10, the pair-wise comparisons are supposed consistent within acceptable random variations. The calculating formula is as the following [12]:

$$CR = \frac{CI}{RI} = \frac{\lambda_{\max} - m}{RI \cdot (m-1)} \quad , \tag{1}$$

which

CR = consistency ratio,

CI = consistency index,

RI = random index,

$\lambda_{\max}$ = principal eigenvalue,

m = the number of possible alternatives.

The consistency of Table 2 can be shown by Table 4, and each CR < 0.1.

Table 4. The analytic hierarchy consistency

CR	Goal	Customers	Financial	IBP	IL
k=1	0.09	0.02	0.03	0.01	0.02
k=2	0.04	0.06	0.09	0.02	0.02
k=3	0.00	0.06	0.06	0.01	0.02

Step 4: For k=1 to p, calculating average score. The individual synthetic scores (ISS) at k=1, 2, 3 can be calculated by product of alternatives and attributes scores. The calculating formula of comprehensive performance index (CPI) can be shown as the following:

$$CPI = \frac{\sum_{k=1}^{p} ISS_k}{p} , \tag{2}$$

which
CPI = comprehensive performance index,
ISS_k = individual synthetic scores at k,
p = number of analytic time sections.

If the adjacent synthetic scores are linear and continuous, the total comprehensive performance index can be got by averaging individual synthetic scores at k=1, 2, 3, and the result can be shown in Table 5.

Table 5. The synthetic scores and the total comprehensive performance indices

	A_1	A_2	A_3	Sum
ISS_1	0.359	0.296	0.345	1
ISS_2	0.305	0.346	0.349	1
ISS_3	0.319	0.369	0.312	1
$\sum_{k=1}^{3} ISS_k$	0.983	1.011	1.006	3
CPI	0.328	0.337	0.335	1

Step 5 : Ranking. From Table 5, A_2 is better choice than A_1 and A_3 in the analytic period.

In case study, A_1 is the most suitable vendor at k=1 for company A. But A_2 is the best choice in the long-term period. Traditional AHP makes the decision at the same time section. If DM may change the scores of attributes and alternatives in the long-term period, the proposed method is better than traditional AHP. The case study shows an example for the main contribution of proposed method.

4 Conclusion and Future Work

This paper proposes a dynamic approach based on AHP and BSC for vendor selection problems. The analytic hierarchy is structure by the 4 major frameworks including customers, financial, internal business processes, and innovation and learning. The main contribution is extend AHP for the changeable scores of pare-wise comparison in long-term decision making problem. Decision maker can estimate the changeable scores in analytic period. The hypotheses are 1.) adjacent scores are linear and continuous, 2.) p□3, 3.) decision maker can evaluate the value of attributes in analytic period. The proposed method is useful in global supply chain management because of the swift changed business environment. In future work, we will try to extend proposed method in fuzzy environment.

References

1. Dickson, G.: An Analysis of Vendor Selection Systems and Decisions. Journal of Purchasing. 2 (1966) 28–41
2. Lamberson, L.R., Diederich, D., Wuori, J.: Quantitative Vendor Evaluation. Journal of Purchasing and Materials Management. 12 (1976) 19-28
3. Roberts, B.J.: A Vendor Delivery Rating Model. Journal of Purchasing and Materials Management. 14 (1978) 12-16
4. Hinkle, C.L., Robinson, P.J., and Green, P.E.: Vendor Evaluation Using Cluster Analysis. Journal of Purchasing. 5 (1969), 49-58
5. Charles, A., Weber, R., Benton, W.C.: Vendor Selection Criteria and Methods. European Journal of Operational Research. 50 (1991) 2-18
6. Thompson, K.N.: Scaling Evaluative Criteria and Supplier Performance Estimates in Weighted Point Prepurchase Decision Models. International Journal of Purchasing and Materials Management. 27 (1991) 27-36
7. Petroni, A., and Braglia, M.: Vendor Selection Using Principal Component Analysis. The Journal of Supply Chain Management. 36 (2000) 63-69
8. Narasimhan, R., Talluri, S., and Mendez, D.: Supplier Evaluation and Rationalization via Data Envelopment Analysis: An Empirical Examination. Journal of Supply Chain Management. 37 (2001) 28-37
9. Simpson, P., Siguaw, J., and White, S.: Measuring the Performance of Suppliers: An Analysis of Evaluation Processes. Journal of Supply Chain Management. 38 (2002) 29-41
10. Muralidhara, C., Anantharaman, N., and Deshmukh, S.G.: A Multi-Criteria Group Decisionmaking Model for Supplier Rating. Journal of Supply Chain Management. 38 (2002) 22-33
11. Chan, F. T. S. and Qi, H. J.: A Fuzzy Basis Channel-Spanning Performance Measurement Method for Supply Chain Management. Proceedings of the I MECH E Part B, Journal of Engineering Manufacture. 216 (2002) 1155-1167
12. Saaty, T.L.: The Analytical Hierarchy Process, McGraw-Hill, New York (1980)
13. Vargas, L. G.: An Overview of the Analytic Hierarchy Process and its Application. European Journal of Operational Research. 48 (1990) 2-8
14. Wedley, W. C. Combining Qualitative and Quantitative Factors - an Analytic Hierarchy Process. Socio-Economic Planning Sciences. 24 (1990), 57-64
15. Nydick, R.L., Hill, R.P.: Using the Analytic Hierarchy Process to Structure the Supplier Selection Procedure. International Journal of Purchasing & Materials Management. 28 (1992) 31-36
16. Barbarosoglu, G., Yazgac, T.: An Application of the Analytic Hierarchy Process to the Supplier Selection Problem. Production & Inventory Management Journal. 38 (1997) 14-21
17. Yahya, S., Kingsman, B.: Vendor Rating for an Entrepreneur Development Programme: a Case Study Using the Analytic Hierarchy Process Method. Journal of the Operational Research Society. 50 (1999) 916-930
18. Tam, C.Y., Tummala, V.M.: An Application of the AHP in Vendor Selection of a Telecommunications System. Omega. 29 (2001), 171-182
19. Chan, F.T.S.: Interactive Selection Model for Supplier Selection Process: an Analytical Hierarchy Process Approach. International Journal of Production Research. 41(2003) 3549-3579

20. Ge, W., Huang, S.H., Dismukes, J.P.: Product-driven supply chain selection using integrated multi-criteria decision-making methodology. International Journal of Production Economics. 91 (2004) 1-15
21. Kaplan, R.S., Norton, D.P.: The Balanced Scorecard- Measures that Drive Performance. Harvard Business Review. (1992) 71-79
22. Kaplan, R.S., Norton, D.P.: Putting the Balanced Scorecard to Work. Harvard Business Review. (1993) 134-142
23. Kaplan, R.S., Norton, D.P.: Using the Balanced Scorecard as a Strategic Management System. Harvard Business Review. (1996) 75-85
24. Clinton, D., Webber, S.A., and Hassel, J.M.: Implementing the Balanced Scorecard Using the Analytic Hierarchy Process. Management Accounting Quarterly. 3 (2002), 1-11
25. Searcy, L. D.W., Aligning the Balanced Scorecard and a Firm's Strategy Using the Analytic Hierarchy Process. Management Accounting Quarterly. 5 (2004), 1-10

A Fuzzy-Expert-System-Based Structure for Active Queue Management

Jin Wu[1,2] and Karim Djemame[3]

[1] School of Computer Science and Engineering, Beihang University,
100083 Beijing, China
jinwu@buaa.edu.cn
[2] Sino-German Joint Software Institute, Beihang University,
100083 Beijing, China
[3] School of Computing, University of Leeds,
LS2 9JT Leeds, United Kingdom

Abstract. In this paper, we demonstrate an example of using artificial intelligent in solving problems with complex and uncertain features in communication networks. The concept of Fuzzy Expert System is used in the design of an Active Queue Management (AQM) algorithm. Expert System and Fuzzy Logic are commonly used methods in solving various kinds of uncertain problems. Network congestion control is a problem with large scale and complexity, where no accurate and reliable model has been proposed so far. We believe Fuzzy Expert System methods have the potential to be applied to congestion control and solve those problems with uncertainties. This research demonstrates the possibility of using Fuzzy Expert System in the network congestion control. In this paper, a fuzzy-expert-system-based structure is proposed for network congestion control and a novel AQM algorithm is introduced. Simulation experiments are designed to show that the fuzzy-expert-system-based AQM algorithm exhibits a better performance than conventional approaches.

1 Introduction

Heterogeneous protocols and complex topologies of communication networks bring huge difficulties to build any accurate abstract model within practical acceptable complexity. Thus it is not easy to design congestion control solutions using conventional approaches like “tidy” mathematical reasoning. Although there has been recently some contributions in the literature that use control theory to design congestion control solutions in communication networks, those solutions either rely too much on the “critical network environment” that applies unrealistic assumptions or are too complex to be widely deployed in practice. It is believed that conventional system analysis approaches based on mathematical model description and reasoning are inefficient in dealing with those complex system control problems with the following features: 1) systems with large uncertainty or interference; 2) highly non-linear systems; 3) complex task control like control of autonomous machines. We believe that the Internet are highly nonlinear systems with large uncertainty, and the conventional control approaches relying on accurate mathematical models are not the best solution.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 266 – 275, 2005.

Therefore, in the paper, we apply intelligent control in the design process of a network congestion controller in order to gain better performance. We concentrate our research in proposing a structure for an Expert Congestion Control structure and apply this structure in queue management. It is shown through simulation experiments that the network performance receives a significant enhancement after applying the Fuzzy Expert Congestion Control structure.

This paper is organized as follows. In section 2, work related to our research is presented. In section 3, a structure for Expert Congestion Controller System is proposed. Following, Section 4 constructs a novel AQM algorithm based on the structure. Simulation experiments are designed in Section 5 to prove that the new algorithm does introduce a significant performance improvement. Finally, some conclusions and acknowledgements are given in Section 7 and Section 8.

2 Related Work

Extensive research has been done on Random Early Detection (RED) since it was introduced by Floyd and Jacobson [3]. There are many arguments on whether RED can improve end-to-end performance. The Internet Engineering Task Force (IETF) recommended its deployment [8], while some researchers showed evidence of opposing its use [4, 13]. Many research papers using mathematical modelling [11, 20, 21] and simulation experiments [19] to evaluate the performance of RED are found in the literature. Other papers treat essentially design guidelines of AQM algorithms [15, 16]. Several RED variations are found in the literature: FRED [17], SRED [12], BRED [18], and ARED [1] are among those that received most attention. FRED, SRED and BRED use per flow queuing algorithms. REM [15, 16, 19] is a new proposed AQM algorithm that received increase attention by the research community. REM uses a (mathematical) duality model to simulate the network congestion control process. REM has two drawbacks that make it hardly fit in current network environments: 1) REM needs to revise the source algorithm when deployed; and 2) REM cannot work along current AQM algorithms such as RED. Model abstraction for communication networks has attracted great interest among the research community [14]. Classic control theory is also used for network modelling [4], [11]. The network congestion control process is converted into a close loop system. This model works well in single switch node system, but not in large-scale networks as there are more model parameters affecting end-to-end performance.

To the best of our knowledge, this research is the first attempt to combine together Expert intelligent control and network congestion management. Moreover, the terms “Intelligent Congestion Controller” and “Expert Congestion Controller” are used here for the first time. Although works on the application of intelligent control to communication networks is found in e.g. [5, 6], they are more alike pure adaptive algorithms which do not prove the use of knowledge structure supporting machine intelligence. Current AI research in the communication networks mostly focuses on human interaction and the application layer. Although not quite active, there is some research on applying AI in the field of Transmission subnet management. Reference [2] presents an Intelligent Agent Architecture and a Distributed Artificial Intelligent based approach for Network Management (NM) where a NM system based on intelligent agents, claimed to be more elastic than conventional centralized approaches, is proposed. Applying fuzzy logic in the congestion control process is also an area that researchers are looking at. References [10, 23] apply fuzzy logic in the Available Bit rate (ABR) congestion control in ATM networks. Reference [22] uses fuzzy logic for the design of an AQM algorithm and proposes Fuzzy RED.

3 Expert Congestion Control System Design

In this section, expert controllers that are originally developed in the field of automation control are extended to be suitable for running in communication network environments in order to control congestion. The structure and data flow of Expert Congestion Control system is introduced as follows.

How to organize the information and knowledge for the Expert Congestion Control System highly depends on its efficiency. A structure is proposed to meet the features of communication network congestion controller. The hierarchical structure model that follows the principle of Increasing Precision with Decreasing Intelligence (IPDI) is depicted in Figure 1.

The observation level can be viewed as independent from the congestion controller and acts as the system identifier in the control system. It traces dynamics of those observable variables of the control objects and passes system dynamic information in the form of a state cognitive vector to the organization level. It also passes system state information to Execution level as to generate control signals. Organization and Coordination levels which contain high level intelligence are intend to perform such operations as planning and high-level decision making. The Organization level is based on the Expert System forming the fundamental of machine intelligence. Based on the knowledge that it contents, the expert system makes decisions of the system patterns' probable of the network system thus achieves the control target set by the upper level which can be a human or other intelligent machine. It transmits the decision result in the form of a State Pattern Vector to the lower level as to indicate the state of the network system that is currently tracing. The Coordination level is an intermediate structure serving as an interface between the Organization and Execution level. Its intelligence relates to its ability to how to advise the low layer a control pattern in its best possible way under certain state patterns given by the upper level. A vector, called the Control Pattern Vector, is defined as the message passing mechanism from the Coordination level to the Execution level. The Control Pattern Vector can be viewed as the result of qualitative analysis. The Organization and Coordination levels obtain information from the network and make qualitative decisions to advise which control mode (or pattern) is most efficient. Moreover, the Coordination level is also responsible to supervise the Observer and direct distributed controls in network system. The Execution level belongs to quantitative analysis where it needs to generate precision control signals. The control elements generate control signals for the network system when activated by the Control Pattern Vector. Control elements can be built up by either the control pattern or control mode and serve as basic control units in the Expert Congestion Control system for congestion control. Details of congestion elements when they are set up in real situations are given in the following section.

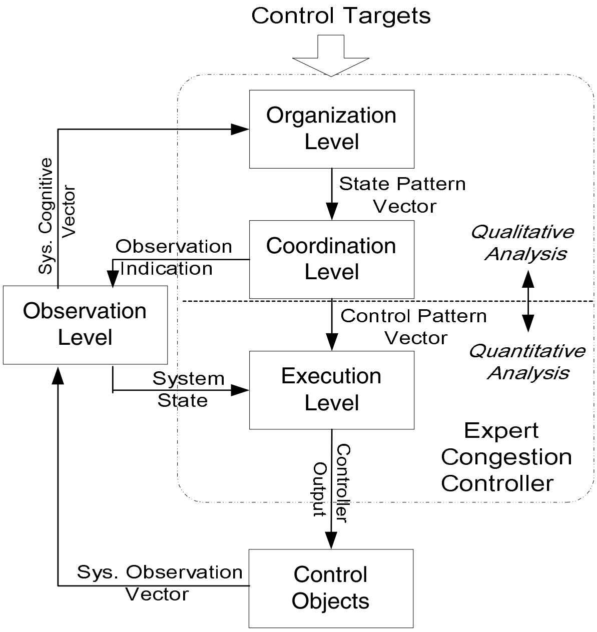


Fig. 1. Hierarchical Structure of Expert Congestion Controller

Altogether, three types of transcendental knowledge are included respectively in the Organization level, Coordination level, and Execution level in Expert Congestion Controller. The knowledge contained in the Organization level offers the ability to realize the system state. The Coordination level includes the knowledge that directs the control missions to correct units. The knowledge in the Execution level decides the most suitable control signal. Roughly, the expert congestion control system is understood as follows: a) the Organization level decides what the system is; b) the Coordination level decides where to control; c) the Execution level decides how to control the system. Therefore, the expert congestion controller works in the way more like what a human does.

4 An Example of Fuzzy-Expert-System-Based AQM Algorithm

In the following, a simple AQM algorithm based on FIFO that runs on TCP/IP networks is proposed. First, the Internet dynamic is studied. Then, the generation process of the new AQM algorithm, Fuzzy Marking (FM), is introduced.

4.1 Queue Dynamic in the Internet

TCP is the most widely used congestion control mechanism in the Internet. For the TCP/IP Protocol Suite, there exists an Explicit Congestion Notification (ECN) [8] mechanism that allows the switch-nodes to notify TCP sources of the congestion states, and thus control the packet sending probability of TCP sources by the probabilistically marking of the ECN bit in the Type of Service (ToS) field of the IP header. The TCP sources slow down the transmission rate when an ECN marked acknowledgement packet is received. For TCP sources, the sending rate can be measured by the following formula [7]:

$$Rate = 0.93 * \frac{MSS}{Rtt\sqrt{p}} \tag{1}$$

where MSS is the segment size being used by the connection, Rtt is the Round Trip Time, and p is the marking probability that the TCP source receives. Marking probability normally stands for the combination rate of ECN-bit marking probability and packet loss ratio. If the packet loss is negligible, then p equals to the ECN marking probability. It is worth pointing out that as long as only the control problem is being considered, it is reasonable to assume that no packet loss occurs in the network system. The packet loss problem is dealt with by other mechanisms. Suppose N TCP connections with the same Rtt access the switch-node k, and the egress bandwidth of the switch-node is BW. The queuing at the switch-node is considered as the M/M/1 queuing system, while following the Queuing Theory, the queue length of the switch-node k is

$$queue_k = \frac{N \cdot rate}{BW - N \cdot rate} \tag{2}$$

Using Equation (1) and (2),

$$\frac{dqueue_k}{dp} < 0 \tag{3}$$

4.2 The Fuzzy Expert Congestion Control Structure in Practise

It is common knowledge with regard to packets queuing in networks that a high queue occupancy brings more queuing delay and a low number of packets in queues harms efficiency. Queue length at gateway is important to network performance. An ideal queue length can in one hand maintains good link efficiency, and on the other hand keeps a reasonable queuing at switch-nodes. Based on this analysis, the queue length at switch node is selected as the System Observation Vector and a low pass filter and a mechanism of deciding ideal queue length are executed in the Observation level. Then, the System Cognitive Vector is set as the offset between average queue length and the ideal queue length. Fuzzy logic is used in the Organization level where five tag patterns (or in terms of lingistical value), w_1, w_2, w_3, w_4, and w_5, are defined. w_1 stands for severely less-queued, w_2 stands for slightly less-queued, w_3 stands for ideal queued, w_4 stands for slightly over-queued, and w_5 stand for severely over-queued. The membership function M is plotted in Figure 2. In this controller, there is only a single controllable element, so the Coordination level does nothing but directly passes the state pattern vector as the control pattern vector and sends it to the output queue of the gateway.

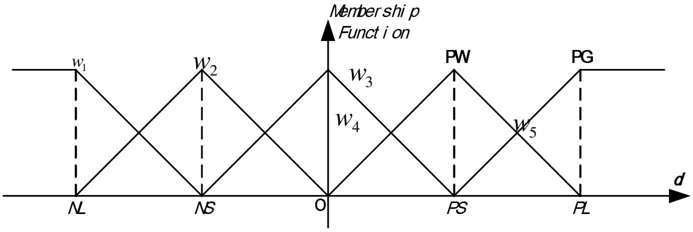


Fig. 2. Membership Function for Fuzzy Expert System

Then, the definition of the Execution level is discussed. Variable *d* represents for the offset between existing queue length and ideal queue length. From equation (3),

$$\frac{dqueue_k}{dt} = -\frac{B}{2} * \frac{dp}{dt}$$

where *B* is a positive coefficient. In the switch-nodes, the discrete additive mechanism is used in tuning marking probability. Four step length h_{NL}, h_{NS}, h_{PS}, h_{PL} are defined as "negative large", "negative small", "positive small", and "positive large" real numbers, respectively. Therefore, a policy set with capacity of 5 is defined as follows:

$$[p^{(t+1)}=p^{(t)}+h_{NL},\ p^{(t+1)}=p^{(t)}+h_{NS}, p^{(t+1)}=p^{(t)}, p^{(t+1)}=p^{(t)}+h_{PS}, p^{(t+1)}=p^{(t)}+h_{PL},]$$

The control output process is defined as Centre-average defuzzification. Therefore, the controller output can be given as

$$p^{(t+1)} = p^{(t)} + h_{NL} \cdot M_{w_1} + h_{NS} \cdot M_{w_2} + h_{PS} \cdot M_{w_4} + h_{PL} \cdot M_{w_5}$$

Thereby, we accomplish the design process of the congestion controller as depicted in Figure 1.

5 Simulation Analysis

Simulation experiments are designed in this section to prove that significant performance improvement is possible even under conventional mechanisms when an fuzzy expert congestion control system is used. Following, FM, RED, and REM are compared through simulation experiments with TCP Reno in terms of queue length at routers. We also compare the end-to-end performance of passing through an FM capable gateway against RED and REM capable switch-nodes. This is to prove that the performance in terms of queue length at switch-nodes can indeed be improved by using fuzzy expert system in queue managements. In this experiment, we trace the queue length at congested gateway for TCP connections accessing it. The bottleneck gateway is set to different congestion states to prove that FM can be successful in improving end-to-end performance by keeping a stable queue length.

5.1 Simulator Settings

The simulation study is based on the network simulator ns-2 [9]. Simulation is performed on the network where an interactive communication shares a bottleneck link with cross traffics (Figure 3). The bottleneck link is set to 15Mb with 10ms propagation delay. Other links are all set to 20Mb bandwidth with 10ms propagation delay. In this experiment, cross traffic is simulated using an N sources configuration consisting of N identical TCP sources and sinks. All sources and sinks are connected to a router with N TCP connections passing through the bottleneck link. All cross traffic sources are supplied by FTP applications. The TCP default packet size is 1KB. The buffer capacity at the router is 100 packets. Packets are served in First In First Out (FIFO) order and are marked with the ECN bit using the probability of the AQM algorithm (RED, REM or AOM).

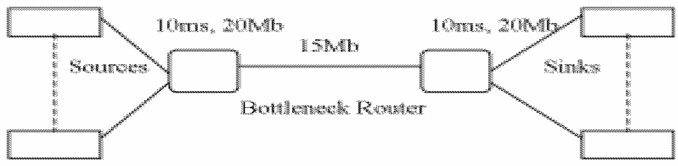


Fig. 3. Network Topology

Two simulation scenarios are used in this experiment, namely simulation 1 and Simulation 2.

Simulation 1

Our first experiment studies the impact of various propagation delays under a number of TCP connections. In this experiment, RED parameters are set as follows: max_{th}=60, min_{th}=20, max_p=0.1 and w=0.002. REM gateway is set with default setting. For FM, we set N_L, N_S, P_S, and P_L as –10, -5, 5, and 10 respectively, and the ideal queue length as 10. The *update_interval* is set to 10ms. The performance metric used in these experiments is Queue Length at gateway. The simulation duration is set to 110 seconds. In order to avoid the impact of the Slow Start phase, the simulation measurements start at time t=10 seconds. All the TCP sources start transmitting at the same time. We set the configuration vector A=[*TCPConn, Delay*] as follows:

TCPconn [10, 50, 75, 100, 125, 150, 175, 200, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500]: Number of TCP connections

Delay [0 10 50 100 150 200 250]: Bottleneck propagation delay (in ms)

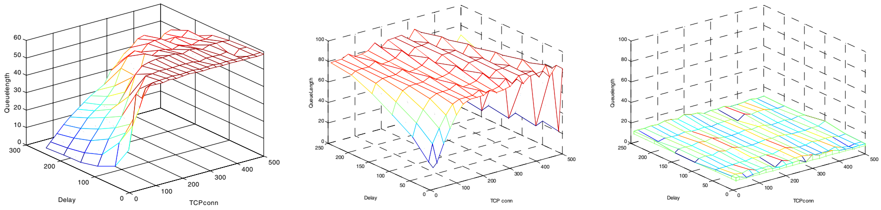
Experiment 1 traces the queue length at the switch-node as the performance matrix.

Simulation 2

RED, REM and FM parameters are set as in simulation 1. The propagation delay of the bottleneck line is set to 10ms. The simulation duration is 70 seconds. To show that FM queue length is more static than RED and REM. We vary the data traffic as follows: the simulation starts at time t=0 and the number of TCP connections is set to 1; from time 0 to 50 seconds, the number of TCP connections is increased by one every second. From time 50 to 70 seconds, the number of TCP connections is kept constant (in this case 50).

5.2 Simulation Results and Discussions

The simulation results are shown in Figure 4. It can be observed that AOM successfully controls the queue occupancy at gateway.

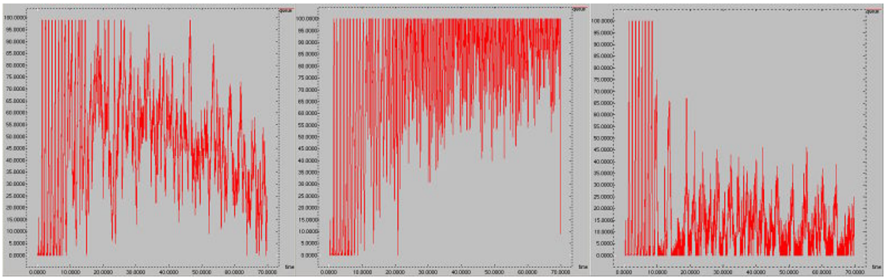


RED REM FM

Fig. 4. Experiment 1 - Performance Comparison RED/REM/FM

In Figure 5, X-axis stands for Time and Y-axis stands for Queue length. Under RED, the queue length at the router increases when the number of active TCP connections increase, and under REM, the capability is weak in controlling the increasing traffics. While under FM, the queue length at the router is not increase under the same traffic conditions. This shows that when fuzzy controller is applied, the router can be more static despite varying traffic load. When more users are accessing the FM capable network, its response time does not receive a significantly increase.

It can be easily observed from Figures 4, 5 that FM provides a lower queuing delay than other AQM algorithms. FM gateway also keeps a more stable queue than RED and REM despite the varying of traffic load. As a conclusion, the FM algorithm that applies expert control and fuzzy logic principles shows a better performance than other widely researched AQM algorithms.



RED REM FM

Fig. 5. Queue Occupancy Comparison between RED, REM & AOM

6 Conclusion

In this paper, the concept of fuzzy expert system is used in the field of communication network congestion control. We bring the concept of expert system and fuzzy logic to tackle the problem of network modelling difficulty. A hierarchical structure of expert congestion controller hence a novel AQM algorithm is proposed. Simulation experiments in NS-2 have shown that the novel AQM algorithm proposed in this paper provides a significant performance improvement against other well-known AQM algorithms. It exhibits the potential of applying the artificial intelligent in solving congestion control and Quality of Service problems by providing a more flexible approach to construct congestion control protocols in order to meet different requirements. Nevertheless, how to program the expert congestion control system and how to maintain it still remain an open issue. Future works will focus on the extension of the expert system and fuzzy logic theorems in order to provide better performance.

Acknowledgement

The author acknowledges the support from the National Natural Science Foundation of China under key project 90412011 and the support from the National 863 Program of China under grant number 2004AA119030.

References

1. Feng, W., Kandlur, D., Shin, K.: A self-configuring RED gateway. In Proc. of INFOCOM' 99, vol 3, pp1280-9, Mar 1999.
2. Gurer, D., Lakshiminarayan, V., Sastry, A.: An Intelligent Agent-Based Architecture for the Management of Heterogeneous Networks, in the Proc. of DSOM '98, 1998.
3. Floyd, S., Jacobson, V.: Random early detection gateways for congestion avoidance. IEEE/ACM Trans. on Networking, Vol.1 No.4:397-41, Aug 1993.
4. Hollot, C. V., Misra, V. , Towsley, D., Gong, Wei-bo.: On designing improving controllers for AQM Routers Supporting TCP Flows. In Proc. of INFOCOM'99, vol 3, pp1736-34, Mar 1999.
5. Park, Young-Keun, Lee Gyungho.: Intelligent congestion control in ATM networks, in Proc. of the 5th Workshop on Future Trends of Distributed Computing Systems, pp369-75, Aug.1995.
6. Hoang, D.B., Yu, Q.: Performance of the fair intelligent congestion control for TCP applications over ATM networks: a simulation analysis, in Proc. of ICATM '99 pp: 390-5. 1999
7. Huston, G.: The Future for TCP, the Internet Protocol Journal, pp.2-27, Vol.3-No.3, September 2000.
8. Braden, B., et al.: "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, April 1998.
9. "ns-2 Network Simulator," Obertain via http://www.isi.edu/nsnam/ns/
10. Pitsillides, A., Sekercioglu, A., Ramamurthy, G.: Effective Control of Traffic Flow in ATM Networks Using Fuzzy Explicit Rate Marking, IEEE JSAC, Vol. 15, Issue 2, pp. 209-25, 1997
11. Hollot, C. V., Misra, V. , Towsley, D. , Gong,Wei-Bo: A Control Theoretic Analysis of RED. In Proc. of INFOCOM01' vol 3, pp1510-9, April 2001.
12. Ott, T. J. , Lakshman, T. V. , Wong, L. H.: SRED: Stabilized RED. In Proc. of IEEE INFOCOM99', vol 3, pp1346-55, March 1999.
13. May, M., Bolot, J., Diot, C., Lyles, B.: Reasons Not to Deploy RED. In Proc. of IWQoS'99, June 1999.

14. Looking Over the Fence at Networks. National Academy Press. ISBN 0-309-07613-7, 2001.
15. Low, S. H.: A Duality Model of TCP and Queue Management Algorithms. under doing, from: http://netlab.caltech.edu
16. Low, S. H., Lapsley, D.: Optimization Flow Control. IEEE/ACM Transactions on Networking, Vol.7 No.6 861-75, Dec. 1999.
17. Lin, D., Morris, R.: Dynamics of random early detection. ACM Computer Communication Review, vol. 27, pp127-37, Oct. 1997.
18. Anjum, F., Tassiulas, L.: Fair Bandwidth sharing among adaptive and non-adaptive flows in the Internet. In Proc. of INFOCOM' 99, vol 3, pp 1412-20, Mar 1999.
19. Athuraliya, S., Low, S. H.: Simulation Comparison of RED and REM, in Proc. of ICON' 00, pp68-72, Sept 2000.
20. Bonald, T., May, M.: Analytic evaluation of RED performance, in Proc. of INFOCOM' 00 vol 3, 1415-24, Mar 2000.
21. Padhye, J., Firoiu, V., Towsley, D., Kurose, J.: Modeling TCP Reno Performance: A Simple Model and Its Empirical Validation, IEEE/ACM Trans. on Networking, Vol.8, No.2, Apr 2000.
22. Loukas, R., Kohler, S., Andreas, P., Phuoc, T-G.: Fuzzy RED: Congestion Control for TCP/IP Diff-Serv, in Proc. of IEEE the 10th Mediterranean Electrotechnical Conference vol.1, pp19-22, Oct 2000.
23. Sekercioglu, Y. A., Pitsillides, A.: Fuzzy Control of ABR Traffic Flow in ATM LANs, in Proc. of IEEE Symposium on Computers and Communications pp 227-32, 1995.

Parameter Identification Procedure in Groundwater Hydrology with Artificial Neural Network

Shouju Li and Yingxi Liu

State Key Laboratory of Structural Analysis for Industrial Equipment,
Dalian University of Technology, Dalian, 116024, China
{Lishouju, Yxliu}@dlut.edu.cn

Abstract. The mathematical model of underground water flow is introduced as basis to identify the permeability coefficients of rock foundation by observing the water heads of the underground water flow. The artificial neural network is applied to estimate the permeability coefficients. The weights of neural network are trained by using BFGS optimization algorithm and the Levenberg-Marquardt approximation which have a fast convergent ability. The parameter identification results illustrate that the proposed neural network has not only higher computing efficiency but also better identification accuracy. According to identified permeability coefficients of the rock foundation, the seepage field of gravity dam and its rock foundation is computed by using finite element method. The numerically computational results with finite element method show that the forecasted water heads at observing points according to identified parameters can precisely agree with the observed water heads.

1 Introduction

A common use of groundwater flow models is to predict the response of an aquifer. While the mathematical and computational aspects of such response predictions are reasonably well developed, the question of how to choose appropriate parameter values for a specific aquifer has not been completely resolved [1]. Traditionally, the determination of aquifer parameters is based on trial-and-error and graphical matching techniques under the assumptions that the aquifer is homogenous and isotropic and a closed-form solution for the governing equation exists [2]. The inverse problem of aquifer parameter identification is often ill-posed. The ill-posedness is generally characterized by the nonuniqueness and instability of the identified parameters. The instability of the inverse solution stems the fact that small errors in heads will cause serious errors in the identified parameters. Classical identification procedures are based on the optimization algorithm. Their drawbacks lie in lacking robustness and global convergence property. With the development of artificial intelligence, Artificial neural network has been widely applied in the inverse problem domain [3,4,5]. Artificial neural networks have gradually been established as a powerful tool in pattern recognition, signal processing, control and complex mapping problems, because of their excellent learning capacity and their high tolerance to partially inaccurate data [6,7]. Oh(2004) proposed the parameter estimation of fuzzy controller and its application to inverted pendulum. The design procedure dwells on

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 276 – 285, 2005.

the use of evolutionary computing and estimation algorithm [8]. Pacella(2004) presented the adaptive resonance theory and the fuzzy network to investigate manufacturing quality control, and analyzed the performances of adaptive resonance theory under the assumption that predictable unnatural patterns are not available [9]. The back propagation network has been successfully in pattern recognition, but the slowness in training still poses some inconveniences for practical applications. Indeed, the convergence of the BP algorithm requires a huge number of iterations, as well as an adequate number of training examples. The proposed approach is based on the BFGS optimization algorithm and the Levenberg-Marquardt approximation that can be capable of fast, stable and cumulative learning. The study is aimed to apply the neural network based on some optimization technologies to parameter estimation of aquifer and to supply model parameter for the seepage field calculation.

2 Calculation of Groundwater Flow Models

Consider a three dimensional steady seepage field, its governing equation can be written as the follows [10]:

$$k_x \partial^2 h / \partial x^2 + k_y \partial^2 h / \partial y^2 k_y + \partial^2 h / \partial z^2 + q = 0 \tag{1}$$

Where k_x and k_y are permeability coefficient in x and y direction, respectively, h is the water head, q *is* the source-sink item. The first kind boundary condition is expressed as the following:

$$h(x, y, z)\,|_{\Gamma_1} = h_0(x, y, z) \tag{2}$$

Where $h_0(x,y,z)$ is the already known head. The second kind boundary condition is written as the following:

$$Q_0(x, y, z)\,|_{\Gamma_2} = k_x \frac{\partial h}{\partial x} l_x + k_y \frac{\partial h}{\partial y} l_y + k_z \frac{\partial h}{\partial z} l_z \tag{3}$$

Where $Q_0(x,y,z)$ is the drainage already known, $l_{x\,,}$ l_y and l_z are the direction cosines of the exterior normal of the boundary in x ,y and z direction , respectively. The following finite element equation can be derived by the principle of the variational method:

$$[G]\{h\} = \{F\} \tag{4}$$

Where $[G]$ is the matrix of the water transmissibility coefficient, which is already known, and $\{F\}$ is the free item. When the boundary condition and the predicted permeability coefficient are determined, the infinite element equation is adopted to compute the distribution of the water head and the drainage in the whole seepage field, which provides modal data to the analysis of the inversion problem of the permeability coefficient.

3 Artificial Neural Networks for the Parameter Identification

An artificial neural network model is a system with inputs and outputs based on biological nerves. The system can be composed of many computational elements that operate in parallel and are arranged in patterns similar to biological neural nets. A neural network is typically characterized by its computational elements, its network topology and the learning algorithm used. Among the several different types of ANN, the feed-forward, multilayered, supervised neural network with the error back-propagation algorithm, the BPN, is by far the most frequently applied neural network learning model, due to its simplicity.

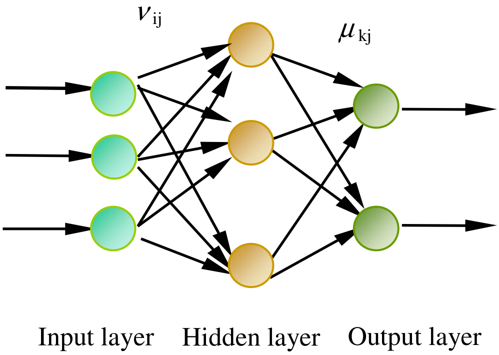


Fig. 1. Topography structure of artificial neural network

The architecture of BP networks, depicted in Figure 1, includes an input layer, one or more hidden layers, and an output layer. The nodes in each layer are connected to each node in the adjacent layer. Notably, Hecht-Nielsen proved that one hidden layer of neurons suffices to model any solution surface of practical interest. Hence, a network with only one hidden layer is considered in this study. Before an ANN can be used, it must be trained from an existing training set of pairs of input-output elements. The training of a supervised neural network using a BP learning algorithm normally involves three stages. The first stage is the data feed forward. The computed output of the i-th node in output layer is defined as follows [6]:

$$y_i = f(\sum_{j=1}^{N_h}(\mu_{ij} f(\sum_{k=1}^{N_i} v_{jk} x_k + \theta_j) + \lambda_i)) \qquad i = 1,2,...N_0 \tag{5}$$

Where μ_{jk} is the connective weight between nodes in the hidden layer and those in the output layer; v_{jk} is the connective weight between nodes in the input layer and those in the hidden layer; θ_{j} or λ_{i} is bias term that represents the threshold of the transfer function *f*, and x_k is the input of the *k*th node in the input layer. Term N_i, N_h, and N_o are the number of nodes in input, hidden and output layers, respectively. The transfer function *f* is selected as Sigmoid function

$$f(\cdot) = 1/[1 + \exp(-\cdot)] \tag{6}$$

The second stage is error Bp through the network. During training, a system error function is used to monitor the performance of the network. This function is often defined as follows

$$E(w) = \sum_{p=1}^{P} (\sum_{i=1}^{N_o} (y_i^p - o_i^p)^2 \tag{7}$$

Where y_i^p and o_i^p denote the practical and desired value of output node i for training pattern p, P is the number of sample. Training methods based on BP offer a means of solving this nonlinear optimization problem based on adjusting the network parameters by a constant amount in the direction of steepest descent, with some variations depending on the flavor of BP being used. The BFGS algorithm is a quasi-Newton optimization technique, in which curvature information is used to prove a more accurate descent direction, without actually calculating the second derivatives. A sequence can be computed according to the formula [7]

$$w(k+1) = w(k) + \alpha(k)d(k) \tag{8}$$

Where w(k) is the vector of network parameters(net weights and element biases) for iteration k, d(k) is the search direction used for iteration k, and $\alpha(k)$ is the step length for iteration k. The search direction will be determined by the following formulas[8]

$$d(k) = -H(k)g(k) + \beta(k)d(k-1) \tag{9}$$

Where *H(k)* is the current approximation to the inverse of the Hessian matrix, and *g(k)* is the current gradient vector. The approximation to *H(*k), *g(k*)and β *(k)* in detail is presented in the references [6,10].

Other algorithm used to train network makes use of the Levenberg-Marquardt approximation. This algorithm is more powerful than the common used gradient descent methods, because the Levenberg-Marquardt approximation makes training more accurate and faster near minima on the error surface. The adjusted weight vector Δw is calculated using a Jacobian matrix ***J*,** a transposed Jacobian matrix $\boldsymbol{J^T}$, a constant μ, a unity matrix **I** and an error vector *e*. The method is as follows [11]

$$\Delta w = (J^T J + \mu I) J^T e \tag{10}$$

The Levenberg-Marquardt algorithm approximates the normal gradient descent method, while if it is small, the expression transforms into the Gauss-Newton method. After each successful step the constant μ is decreased, forcing the adjusted weight matrix to transform as quickly as possible to the Gauss-Newton solution. When after a step the errors increase the constant μ is increased subsequently.

4 Application of ANN to Identification of Permeability Coefficients of Dam Foundation

Baishan Hydropower Station, as shown as Fig.2, is located in the Second Songhuajiang River in Jilin province, China. It consists of a 149.5-meter-high concrete heavy-pressure dam, a weir with four 12×13 meter tunnels on top of the 404-meter-high spillway dam, three 6 × 7 meter tunnels for discharging water are 350 meters high, an underground powerhouse with an installed generating capacity of 900,000 KW and another powerhouse on the surface with an installed generating capacity of 600,000 KW. The dam is 423.5 meters high and the reservoir has a storage capacity of 6.812 billion cubic meters. Its highest normal storage water level is 413 meters. The capacity for water control storage is 3.54 billion cubic meters while the flood control storage capacity is 950 million cubic meters. Cross-section of Baishan dam at block 18 is shown in Figure 3. Figure 4 shows the disposition of observation holes for dam uplift pressure at block 18.

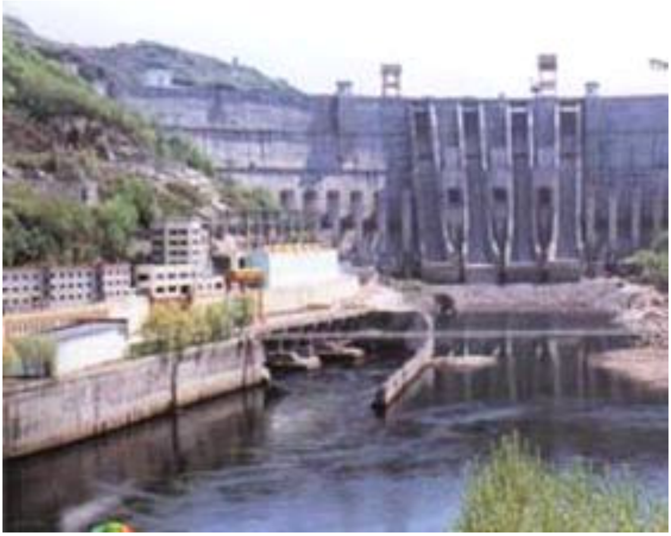


Fig. 2. Baishan Hydropower Station

In order to identify the permeability coefficients of rock foundation, the three-dimensional finite element model for seepage calculation is carried out. The seepage fields of the dam and its rock foundation at different load cases are computed. According to the prior information of pumping water test in field, the domains of identification parameters are determined. The training sample pairs are got basing on finite element analysis. The rock foundation is divided into 3 sub-regions, rock base before concrete certain, concrete certain and rock base after concrete certain. The number of input neurons is determined as 4 according to the number of observing points. And the number of hidden neurons is equal to 8. The number of output neurons is equal to 3, which is equal to the number of sub-regions for back-analysis.

After training ANN with BFGS optimization algorithm, based on measured data of water heads in the observation holes, shown in table1, the permeability coefficients of rock foundation are obtained. Table 2 is identification results of permeability coefficients. Table3 is the comparison between measured and forecasted water heads.

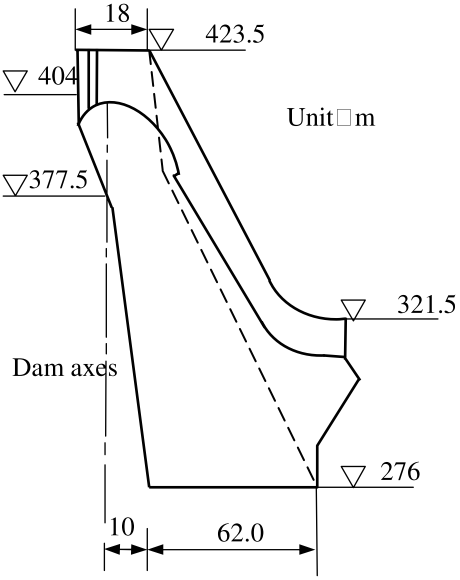


Fig. 3. Cross-section of Baishan dam at block 18

The water heads of the seepage fields must be normalized to a rang of [-1,1] using the equation below in order to identify the parameters of the aquifer. Otherwise, measuring data of water heads have same quantity level, and eventually some of them are almost same because the differences among measuring data are so small that they can be easily distinguished.

$$\hat{x}_n = \frac{2(x_n - x_{n\min})}{(x_{n\max} - x_{n\min})} - 1 \tag{11}$$

Where $\hat{x}_n$ is the normalized network input, x_{nmin} and x_{nmax} are respectively minimum and maximum of water heads of the *n*-th observing point of water heads in all training pairs. As described previously, each unit is a computational element that forms a weighted sum of inputs and passes the result through a sigmoid function is [0,1], outputs of training patterns must be normalized into a unit range of [0,1] before given to the network as teaching signals.

$$\hat{y}_m = \frac{(y_m - y_{m\min})}{(y_{m\max} - y_{m\min})} \tag{12}$$

Where $\hat{y}_m$ is the normalized network input, y_{mmin} and y_{mmax} are respectively minimum and maximum of network output of the m-th neuron in output layer in all training pairs. It is difficult to train the network without such normalization. The normalized output values must then be re-normalized again to obtain the parameters of aquifer. The re- normalization is shown as follows

$$k = \frac{\hat{y}}{(y_{m\max} - y_{m\min})} + y_{m\min} \tag{13}$$

Where k is the identified model parameter.

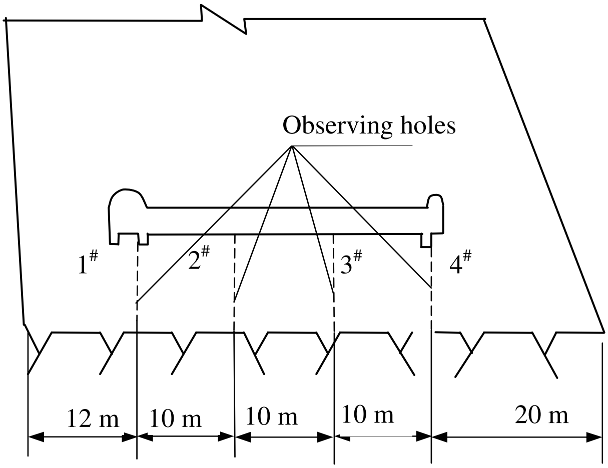


Fig. 4. Disposition of observation holes for dam uplift pressure at block 18

Table 1. Measured data of water heads in the observation holes

Measuring date	Upstream water elevation	Downstream water elevation	water head h_1	water head h_2	water head h_3	water head h_4
19980910	413.00	290.80	291.82	283.59	284.74	281.72

Table 2. Identification results of permeability coefficients

Rock foundation(I) k_1 /10^{-9}m·s^{-1}	Concrete certain k_2 /10^{-9}m·s^{-1}	Rock foundation(II) k_3 /10^{-9}m·s^{-1}
44.05	6.16	52.80

To investigate the influences of the training algorithms on the convergence speed, the three iterative algorithms, including the classical back propagation, BFGS optimization method and the Levenberg-Marquardt approximation, are implemented to train neural networks. The training error versus number of epochs by using different iterative algorithms is shown in Fig. 5.

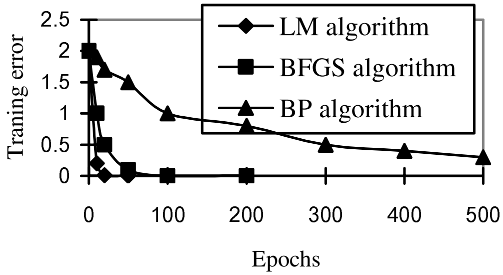


Fig. 5. Comparison among different iterative algorithms

The performance of a trained network can be measured to some extent by the errors on the training, validation and test sets, but it is often useful to investigate the network response in more detail. One option is to perform a regression analysis between the network response and the corresponding targets. The routine post-processing is designed to perform this analysis.

$$A = mT + b \tag{14}$$

Where A is the network output; and T is the corresponding targets to post-processing. It returns three parameters. The first two, m and b, correspond to the slope and the y-intercept of the best linear regression relating targets to network outputs. If we had a perfect fit (outputs exactly equal to targets), the slope would be 1, and the vertical intercept would be 0. From figure 6, we can see that the numbers are very close. The third variable returned by post-processing is the correlation coefficient (R-value) between the outputs and targets. It is a measure of how well the variation in the output is explained by the targets. If this number is equal to 1, then there is perfect correlation between targets and outputs. In our example, the number is very close to 1, which indicates a good fit. The figure 6 illustrates the graphical output provided by post-processing. The network outputs are plotted versus the targets as dots. The best

Table 3. Comparison between measured and forecasted water heads

Measuring date	No.1 measured point	No.2 measured point	No.3 measured point	No.4 measured point
19971015	291.82/291.78	283.54/283.49	284.23/283.19	281.72/281.13
19980108	291.31/291.29	283.59/283.44	282.19/283.15	281.82/282.14
19980901	291.82/291.78	283.59/283.49	284.74/283.19	281.72/282.12
19981012	291.82/291.78	283.49/283.49	284.74/283.19	281.72/282.12

Note: measured water heads/computed water heads

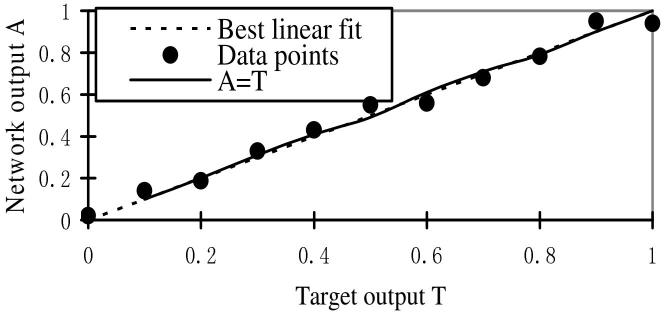


Fig. 6. Training characteristics of neural network

linear fit is indicated by a dashed line. The perfect fit (output equal to targets) is indicated by the solid line. From figure 6, it is difficult to distinguish the best linear fit line from the perfect fit line, because the fit is so good.

5 Conclusion

Based on the finite element model of groundwater flow and the measured water heads of uplift pressure at dam tunnel, artificial neural network is applied to estimate the permeability coefficients of rock foundations of concrete dam. The BFGS optimization method and the Levenberg-Marquardt approximation are used to train and adjust the weights of neural network. It was found that these algorithms can improve the rate of convergence of neural network. According to the identified permeability coefficients of rock foundations of concrete dam and finite element analysis, the forecasted water heads can precisely agree with the observed water heads.

Acknowledgements

This research is funded by the National Natural Science Foundation of China(No. 10472025).

References

1. Carrera, J.:Estimation Of Aquifer Parameters Under Transient And Steady State Conditions. Water Resources Research. 22(1986)199-210
2. Yeh, W.: Review Of Parameter Identification Procedures In Groundwater Hydrology: The Inverse Problem. Water Resources Research. 22(1986)95-108
3. Huang, Yi.: Application Of Artificial Neural Networks To Predictions Of Aggregate Quality Parameters. Int. J. of Rock Mechanics and Mining Sciences. 36(1999)551-561
4. Najjar, Y. M.: Utilizing Computational Neural Networks For Evaluating The Permeability Of Compacted Clay Liners. Geotechnical and Geological Engineering. 14(1996) 193-212
5. Cao, X.: Application Of Artificial Neural Networks To Load Identification. Computers & Structures. 69(1998)63-78

6. Huang, C. S.: A Neural Network Approach For Structural Identification And Diagnosis Of A Building From Seismic Response Data. Earthquake Engineering and Structural Dynamics. 32(2003)187-206
7. Lightbody, G.: Multi-Layer Perceptron Based Modeling Of Nonlinear Systems. Fuzzy Sets and System. 79(1996) 93-112
8. Oh, S. K.: Parameter Estimation Of Fuzzy Controller And Its Application To Inverted Pendulum. Engineering Applications of Artificial Intelligence, 17(2004)37-60
9. Paccella, M.: Manufacturing Quality Control By Means Of A Fuzzy ART Network Trained On Natural Process Data. Engineering Applications of Artificial Intelligence. 17(2004)83-96
10. Denton, J. W.: A Comparison Of Nonlinear Optimization Methods For Supervised Learning In Multilayer Feedforward Neural Networks. European Journal of Operational Research. 93(1996) 358-368
11. Meulenkamp, F.: Application Of Neural Networks For The Prediction Of The Unconfined Compressive Strength From Equotip Hardness. Int. J. of Rock Mechanics and Mining Sciences. 36(1999)29-39

Design of Intelligent Predictive Controller for Electrically Heated Micro Heat Exchanger

Farzad Habibipour Roudsari[1], Mahdi Jalili-Kharaajoo[2], and Mohammad Khajepour[1]

[1] Iran Telecom Research Center, Ministry of ICT, Tehran, Iran
[2] Young Researchers Club, Islamic Azad University, Tehran, Iran
roudsari@itrc.ac.ir, mahdijalili@ece.ut.ac.ir

Abstract. An intelligent predictive control with locally linear neurofuzzy identifier and numerically optimization procedure has been proposed for temperature control of electrically heated micro heat exchanger. To this end, first the dynamics of the micro heat exchanger is identified using Locally Linear Model Tree (LOLIMOT) algorithm. Then, the predictive control strategy based on the LOLIMOT model of the plant is applied to provide set point tracking of the output of the plant. Some computer simulation is provided to show the effectiveness of the proposed controller. Simulation results show better performance for the proposed controller in comparison with PID controller.

1 Introduction

Most of the existing predictive control algorithms use an explicit process model to predict the future behavior of a plant and because of this, the term model predictive control (MPC) is often utilized [1,2] for this control strategy. An important characteristic, which contributes to the success of the MPC technology, is that the MPC algorithms consider plant behavior over a future horizon in time. Thus, the effects of both feedforward and feedback disturbances can be anticipated and eliminated, fact, which permits the controller to drive the process output more closely to the reference trajectory.

Although industrial processes usually contain complex nonlinearities, most of the MPC algorithms are based on a linear model of the process. Linear models such as step response and impulse response models derived from the convolution integral are preferred, because they can be identified in a straightforward manner from process test data. In addition, the goal for most of the applications is to maintain the system at a desired steady state, rather than moving rapidly between different operating points, so a precisely identified linear model is sufficiently accurate in the neighborhood of a single operating point. As linear models are reliable from this point of view, they will provide most of the benefits with MPC technology. Even so, if the process is highly nonlinear and subject to large frequent disturbances, a nonlinear model will be necessary to describe the behavior of the process [3-5].

Recently, neural networks have become an attractive tool in the construction of models for complex nonlinear systems [6-8]. Most of the nonlinear predictive control

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 286 – 295, 2005.

algorithms imply the minimization of a cost function, by using computational methods for obtaining the optimal command to be applied to the process. The implementation of the nonlinear predictive control algorithms becomes very difficult for real-time control because the minimization algorithm must converge at least to a sub-optimal solution and the operations involved must be completed in a very short time (corresponding to the sampling period). In this paper, we will apply a predictive controller to output temperature tracking problem in a electrically heated micro heat exchanger plant [9,10]. First, the nonlinear behavior of the process is identified using a Locally Linear Model Tree (LOLIMOT) network [11,12] and then predictive Controller is applied to the plant. Using the proposed strategy, the tracking problem of the temperature profile will be tackled. The performance of the proposed controller is compared with that of a classic PID controller, which simulation results show better match for the proposed predictive controller.

2 Electrically Heated Micro Heat Exchanger

Electrically heated micro heat exchangers have been developed to accelerate the fluid and gas heating in a reduced space [9,10]. This system consists of a diffusion bonded metal foil stack with many grooves, the heating element are placed between the foils (Fig. 1). In a small volume, powers to 15 kW can be converted. The advantages of this heat exchanger are

- Fluids and gas heated by electrical power and not by additional flow cycle
- Efficient transformation of electrical energy in thermal energy
- Fast temperature change of the media and temperature good fit for sensitive media
- Compact construction due to micro system technology

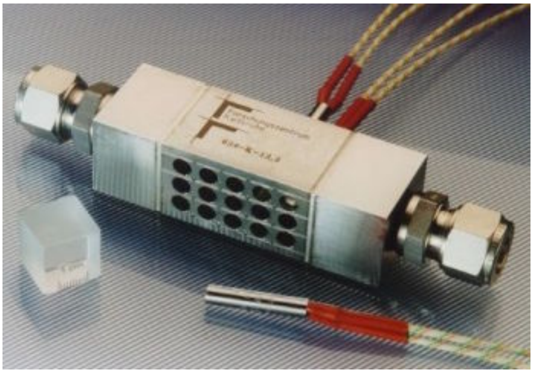


Fig. 1. Electrically heated micro heat exchanger

3 Locally Linear Model Tree Identification of Nonlinear Systems

The network structure of a local linear neuro-fuzzy model [11,12] is depicted in Fig. 2. Each neuron realizes a local linear model (LLM) and an associated validity function that determines the region of validity of the LLM that are normalized as

$$\sum_{i=1}^{M} \varphi_i(\underline{z}) = 1 \tag{1}$$

for any model input $\underline{z}$.

The output of the model is calculated as

$$\hat{y} = \sum_{i=1}^{M} (w_{i,o} + w_{i,1}x_1 + ... + w_{i,n_x} x_{n_x})\varphi_i(\underline{z}) \tag{2}$$

where the local linear models depend on $\underline{x} = [x_1,...,x_{n_x}]^T_.$ and the validity functions depend on $\underline{z} = [z_1,...,z_{n_z}]^T$. Thus, the network output is calculated as a weighted sum of the outputs of the local linear models where the $\varphi_i(.)$ are interpreted as the operating point dependent weighting factors. The network interpolates between different Locally Linear Models (LLMs) with the validity functions. The weights w_{ij} are linear network parameters. The validity functions are typically chosen as normalized Gaussians. If these Gaussians are furthermore axis- orthogonal the validity functions are

$$\varphi_i(\underline{z}) = \frac{\mu_i(\underline{z})}{\sum_{j=1}^{M} \mu_j(\underline{z})} \tag{3}$$

with

$$\mu_i(\underline{z}) = \exp(-\frac{1}{2}(\frac{(z_1 - c_{i,1})^2}{\sigma_{i,1}^2} + ... + \frac{(z_1 - c_{i,n_z})^2}{\sigma_{i,n_z}^2})) \tag{4}$$

The centers and standard deviations are *nonlinear* network parameters. In the fuzzy system interpretation each neuron represents one rule. The validity functions represent the rule premise and the LLMs represent the rule consequents. One-dimensional Gaussian membership functions

$$\mu_{i,j}(z_j) = \exp(-\frac{1}{2}(\frac{(z_j - c_{i,j})^2}{\sigma_{i,j}^2})) \tag{5}$$

can be combined by a t-norm (conjunction) realized with the product operator to form the multidimensional membership functions in (3). One of the major strengths of local linear neuro-fuzzy models is that premises and consequents do not have to depend on identical variables, i.e. $\underline{z}$ and $\underline{x}$ can be chosen independently.

The LOLIMOT algorithm consists of an outer loop in which the rule premise structure is determined and a nested inner loop in which the rule consequent parameters are optimized by local estimation.

1. *Start with an initial model:* Construct the validity functions for the initially given input space partitioning and estimate the LLM parameters by the local weighted least squares algorithm. Set *M* to the initial number of LLMs. If no input space partitioning is available a-priori then set *M* = 1 and start with a single LLM which in fact is a global linear model since its validity function covers the whole input space with $\varphi_i(\underline{z}) = 1$.
2. *Find worst LLM:* Calculate a local loss function for each of the *i*=1,…,*M* LLMs. The local loss functions can be computed by weighting the squared model errors with the degree of validity of the corresponding local model. Find the worst performing LLM.
3. *Check all divisions:* The LLM *l* is considered for further refinement. The hyperrectangle of this LLM is split into two halves with an axis-orthogonal split. Divisions in each dimension are tried. For each division $\dim = 1,\ldots,n_z$ the following steps are carried out:
   (a) Construction of the multi-dimensional MSFs for both hyperrectangles.
   (b) Construction of all validity functions.
   (c) Local estimation of the rule consequent parameters for both newly generated LLMs.
   (d) Calculation of the loss function for the current overall model.
4. *Find best division:* The best of the n_z alternatives checked in Step 3 is selected. The validity functions constructed in Step 3(a) and the LLMs optimized in Step 3(c) are adopted for the model. The number of LLMs is incremented $M \rightarrow M+1$.
5. *Test for convergence:* If the termination criterion is met then stop, else go to Step 2.

For the termination criterion various options exist, e.g., a maximal model complexity, that is a maximal number of LLMs, statistical validation tests, or information criteria. Note that the *effective* number of parameters must be inserted in these termination criteria. Fig. 3 illustrates the operation of the LOLIMOT algorithm in the first four iterations for a two-dimensional input space and clarifies the reason for the term ”tree” in the acronym LOLIMOT. Especially two features make LOLIMOT extremely fast. First, at each iteration not all possible LLMs are considered for division. Rather, Step 2 selects only the worst LLM whose division most likely yields the highest performance gain. For example, in iteration 3 in Fig. 3 only LLM 3-2 is considered for further refinement. All other LLMs are kept fixed. Second, in Step 3c the local estimation approach allows to estimate only the parameters of those two LLMs which are newly generated by the division. For example, when in iteration 3 in Fig. 3 the LLM 3-2 is divided into LLM 4-2 and 4-3 the LLMs 3-1 and 3-3 can be directly passed to the LLMs 4-1 and 4-3 in the next iteration without any estimation.

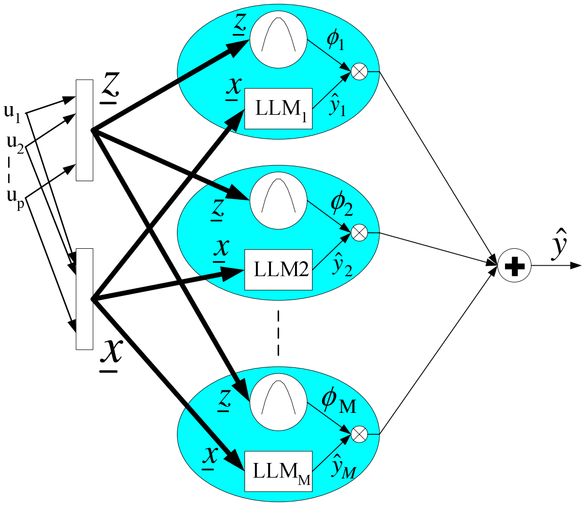


Fig. 2. Network structure of a local linear neurofuzzy model with *M* neurons for *nx* LLM inputs *x* and *nz* validity function inputs *z*

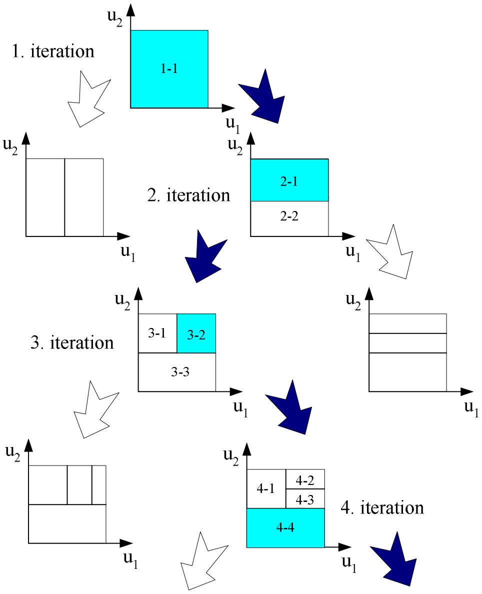


Fig. 3. Operation of the LOLIMOT structure search algorithm in the first four iterations for a two-dimensional input space ($p = 2$)

Using the above strategy, a locally linear model is adopted to the system for input data as shown in fig. 4. The identified and actual outputs can be seen in Fig. 5. As it can be seen, the error between these two values is not considerable and the identified model can match the system well. In the next section, we will use of this model in predictive control block.

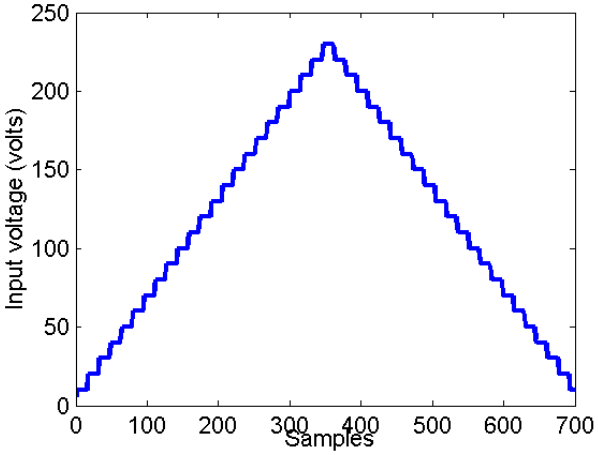


Fig. 4. Input voltage for system identification

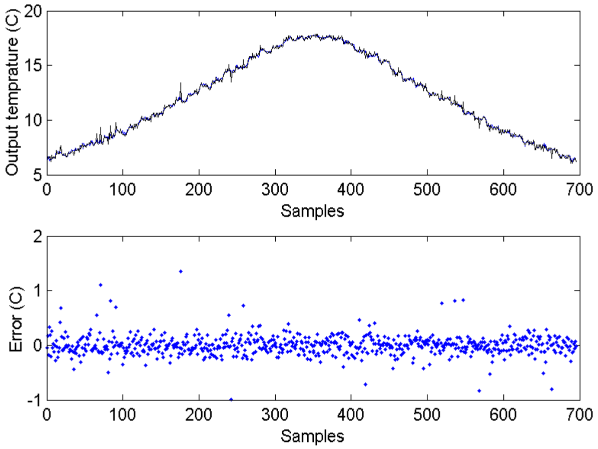


Fig. 5. The identified and actual output temperature with error between theses two values

4 Predictive Controller Design

The objective of the predictive control strategy using LOLIMOT predictors is twofold: *(i)* to estimate the *future output* of the plant and *(ii)* to minimize a *cost* *function* based on the error between the predicted output of the processes and the reference trajectory. The cost function, which may be different from case to case, is minimized in order to obtain the optimum control input that is applied to the nonlinear plant. In most of the predictive control algorithms a quadratic form is utilized for the cost function:

$$J=\sum_{i=N_1}^{N_2}\left[y(k+i)-r(k+i)\right]^2+\lambda\sum_{i=1}^{N_u}\Delta u^2(k+i-1) \tag{6}$$

with the following requirements

$$\Delta u(k+i-1)=0 \qquad 1\le N_u<i\le N_2 \tag{7}$$

where N_u is the control horizon, N_1 and N_2 are the minimum and maximum prediction horizons respectively, i is the order of the predictor, r is the reference trajectory, λis the weight factor, and Δ is the differentiation operator.

The command u may be subject to amplitude constraints:

$$u_{\min}\le u(k+i)\le u_{\max} \qquad i=1,2,\ldots,N_u \tag{8}$$

The cost function is often used with the weight factor λ=0. A very important parameter in the predictive control strategy is the control horizon N_u, which specifies the instant time, since when the output of the controller should be kept at a constant value.

The output sequence of the optimal controller is obtained over the prediction horizon by minimizing the cost function J with respect to the vector U. This can be achieved by setting

$$\frac{\partial J}{\partial U}=0 \qquad U=\left[u(k-d),\ldots,u(k-d+N_u-1)\right]^T \tag{9}$$

However, when proceeding further with the calculation of $\partial J/\partial U$, a major inconvenience occurs. The *analytical approach* to the optimization problem needs for the differentiation of the cost function and, finally, leads to a nonlinear algebraic equation; unfortunately this equation cannot be solved by any analytic procedure. This is why a *computational method* is preferred for the minimization of the cost function, also complying with the typical requirements of the real-time implementations (guaranteed convergence, at least to a sub-optimal solution, within a given time interval).

The advantage of this nonlinear neural predictive controller consists in the implementation method that solves the key problems of the nonlinear MPC. The implementation is robust, easy to use and fulfills the requirements imposed for the minimization algorithm. Changes in the parameters of the neural predictive controller (such as the prediction horizons, the control horizon, as well as the necessary constraints) are straightforward operations. A simple block diagram of predictive control strategy is depicted in Fig. 6.

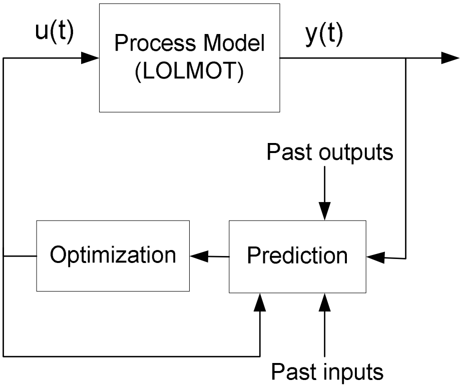


Fig. 6. The scheme of predictive control

The closed-loop system response using the predictive control algorithm based on LOLIMOT model is shown in Fig. 7. In order to investigate the performance of the predictive controller, we will provide another simulation using conventional PID controller. Using trail and error algorithm, the best values for PID controller in which the closed-loop system is stable and has almost satisfactory performance is adopted as follows

$$\text{PID Controller} \Rightarrow k_p + \frac{k_i}{s} + k_d s$$

$$k_p = 1.24 \quad k_i = 2.27 \quad k_d = 0.17$$

The closed-loop system response using PID controller with above parameters is shown in Fig. 8. Comparing Fig. 7 with Fig. 8, we can see that the performance of the system using the predictive controller is much better than that of PID one.

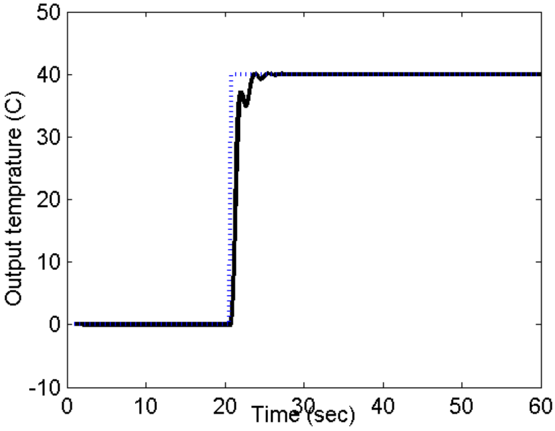


Fig. 7. Closed-loop system response using proposed predictive controller

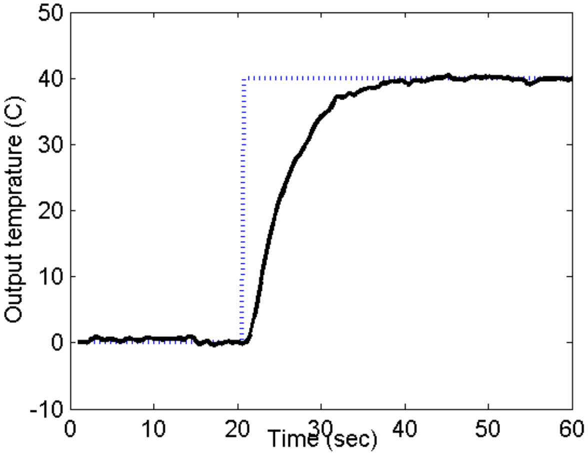


Fig. 8. Closed-loop system response using PID

5 Conclusion

In this paper, a predictive Controller (BELBIC) was applied to electrically heated micro heat exchanger, which is a highly nonlinear plant. To this end, the dynamics of the system was identified using Locally Linear Model Tree (LOLIMOT) algorithm. Then, a controller based on predictive strategy was applied to the system to tackle the output temperature tracking problem. The closed-loop system performance using the proposed predictive controller was compared with that of PID one, which the result of predictive controller was much better than that of PID controller.

References

1. Camacho, E.F.: Model Predictive Control, Springer Verlag (1998)
2. Garcia, C.E., Prett, D.M., Morari, M.: Model Predictive Control: Theory and Practice- a Survey, Automatica 25 (3) (1989) 335-348
3. Badgwell, A.B., Qin, S.J.: Review of Nonlinear Model Predictive Control Applications, In Nonlinear predictive control theory and practice, Kouvaritakis, B, Cannon, M (Eds.), IEE Control Series (2001) 3-32
4. Parker, R.S., Gatzke E.P., Mahadevan, R., Meadows, E.S., Doyle, F.J.: Nonlinear Model Predictive Control: Issues and Applications, In Nonlinear predictive control theory and practice, Kouvaritakis, B, Cannon, M (Eds.), IEE Control Series (2001) 34-57
5. Babuska, R., Botto, M.A., Costa, J.S.D., and Verbruggen, H.B.: Neural and Fuzzy Modeling on Nonlinear Predictive Control, A Comparison Study, Computatioinal Engineering in Systems Science (1996)
6. Nelles, O.: Nonlinear System Identification: from Classical Approach to Neuro-fuzzy Identification, Springer Verlag (2001)

7. Narendra, K. S., and Parthasarathy, K.: Identification and Control of Dynamic Systems Using Neural Networks. IEEE Transactions on Neural Networks 1 (1990) 4–27
8. Jalili-Kharaajoo, M. and Araabi, B.N.: Neural Network Control of a Heat Exchanger Pilot Plant, to appear in IU Journal of Electrical and Electronics Engineering (2004)
9. Brander, J., Fichtner, M., Schygulla, U. and Schubert, K.: Improving the Efficiency of Micro Heat Exchangers and Reactors. In Irven, R. [Hrsg.] Micro reaction Technology: 4th International Conference; AIChE Spring Nat. Meeting, Atlanta, Ga., March 5-9 (2000)
10. Stief, T., Langer, O.U. and Schuber, K.: Numerical Investigations on Optimal Heat Conductivity in Micro Heat Exchangers. In Irven, R. [Hrsg.] Micro reaction Technology: 4th International Conference; AIChE Spring Meeting, Atlanta, Ga., March 5-9 (2000)
11. Nelles, O.: Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models. Springer (2001)
12. Nelles, O.: Local Linear Model Tree for on-line Identification of Time Variant Nonlinear Dynamic Systems. Proc. International Conference on Artificial Neural Networks (ICANN), Bochum, Germany (1996) 115-120

A Group Based Insert Manner for Storing Enormous Data Rapidly in Intelligent Transportation System

Young Jin Jung and Keun Ho Ryu

Database, Bioinformatics Laboratory, Chungbuk National University, Korea
{yjjeong, khryu}@dblab.chungbuk.ac.kr

Abstract. The flood of data is occurred in ITS(intelligent transportation system) according to the progress of wireless telecommunication and sensor network. To deal with large data and provide suitable services smoothly, it is necessary for an index technique to store and search bulk object data rapidly. However the existing indices require a lot of costs to insert a huge amount of data because they store every position data into the index directly. To solve this problem in this paper, we propose a buffer node operation and design a GU-tree(Group Update tree). The proposed buffer node manner reduces the input cost effectively since it stores the moving object location in a group. And then we confirm the effect of the buffer node manner which reduces the insert cost and increases the search performance in a time slice query from the experiment to compare the operation with some existing indices.

Keywords: Moving Objects, Moving Object Index, Index Manager, Buffer Node, Storage Management.

1 Introduction

To recognize positions and maps has made much progress with histories through travel, exploration, transportation, war and so on. Nowadays, it is actively researched to deal with and utilize positions of objects relying on the progress of location management technologies and a wireless communication network, the miniaturization of terminal devices. Specially, it is becoming important issues to track the objects such as a vehicle and an aircraft and to provide suitable services to mobile users depending on changed locations[1].

It is very essential to deal with moving objects are spatial ones changing their locations over time[2, 3]. And the quality of the services closely depends on the locations in mobile environment. It is necessary to manage an amount of location data, since the location information has been increasing massively over time. Therefore the DBMS requires the index technique to effectively search the plentiful spatial elements about moving objects. And it is actively continued to research the indices of moving objects[2].

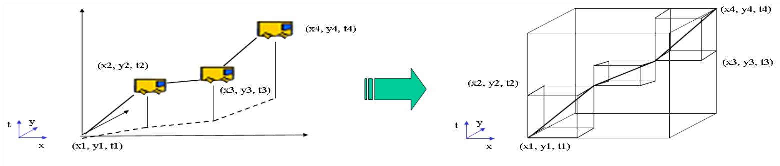
Most of moving object indices based on the R-tree[4], height balanced tree extended from B-tree, are proposed frequently[5]. To reduce not only search cost but also input cost is considerable, because data insertion is more frequent than data search in wireless telecommunication applications to transmit moving object

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 296 – 305, 2005.

information endlessly. Two methods are used to insert data in moving object indices briefly. One is top-down approach used generally, and the other is bottom-up approach to reduce insertion cost as utilizing a hash table. However the bottom-up manner requires to maintain the hash table and its search performance is slower than the top-down manner[6]. In this previous works, a moving object index requires a lot of insert costs, because the insertion is occurred frequently whenever the data is transmitted. To solve this problem in this paper, we propose a GU-tree(Group Update tree) to insert data in a group and design a index manager to improve the performance of the mobile data management system.

2 Related Works

Moving objects changing their spatial information continuously are divided into moving points considering only locations and moving regions containing shapes and positions[7]. In order to provide appropriately services depending on location, the management system should have the index techniques to search the information of an object rapidly as changing their spatial position or shape over the flow of time and as increasing the volume of data[2]. Existing indices which handle continuously moving object location are roughly divided into two types according to their fields of applications[8]. One is dealing with moving object's history and trajectory and the other is handling current position and its position in the near future. Our researches focus on the indices to deal with the history and trajectory information of moving objects.



< The trajectory of a vehicle >   < The trajectory in the R-tree >

Fig. 1. Moving object trajectory representation with the R-tree

Most of moving object indices are based on the R-tree and handle the trajectory information of moving objects with line segment split like Figure 1. The insert manners of existing indices are roughly distinguished into two manners such as top-down and bottom-up. The indices using the top-down approach are STR-tree(Spatio-Temporal R-tree)[9], TB-tree(Trajectory Bundle Tree)[8], TPR-tree(Time Parameterized R-tree)[10], MP-tree(Moving Point tree)[11], and so on. The STR-tree considers both spatial attributes and the trajectory protection of moving objects, but search performance is not good. The TB-tree considering the only trajectory as connecting nodes included in same trajectory. The TPR-tree reduces frequent updates as utilizing a time parameterized bounding with the function of time. The MP-tree improves the search performance as using the projection storage.

The bottom-up manner could fast insert and update the data as employing a hash table. For example, the indices using the bottom-up method are the LUR-tree[5], the bottom-up approach[6], and the TB*-tree[12]. The LUR-tree reduce the insert cost as extending a MBR according to inserted data but its search performance is worse than that of the R-tree. And it needs some storage space such as the hash table, the pointers to update the boundaries of parent nodes and also requires the more cost to index maintenance.

The previous works of moving objects indices deal with data as the leaf node presenting a line segment as inserting and updating data. Consequently, it makes increase the cost for process and maintenance. Besides, although there are many researches about new index method in various environments, there is no system utilizing index manager to improve the search performance of the index and to keep the safety of data. In order to solve this problem, So, we proposes not only the group based storing method to decrease the insert cost but also the index manager for handling continuous data and keeping good search speed in this paper.

3 Mobile Data Management System

Our researches focus on dealing with the history and current information under the assumption that moving object data are continuously transmitted to a mobile object management server over time. And we consider the only moving point's movement, not moving region. This section describes the structure of the mobile data management system and its index manner, index manager, and query processor.

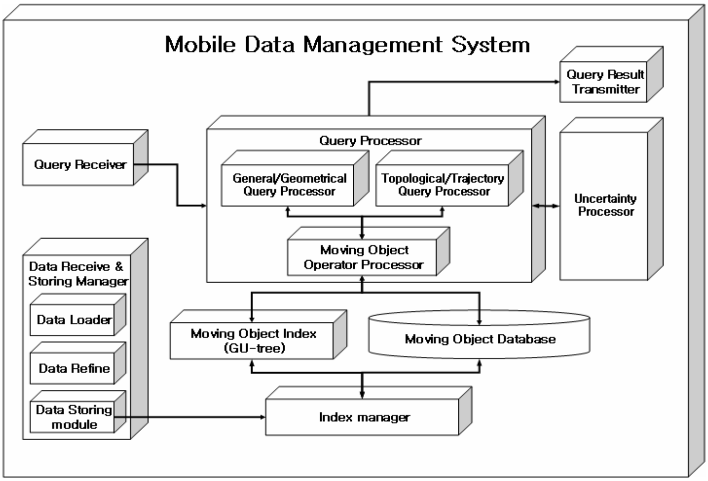
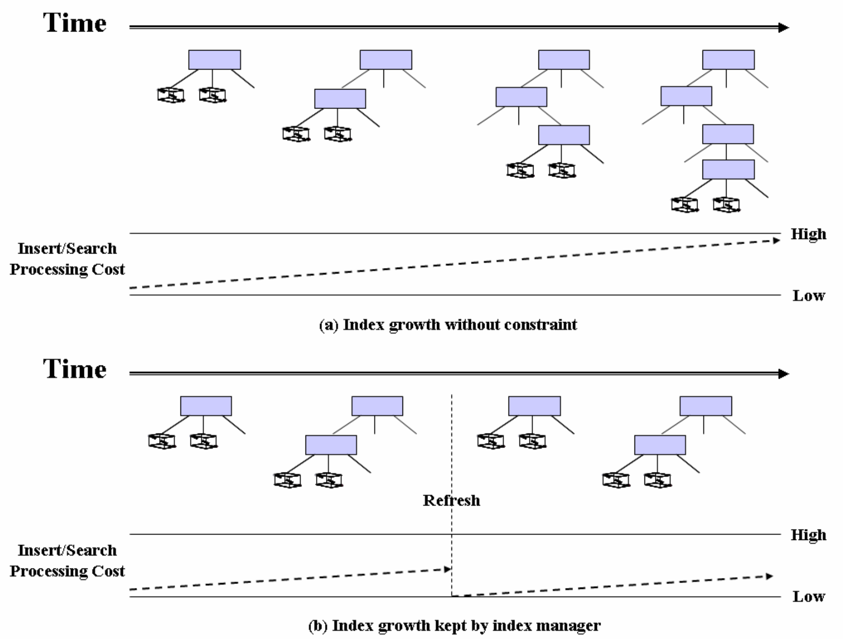

Mobile Data Management System
Query Result Transmitter
Query Receiver
Query Processor
General/Geometrical Query Processor
Topological/Trajectory Query Processor
Moving Object Operator Processor
Uncertainty Processor
Data Receive & Storing Manager
Data Loader
Data Refine
Data Storing module
Moving Object Index (GU-tree)
Moving Object Database
Index manager

Fig. 2. The structure of the mobile data management system

Figure 2 shows the structure of mobile data management system. The system consists of data storing manager, query processor, index manager, moving object index and so on. The data storing manager receives and refines the transmitted location data from moving vehicles. The index manager stores the location data into the index required to search data rapidly and database to keep the data safely. The index manager keeps the performance of index good through refreshing the index per a specific period such as a day, a week etc. The query processor contains the general and geometrical query processor, the topological and trajectory query processor, the moving object operator, and the uncertainty processor. We focus on data repository management of the system in this paper.

The index manager is required to store the position data into the index, database and to keep the search performance of the index good. It is not recommended to combine the index with query processor without any strategy to maintain the index condition in the system. It causes to worsen the search performance of the index, because the index grows up over time. So, it is required to remove and refresh the data in the index. The history of data is maintained, since the back up version of removed data in the index is remained in the database.

Figure 3 shows the process of index maintenance using index manager. The performance of data search is worsened depending on the index growth. The more data inserts, the more cost increases. So, refreshing the index is required whenever the cost of insert and search process is too high to provide a location based services quickly. To refresh the index, we consider two elements - the size of data and time period - to decide standard for updating the index. Considering the data size is updating the index when the volume of the data records approaches a specific size. The other is refreshing the index by maintenance periods. In the result, it is selected to update the index according to a time period for repairing the index automatically as occurring errors like uncertainty boundary of data in the index.



(a) Index growth without constraint

(b) Index growth kept by index manager

Fig. 3. The moving object index management

4 Moving Object Index Using Group Update

In the existing indices of moving objects, whenever a line segment describing the movement of mobile objects is transmitted, it is inserted into the index immediately. Therefore, inserting and updating cost are very high. Moreover, the index structure needs to be maintained in the balance state frequently. In order to solve the problem, the proposed buffer node insert manner use a non-leaf node instead of a leaf node as an insert unit like Figure 3 as inserting data into an index.

The proposed GU-tree is an index to handle moving points to reduce a frequent insert cost and to process a time slice and time range queries effectively as utilizing the buffer node manner in existing indices. The proposed GU-tree consists of leaf node, non-leaf node, and the buffer node to store the set of leaf nodes heaped.

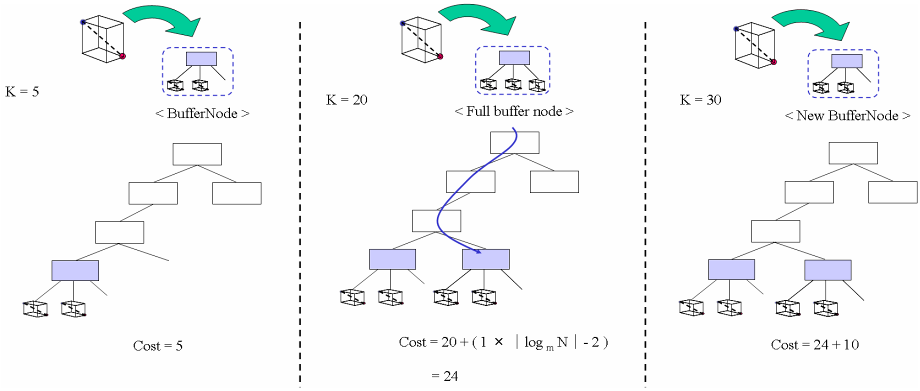


Fig. 4. The buffer node insert manner to insert data group instead of each data

Figure 4 describes the buffer node manner to insert data in groups. The buffer node heaps up leaf nodes until filling the buffer node with leaf nodes. And the full buffer node is inserted into an index. Therefore the number of inserted buffer nodes is reduced to at most K / M + M (∵ K : all inserted vehicle records, M : maximum number of entries in a node). The GU-tree growing in the direction of time flow starts to process queries by checking the time condition of queries first, in order to remove unnecessary candidates within the relatively wider range of time.

The insert process of the top-down manner is finished as reaching a leaf node level in an index. However the process of the buffer node manner is completed as arriving in the parent node of a leaf node. Therefore when the equal numbers of records are inserted into the indices using two manners, the number of node access in the insert process of the index using buffer node is fewer than the other. The GU-tree growing in the direction of time flow starts to process queries by checking the time condition of queries first, in order to remove unnecessary candidates within the relatively wider range of time.

```
Algorithm Insert_buffer node(class node *root, class node *entry)
Input : root  // Node of a Tree
          entry // Information of moving objects
Method :
if BufferNode is null then        // BufferNode is global variable pointer
        make new BufferNode
else
        if the entry's boundary is beyond the BufferNode 's boundary then
                modify BufferNode 's boundary with the entry's boundary
        endif
        if the BufferNode 's count < M then
                insert the entry into BufferNode
        else
                insert the entry into BufferNode        // BufferNode is full
                insert_buffer node_into_tree(root, BufferNode)
                make BufferNode null
        endif
endif
End
```

Algorithm 1. The buffer node insert algorithm

Algorithm 1 shows the buffer node insert process in the GU-tree. When a node in the GU-tree is full, the index does not use split operation, but makes a new node to insert data.

The insert process of the top-down manner is finished as reaching a leaf node level in an index. However the process of the buffer node manner is completed as arriving in the parent node of a leaf node. Therefore when the equal numbers of records are inserted into the indices using two manners, the number of node access in the insert process of the index using buffer node is fewer than the other. So, the insert cost could be calculated under the assumption that K records are inserted into the index containing N index records. The merit of buffer node is illustrated with calculated formula and virtual values in below table.

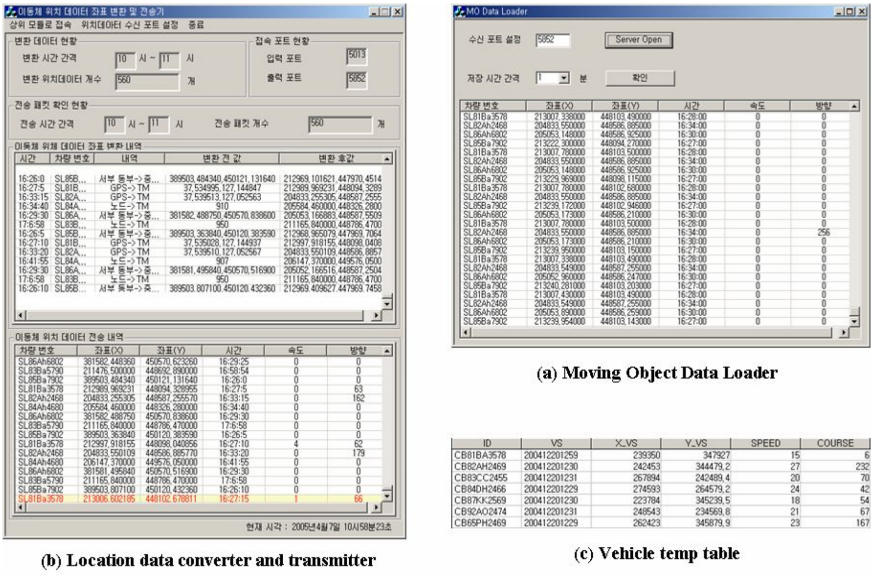
Table 1. The insert cost comparison between the top-down and buffer node manner

Insert manner	Cost analysis
Assumption	$K = 500{,}000,\ m = 20,\ M = 40,\ N = 64{,}000{,}000$ $\lvert \log_m N \rvert - 1 = \lvert \log_{20} 64000000 \rvert - 1 = \lvert \log_{20} 20^6 \rvert - 1 = 5$
Top down	Total insert cost $= K \times (\lvert \log_m N \rvert - 1) = 500000 \times 5 = 2{,}500{,}000$
Buffer node	Total insert cost $= K + \{ K/M \times (\lvert \log_m N \rvert -2) \}$ $= 500000 + (500000/40) \times (6-2) = 550{,}000$
Difference between top-down and buffer node	Cost difference $= (K - K/M) \times (\lvert \log_m N \rvert -2)$ $= \{500000 - (500000/40)\} \times (6-2) = 1{,}950{,}000$

The effect of buffer node is confirmed in the difference between two indices in the table 1 with sample values. The more K and M increase, the better the effect of the buffer node manner gets. In other words, if K and M increase, the effect of an index using the buffer node will be also better. The performance of the buffer node insert manner is also better in the comparison with the bottom-up approach in section 6.

5 Implementation

In this chapter, we describe the process dealing with the data of moving vehicles in the data storing management parts of the system such as an index manager, a moving object index, and a moving object data loader, etc. The locations are processed and stored in the oracle and access database in the order of vehicle position data → a data collector → a data converter and transmitter → a MO data loader → a database and an index.


MO Data Loader
Server Open
확인
이동체 위치 데이터 좌표 변환 및 전송기
상위 모듈로 접속
위치데이터 수신 포트 설정
종료

수신 포트 설정 5852

저장 시간 간격 1 분

차량 번호	좌표(X)	좌표(Y)	시간	속도	방향
SL81Ba3578	213007,338000	448103,490000	16:28:00	0	0
SL82Ah2468	204833,550000	448586,885000	16:34:00	0	0
SL86Ah6802	205053,148000	448586,925000	16:30:00	0	0
SL85Ba7902	213222,300000	448094,270000	16:27:00	0	0
SL81Ba3578	213007,780000	448103,500000	16:28:00	0	0
SL82Ah2468	204833,550000	448586,885000	16:34:00	0	0
SL86Ah6802	205053,148000	448586,925000	16:30:00	0	0
SL85Ba7902	213229,969000	448098,115000	16:27:00	0	0
SL81Ba3578	213007,780000	448102,680000	16:28:00	0	0
SL82Ah2468	204833,550000	448586,885000	16:34:00	0	0
SL85Ba7902	213239,172000	448102,946000	16:27:00	0	0
SL86Ah6802	205053,173000	448586,210000	16:30:00	0	0
SL81Ba3578	213007,780000	448103,500000	16:28:00	0	0
SL82Ah2468	204833,550000	448586,885000	16:34:00	0	256
SL86Ah6802	205053,173000	448586,210000	16:30:00	0	0
SL85Ba7902	213239,950000	448103,150000	16:27:00	0	0
SL81Ba3578	213007,338000	448103,490000	16:28:00	0	0
SL82Ah2468	204833,549000	448587,255000	16:34:00	0	0
SL86Ah6802	205052,960000	448586,247000	16:30:00	0	0
SL85Ba7902	213240,281000	448103,203000	16:27:00	0	0
SL81Ba3578	213007,430000	448103,490000	16:28:00	0	0
SL82Ah2468	204833,549000	448587,255000	16:34:00	0	0
SL86Ah6802	205053,890000	448586,259000	16:30:00	0	0
SL85Ba7902	213239,954000	448103,143000	16:27:00	0	0

(a) Moving Object Data Loader

변환 데이터 현황

변환 시간 간격 10 시 ~ 11 시

변환 위치데이터 개수 560 개

접속 포트 현황

입력 포트 5013

출력 포트 5852

전송 패킷 확인 현황

전송 시간 간격 10 시 ~ 11 시 전송 패킷 개수 560 개

이동체 위체 데이터 좌표 변환 내역

시간	차량 번호	내역	변환 전 값	변환 후값
16:26:0	SL85B...	서부 동부->중...	389503,484340,450121,131640	212969,101621,447970,4514
16:27:5	SL81B...	GPS->TM	37,534995,127,144847	212989,969231,448094,3289
16:33:15	SL82A...	GPS->TM	37,539513,127,052563	204833,255305,448587,2555
16:34:40	SL84A...	노드->TM	910	205584,460000,448326,2800
16:29:30	SL86A...	서부 동부->중...	381582,488750,450570,838600	205053,166883,448587,5509
17:6:58	SL83B...	노드->TM	950	211165,840000,448786,4700
16:26:5	SL85B...	서부 동부->중...	389503,363840,450120,383590	212968,965079,447969,7064
16:27:10	SL81B...	GPS->TM	37,535028,127,144937	212997,918155,448098,0408
16:33:20	SL82A...	GPS->TM	37,539510,127,052567	204833,550109,448586,8857
16:41:55	SL84A...	노드->TM	907	206147,370000,449576,0500
16:29:30	SL86A...	서부 동부->중...	381581,495840,450570,516900	205052,166516,448587,2504
17:6:58	SL83B...	노드->TM	950	211165,840000,448786,4700
16:26:10	SL85B...	서부 동부->중...	389503,807100,450120,432360	212969,409627,447969,7458

이동체 위치 데이터 전송 내역

차량 번호	좌표(X)	좌표(Y)	시간	속도	방향
SL86Ah6802	381582,448360	450570,623260	16:29:25	0	0
SL83Ba5790	211476,500000	448692,890000	16:58:54	0	0
SL85Ba7902	389503,484340	450121,131640	16:26:0	0	0
SL81Ba3578	212989,969231	448094,328955	16:27:5	0	63
SL82Ah2468	204833,255305	448587,255570	16:33:15	0	162
SL84Ah4680	205584,460000	448326,280000	16:34:40	0	0
SL86Ah6802	381582,488750	450570,838600	16:29:30	0	0
SL83Ba5790	211165,840000	448786,470000	17:6:58	0	0
SL85Ba7902	389503,363840	450120,383590	16:26:5	0	0
SL81Ba3578	212997,918155	448098,040856	16:27:10	4	62
SL82Ah2468	204833,550109	448586,885770	16:33:20	0	179
SL84Ah4680	206147,370000	449576,050000	16:41:55	0	0
SL86Ah6802	381581,495840	450570,516900	16:29:30	0	0
SL83Ba5790	211165,840000	448786,470000	17:6:58	0	0
SL85Ba7902	389503,807100	450120,432360	16:26:10	0	0
SL81Ba3578	213006,602185	448102,678811	16:27:15	1	66

현재 시각 : 2005년4월7일 10시58분23초

(b) Location data converter and transmitter

ID	VS	X_VS	Y_VS	SPEED	COURSE
CB81BA3578	200412201259	239350	347927	15	6
CB82AH2469	200412201230	242453	344479,2	27	232
CB83CC2455	200412201231	267894	242489,4	20	70
CB84DH2466	200412201229	274593	264579,2	24	42
CB87KK2569	200412201230	223784	345239,5	18	54
CB92AO2474	200412201231	248543	234569,8	21	67
CB65PH2469	200412201229	262423	345879,9	23	167

(c) Vehicle temp table

Fig. 5. Vehicle location data converted to transmit to the server

Figure 5 shows the transmission and storing process of vehicle data in the moving object data loader in the system. The modified vehicles location depending on infrastructure such as roads, buildings is transmitted into the moving object data loader(a) in the system. The loader can store the location per a specific period such as a minute, 10 minutes. The last vehicle data in database(c) can be used to predict positions in near future or to search current positions.

6 Performance Studies

The spatial objects such as national borders, lakes, roads and parks can be obtained from the public web. However moving objects such as a ship or a vehicle have no publicized actual data, therefore in the tests for application fields related to moving objects, researchers commonly use the data generator such as the GSTD(Generator of Spatio-Temporal Dataset)[13, 14], City Simulator[15], and Oporto[16], and so on. So, the data generator used in this test is similar to the GSTD, and it can create the positions and the speed of moving point data per 10 minutes randomly. All used indices deal with history and current moving object data under the assumption that space is limited and time is infinite. The number of node access is checked according to the number of moving objects, changed spatial and temporal scope. In order to analyze the insert cost, the node access number is checked with 100,000 ~ 1,000,000 records. The used indices are GU-T, GU-B, R-T, and R-B to describe GU-tree(top-down), the GU-tree(bottom-up), the R-tree(top-down), and the R-tree(bottom-up) in the figures.

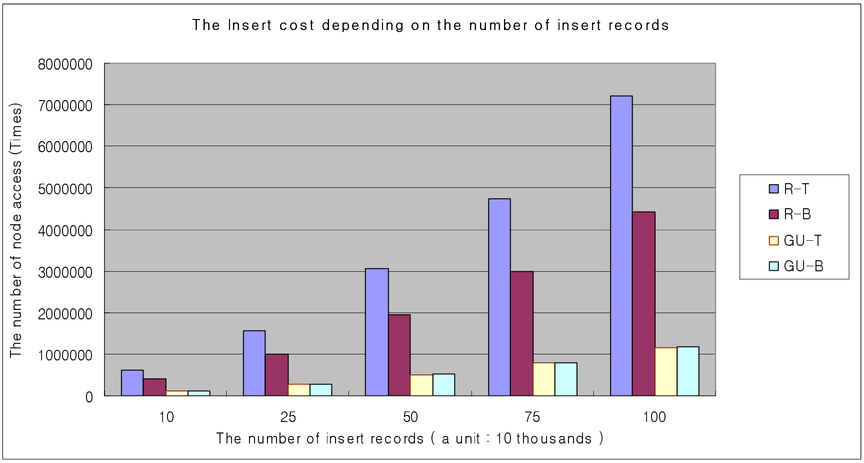


Fig. 6. The insert cost analysis of indices

The insert cost grows according to increased insert records in Figure 6. The R-B is better than the R-T, because the bottom-up manner can insert records into a leaf node using a hash table, and the update cost to modify the boundaries of parent nodes can be fewer than the top-down. The GU-tree utilizing the buffer node shows better performance than the R-tree in top-down and bottom-up manner. Figure 6 proves the advantage of the buffer node manner to insert records into an index as a group of leaf nodes, not as a leaf node. The node approach number is GU-T ≒ GU-B < R-B < R-T.

The search performance of the GU-tree is also analyzed, because there is a trade-off like the LUR-tree. The LUR-tree shows good insert performance, but its search performance is worse than the R-tree.

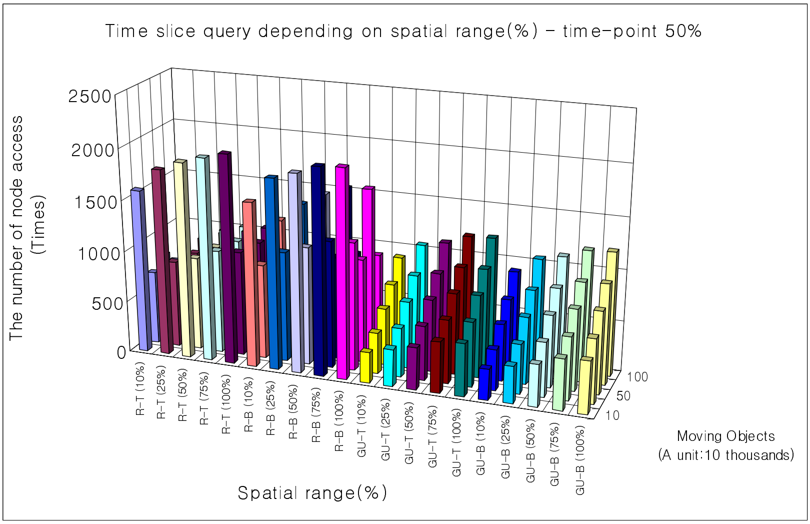


Fig. 7. The search cost for the time slice query depending on changed spatial range

Figure 7 shows the node access number to process the time slice query depending on the changed spatial range such as 10%, 25%, 50%, 75%, 100% in entire spatial area. We can recognize that the GU-tree has node access number fewer than existing top-down approaches like the formulas in section 4 from experimental results tested. The more spatial range widens, the more the cost increases. The GU-tree is also better than the R-tree in the time slice query changing spatial range. The number of node access containing the changed spatial range is GU-T ≒ GU-B < R-T ≒ R-B.

In most test to search data, the more spatial and temporal range widens, the more the cost of indices becomes similar. In addition, this buffer node manner is useful in a moving object management system to insert data frequently. However, it is needed to study for applying to the indices dealing with the current and future locations

7 Conclusion

A moving object management system requires effective data insert and update techniques, because data insert requirement is more frequent than data search requirement. To insert and update data, the existing moving object indices use the top-down and the bottom-up manner roughly. However both methods have the weak point in the regard of huge insert cost, because they insert data records as a unit presenting a line segment. Therefore to solve the problem, the buffer node insert manner is proposed to store records into an index as a unit grouping records. To improve the search performance of the index and the data utilization in database, the index manager refreshing the index is utilized. We explained data management process in storage parts and proved the considerable reduction of the insert and search cost in index performance comparison between the GU-tree and the R-tree in section 6.

Future work is to study various applications of the buffer node, to extend the index manager to the stream data database to consider some storage level, and to improve the performance of the systems in a variety of environments.

Acknowledgements

This work was supported by the RRC program of MOCIE and ITEP in Korea.

References

1. Reed, J.H., Krizman, K.J., Woerner, B.D., Rappaport, T.S.: An Overview of the Challenges and Progress in Meeting the E-911 Requirement for Location Service. IEEE Communication Magazine (1998) 33-37
2. Mokbel, M.F., Ghanem, T.M., Aref, W.G.: Spatio-temporal Access Methods. IEEE Data Engineering Bulletin, Vol. 26, No. 2, (2003) 40-49
3. Guting, R.H., Bohlen, M.H., Erwig, .M., Jensen, C.S., Lorentzos, N.A., Schneider, M., Vazirgiannis, M.: A Foundation for Representing and Querying Moving Objects. ACM Transactions on Database Systems, Vol. 25, No. 1, (2000) 1-42
4. Guttman, A.: A.:R-trees: a Dynamic Index Structure for Spatial Searching. ACM-SIGMOD, (1984) 47-57
5. Lee, M.L., Hsu, W., Jensen, C.S., Cui, B., Teo K.L.: Supporting Frequent Updates in R-Trees: A Bottom-Up Approach. VLDB, (2003) 608-619
6. Kwon, D.S., Lee, S.J., Lee, S.H.: Indexing the Current Positions of Moving Objects Using the Lazy Update R-Tree. Mobile Data Management (2002) 113-120
7. Forlizzi, L., Guting, R.H., Nardelli, E., Schneider, M.: A Data Model and Data Structures for Moving Objects Databases. ACM SIGMOD (2000) 319-330
8. Pfoser, D., Theodoridis, Y., Jensen, C.S.: Indexing Trajectories of Moving Point Objects. CHOROCHRONOS TECHNICAL REPORT CH-99-03 (1999)
9. Pfoser, D., Jensen, C.S, Theodoridis, Y.: Novel Approaches in Query Processing for Moving Objects. CHOROCHRONOS TECHNICAL REPORT CH-00-03 (2000)
10. Saltenis, S., Jensen, C., Leutenegger, S., Lopez, M.: Indexing the Positions of Continuously Moving Objects. ACM-SIGMOD (2000) 331-342
11. Jung, Y.J., Lee, E.J., Ryu, K.H.: MP-tree : An Index Approach for Moving Objects in Mobile Environment. ASGIS (2003) 104-111
12. Lee, E.J., Ryu, K.H., Nam, K.W.: Indexing for Efficient Managing Current and Past Trajectory of Moving Object. Apweb, Hangzhou (2004) 781-787
13. Theodoridis, Y., Nascimento M.A.: Generating Spatiotemporal Datasets. SIGMOD Record, Vol. 29, No. 3, (2000) 39-43
14. Pfoser, D., Jensen, C.S.: Querying the Trajectories of On-Line Mobile Objects. CHOROCHRONOS TECHNICAL REPORT CH-00-57 (2000)
15. Brinkhoff, T.: Generating Traffic Data. IEEE Data Engineering Bulletin, Vol. 26, No. 2, (2003) 19-25
16. Saglio, J.M., Moreira, J.: Oporto: A Realistic Scenario Generator for Moving Objects. DEXA Workshop (1999) 426-432

A RDF-Based Context Filtering System in Pervasive Environment*

Xin Lin[1], Shanping Li[1], Jian Xu[2], and Wei Shi[1]

[1] College of Computer Science, Zhejiang University, Hangzhou, P.R.China 310027
[2] College of Computer Science, Hangzhou Dianzi University,
Hangzhou, P.R.China 310000
{alexlinxin, cnxujian}@hotmail.com, shan@cs.zju.edu.cn,
shiwei@zj165.com

Abstract. In pervasive computing community, there is a high interest on context-aware computing. Much work focuses on context reasoning, which deals with high-level abstraction and inference of pervasive contextual information and several prototype systems have been proposed. However, overwhelming contextual information in pervasive computing makes these systems inefficient, even useless. In this paper, we propose an application-oriented context filtering system (ACMR) to deal with above problem. To prevent the pervasive applications from being distracted by trashy contexts, ACMR system only deals with application-related contextual information rather than all the available contextual information. Experiments about ACMR system demonstrate its higher performance than those of previous systems.

1 Introduction

Mark Weiser envisioned that the computing environments of 21st Century are so pervaded with computing devices [1]. In his vision, computing entities will disappear into background so that people can focus on their daily tasks rather than the underlying technologies. To accomplish such invisibility, applications in pervasive computing should be aware of context, which is defined as any information that can be used to characterize the situation of the environment [2], to reduce inputs from users. As a result, it is widely acknowledged that context-awareness is an important feature in pervasive computing.

There have appeared several context modeling and reasoning systems, but these systems take in all the raw contextual information without classification. Such huge amount of contextual information renders the modeling and reasoning process inefficient, even useless. To deal with this difficulty, we propose an application-oriented context filtering system (ACMR) in this paper. The main idea of ACMR is that the system only deals with the application-related contextual information, rather than all the available contextual information. For example, Tom's mobile phone only cares about the contextual information of Tom not that of Jerry. In ACMR system, a con-

* This work is supported by National Natural Science Foundation of China (No. 60473052).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 306 – 315, 2005.

text filter is designed to discard the trashy contextual information. The Concerned Value (CV) algorithm is proposed to evaluate the importance of specific contexts by gathering the information from pervasive applications and the context reasoner. Experiments about ACMR system demonstrate its higher performance than those of previous systems.

The rest of this paper is organized as follows: Section 2 compares ACMR system with related work. The overview of our system is depicted in Section 3. Section 4 presents the detail of AMCR system. We evaluate the performance in Section 5 and the conclusions are drawn in Section 6.

2 Related Work

Much effort has been put into the context-awareness computing. The first research investigating context-aware computing was the Olivetti Active Badge system [4] and on its footsteps a lot of projects emerged. Dey et.al. developed Context Toolkit [5] and represented contexts as attribute-value pairs. The Cooltown project [6] considers that everything has a web presence and relies on a web-based context model in which every entity (device, person and location) has a corresponding URL. However, these projects pay more attention to gathering and representing the contexts in pervasive computing. These contexts are mostly raw and context-aware applications cannot consume them directly. As a result, context reasoning intrigues some researchers. In this area, works focus on how to inference high-level contexts from raw context. H.Wang [10] [9] proposed ontology-based context modeling and reasoning. Anand [3] use probabilistic and fuzzy logic to deal with reasoning about uncertain contexts in pervasive environment, which is also on basis of ontology. However, their works only focus on how to get high-level contexts from raw contexts. They did not care about how to consume the high-level contexts. Moreover, in their reasoning mechanism, all raw contexts are adopted as the input of reasoner, no matter whether they are useful for applications. Such solutions may impose high overhead on the system. To deal with this problem, we proposed ACMR system, in which applications only subscribe the contexts that they are concerned about from Context Knowledge Base (CKB) and send them to the reasoner as input.

3 Overview of ACMR

3.1 Architecture of ACMR System and Main Data Structure

Fig.1 shows the architecture of ACMR system. The system contains 4 components: context modeler, context knowledge base (CKB), context reasoner and context filter. Next, we will introduce these components:

CKB stores gathered contexts and provides interfaces to context reasoner and context filter. RDF triple [7], whose format is like <subject predicate object>, is imported in ACMR to represent contexts in pervasive environment. E.g, the RDF triple <Tom

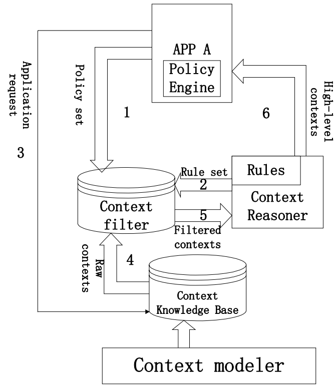


Fig. 1. The architecture of ACMR system

locatedIn RoomA> means "Tom is in the RoomA.". Because the contexts in pervasive computing environment vary with high change rate, the CKB should regularly refresh the contexts stored in it. Moreover, the CKB should check the consistency of these contexts periodically. For example, context "The user A is in room R" and "The user A is in room M" are not consistent because the user A cannot be in different rooms at the same time.

Context modeler is the interface between environmental pervasive devices and ACMR system. By originating form diverse sources, the contextual information presents heterogeneous formats. For example, a system clock shows the time as "Tuesday, March 15, 2005" and a PDA gives the user's schedule as a to-do list. To deal with such heterogeneity, diverse interfaces have been designed in context modeler. As mentioned in the previous subsection, it transforms captured contextual information to RDF triples and sends them to CKB.

Context reasoner infers high-level contexts from raw contexts using defined rule set. An informal description of rule set syntax is shown in Fig.2a:

RuleSet := {Rule}$^+$
Rule := RuleCondition '=>' RuleConclusion
RuleCondition:= RuleCondition ∧ Triple | Triple
RuleConclusion := Triple
Triple : = <subject predicate object>

(a)

type (?u) = Person type(?w) = Washroom
type (?c) = Car
<?u locatedIn ?w> => <?u status UsingWashingroom>
<Person:?u own ?c> ∧ <?c status Running> => < ?u status driving>

(b)

Fig. 2. Syntax of rule set

For the sake of simplicity, the disjunction sign "∨" is not available in the syntax because a rule " A ∨ B => C" can be transformed into two rules: " A =>C" and

"B=>C". The syntax of the rule set is useful for the Concerned Value (CV) algorithm, which will be discussed in Section 4. Note that the elements of the triples in rule set are class-based, while the triples in policy engine and the CKB are instance-based. The difference between class-based triples and instance-based triples is that the elements of the former represent a class while the later only stand for a specific instance. Fig.2b shows an example rule set. In this rule set, ?u stands for all instances of class Person, not a specific instance.

Context filter records the applications' Concerned Value (CV) about specific instance and discards the trashy contexts. Concerned Values, ranging from 0 to 1, map applications to contextual instances. Higher CV denotes the contextual instance is more useful for the application. Each application is assigned a CV table in context filter. An example CV table is shown in Tab.1. In the CV table, we define the entry that is both in row R and column C as Entry (R,C).

Table 1. Concerned value table in context filter

APPA	Subject	Predicate	Object
WashroomA			0.05
Tom	0.07		0.03
locatedIN		0.04	

Note that the CV tables record the applications' concerned value about an instance (the first column of each row) not a context triple. There are 3 columns in each table: subject, predicate and object, which correspond to the three parts of RDF triple respectively. If an instance appears in the different part of a context triple, the application's CV about it may be different. In above example, the first row means the AppA's CV about instance WashroomA is 0.05, only if WashroomA is the object of a context. The AppA are not concerned about WashroomA if it is a subject or predicate of a context. Similarly, the second row means the AppA's CVs about instance Tom are 0.07 or 0.03, if Tom is subject or object of a context respectively. The application' CVs about unlisted contexts instance are 0. In Section 4 we will presents how can we get the CVs and how can we use the CVs.

type (Tom) = Person
<Tom status WatchingTv> => Volume of Rings Turn Up (1)
<Tom status Sleeping> => Rings Shut Off (2)
<Tom status GivingLecture> => Rings Shut off and turn vibration on (3)
<Tom status Driving> $\wedge$ <Tom drives CarA> $\wedge$ <CarA SpeedExceed 60mph> => Forward
Incoming Call To Voice Mail Box (4)

Fig. 3. The policy set of Tom's mobile phone

The above presents the elements of ACMR system. Outside ACMR, we assume each pervasive application contains a ***policy engine***, which help the application choosing a context-awareness policy. We use policy set to represent the information in policy engine. The syntax of policy set is similar with that of rule set in context reasoner. There are two differences between them: (1) the conclusion (right side of deducing sign "=>") of policy is not triple but the action to be adopted;(2) the condition (triple in the left side of "=>") of policy is instance-based. We give partial policy set of Tom's mobile phone in Fig.3.

4 Application-Oriented Context Filtering

The main contribution in ACMR system is Application-Oriented Context Filtering (AOCF) technology, which release the overhead of the context reasoner. In this section, we firstly present the workflow of AOCF. And we will discuss the detail of concerned value algorithm in subsection 4.2.

4.1 Workflow of AOCF

The workflow of AOCF is also illustrated in Fig.1. The order of numbers in the figure depicts the process of our system. When a new application, say AppA, enters the ACMR system, it initially finds out the context filter and sends its policy set to the context filter. Then the context filter gets the rule set from the context reasoner. According to the Concerned Value algorithm, the context filter figures out APPA's concerned value table by collected policy set and rule set. After establishing the Concerned Value table, the initial configuration is finished. If AppA wants to set up context-aware policies, it firstly sends an application request to the CKB. After accepting the request, the CKB send the raw contexts stored in it to the context filter. Then the context filter works out the AppA's Concerned Value about these raw contexts using the established Concerned Value table. The contexts with low concerned value are regarded as trashy and should be discarded. After the filtering, the context filter sends the remaining contexts to the context reasoner as the input of inference. After the inference finished, the context reasoner outputs the high-level contexts and sends them to AppA to consume.

4.2 Concerned Value (CV) Algorithm

The main idea of the CV algorithm is: if a specific application is concerned about a given context, this context or the high-level contexts deduced from it are likely to exist in the policy set of the application. Prior to describing the detail of CV algorithm, we assume a scenario of Tom's mobile phone, which is smart enough to adapt to the user's contexts. For example, if the user is at meeting, the phone shuts off the rings and turns the vibration on; if the user is watching TV, the volume should be turned up. The policy set of this phone has been given in Fig.3 and the rule set is shown in Fig.4.

The CV algorithm can be divided into 2 processes: creating the CV table and filtering contexts. Both of them are carried out in the context filter, which firstly gather the phone's policy set and the rule set of the context reasoner.

type (?u) = person
<?u locatedIn LivingRoom> ∧ <TVSet locatedIn LivingRoom> ∧ <TVSet status ON> => <?u status WatchingTV> (1)
<?u locatedIn Bedroom> ∧ <Bedroom lightLevel Low> ∧ <?u AroundVolume Low> => <?u status Sleeping> (2)
<?u locatedIn ClassRoom> ∧ <?u career Teacher> ∧ <?u status Speaking> => <?u status GivingLecture> (3)
<?u locatedIn Kitchen> ∧ <ElectricOven locatedIn Kitchen> ∧ <ElectricOven status ON> => <?u status Cooking> (4)

Fig. 4. Rule set of our assumed scenario

4.2.1 Creating the CV table

Above all, we introduce Triple-CV set, a temporary data structure used in this process. Each entry of Triple-CV set is a key-value pair, in which the key is instanced-based triple and the value is the given application's CV about this triple. For the sake of simplicity, we represent such entry as (triple, CV (triple)), e.g., (<Tom status Sleeping>, 0.2). Now we describe the process of creating the CV table step by step.

Table 2. Initial Triple-CV set

triple	CV (triple)
<Tom status WatchingTv>	0.25
<Tom status Sleeping>	0.25
<Tom status GivingLecture>	0.25
<Tom status Driving>	0.083
<Tom drives CarA>	0.083

Step 0: Initialization. The triples in the condition of policy set are inserted into Triple-CV set in this step. For the sake of simplicity, the policies in the policy set are assumed to be equally important in this paper. In our future work, we will improve the AOCF mechanism by analyzing the importance of a policy. For a given policy, say $\mathrm{CODITION} \Rightarrow \mathrm{ACTION}$, if there is only one triple T in the CONDITION, the CV (T) is assigned as 1/N, where N is the number of policies. If there is any conjunction sign ("∧") existing in the CONDITION, such as $\mathrm{T}_1 \wedge \mathrm{T}_2 \wedge \ldots\ldots \wedge \mathrm{T}_\mathrm{N} \Rightarrow \mathrm{ACTION}$ each triple in this policy is assigned as: $\mathrm{CV}(\mathrm{T}_1) = \mathrm{CV}(\mathrm{T}_2) = \ldots\ldots = \mathrm{CV}(\mathrm{T}_\mathrm{N}) = 1/((\mathrm{M}+1) * \mathrm{N})$, where M is the number of conjunction sign. Considering the policy set given in Fig.3, CV (<Tom status WatchingTv>) =1/4 and CV (<Tom status Driving>) = CV (<Tom drives CarA>) = 1/(2+1)*4 = 1/12. Tab.2 gives an initial Triple-CV set of our scenario.

Step 1: Comparing the Triple-CV set with rule set. If some triple (say T) in Triple-CV set "matches" a triple (say R) in the conclusion of a rule in the rule set, then system turns to step 2. Otherwise, it goes to step 3. Note that the triples in the Triple-CV set is instance-based, as that in the rule set is class-based. So, here triple T "matches" triple R means T is an instance of R. For example, <Tom status GivingLecture> is an instance of <?u status GivingLecture>, because the type of "Tom" and "?u" are both Person.

Step 2: Expanding the Triple-CV set. As it is detected that triple T matches triple R in step 1, we then produce an instantiation of the whole rule of R and the triples in the condition part of the instantiation are added to the Triple-CV set. Assume the number of conjunction sign in the condition part is M, each new triple's CV is assigned as CV(T) / (1+M). For example, it is detected that <Tom status GivingLecture> is an instance of <?u status GivingLecture>. So we produce an instantiation of rule (3) as follow:

Tom locatedIn ClassRoom> ∧ <Tom career Teacher> ∧ <Tom status Speaking> =>
<Tom status GivingLecture>

The number of conjunction sign "∧" in the condition part is 2. So each triple in the condition is assigned a CV with 0.25 / (2+1). The Triple-CV set is appended with 3 pairs (see Tab.3). After the expanding, the process goes back to step 1 again. To eliminate endless loops, we define triple T doesn't match the triple R any more.

Table 3. Expanded Triple-CV set

triple	CV (triple)
…	…
<Tom locatedIn ClassRoom>	0.083
<Tom career Teacher>	0.083
<Tom status Speaking>	0.083

After the expanding, the process goes back to step 1 again. To eliminate endless loops, we define triple T doesn't match the triple R any more.

Step 3: Transforming the Triple-CV set to CV table. Since it has been proved in step 1 that there is no more match existing between Triple-CV set and rule set, the expanding of Triple-CV set is finished. The final Triple-CV set is shown in Tab.4. The main task in this step is creating the CV table from the final Triple-CV set. As aforementioned, the triples in Triple-CV set is composed of three instances, which are subject, predicate and object respectively. We firstly transform the Triple-CV set pair into Instance-CV triple (instance, part, CV (instance, part)), where part can be subject, predicate or object The CV (instance, part) is assigned as CV (triple) /3. For example, the pair (<Tom career Teacher>, 0.083) is decomposed to three Instance-CV triples: (Tom, subject, 0.028), (career, predicate, 0.028), and (Teacher, object, 0.028). For each Instance-CV triple (I, P, C), the C is added to the Entry (I, P). After above operations, the CV table of Tom's mobile phone is established. We give partial CV table in Tab.5.

Table 4. Final Triple-CV set

triple	CV (triple)	triple	CV (triple)
<Tom status WatchingTv>	0.25	<Tom status Speaking>	0.083
<Tom status Sleeping>	0.25	<Tom locatedIn LivingRoom>	0.083
<Tom status GivingLecture>	0.25	<TVSet locatedIn LivingRoom>	0.083
<Tom status Driving>	0.083	<TVSet status ON>	0.083
<Tom drives CarA>	0.083	<Tom locatedIn Bedroom>	0.083
<CarA SpeedExceed 60mph>	0.083	<Bedroom lightLevel Low>	0.083
<Tom locatedIn ClassRoom>	0.083	<Tom AroundVolume Low>	0.083
<Tom career Teacher>	0.083		

Table 5. Partial CV table of Tom's mobile phone.

Tom's phone	Subject	Predicate	Object
CarA	0.028		0.028
Tom	0.46		
locatedIn		0.11	
LivingRoom			0.056
…	…	…	…

4.2.2 Filtering Contexts

Above all, we give a formula to compute a raw context stored in the CKB. If a raw context is sent to context filter, its CV is computed by formula (1) firstly.

$$CV\ (<S, P, O>) = Entry\ (S, Subject) + Entry\ (P, Predicate) + Entry\ (O, Object) \qquad (1)$$

If CV(<S,P,O>) is larger than a preset value K, context <S,P,O> is regarded as useful information and sent to the context reasoner Otherwise, it is regarded trashy and discarded. However, the K has to be chosen properly. An overly large K leads to loss of useful contexts, while too small K compromise the efficiency of the context filter. This problem will be discussed in the next section.

5 Simulations

To present the performance of ACMR system, we will give the results of our prototype experiment in this section. The context reasoner is built based on Jena 2 Semantic Web Toolkit [8], which supports rule-based inference over OWL/RDF graphs. The experiment is run on the Windows 2000 operating system with the P4-2.4GHZ CPU and 256M main memory. To demonstrate the AOCF's improvement over pervious works, we simulate large raw context sets ranging from 500 to 7000 RDF triples in the CKB.

The first experiment is about the threshold K, which is mentioned in the previous section. When selecting a proper K, there are two factors to be considered, namely, availability and the performance of the system. If the context filter discards some useful contextual information, the policy engine of application may fail to carry out a context-awareness policy successfully. It may reduce the availability of our system. For a given K, the AOCF is simulated for several times. The availability of ACMR system can be represented as a newly introduced term successful rate, which means the proportion of successful simulations. On the other hand, the rule set and policy sets in ACMR system may be relatively static, while the raw contexts in the CKB change at any moment. As a result, the CV table can be changeless for a long time and cache mechanism can be imported to optimize the process of filtering context. Since the context reasoning should be accomplished frequently because of the high change rate of contextual information, the bottleneck of the performance in ACMR system is in the context reasoner. If too many trashy contexts are not discarded, the performance of the context reasoner will be low. In this experiment, *response time* of the context reasoner is used to reflect the performance of our system. In this experiment, we design three simulations with different policy sets and rule sets. Diverse CV tables are produced by these policy sets and rule sets. The smallest entry in CV table is defined as L and the Ls in the three simulations are 0.021, 0.028 and 0.035 respectively. Fig.5 shows the successful rate for each CV table. The successful rate decreases while the K increases. Note that there is a sharp drop of the successful rate when K = 2L Fig.6 illustrates the response time for each CV table. If K is set to 2L,

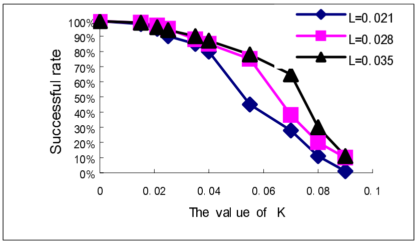


Fig. 5. The Successful rate of ACMR

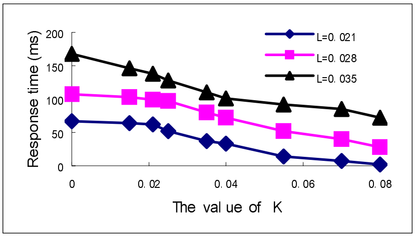


Fig. 6. The response time of ACMR

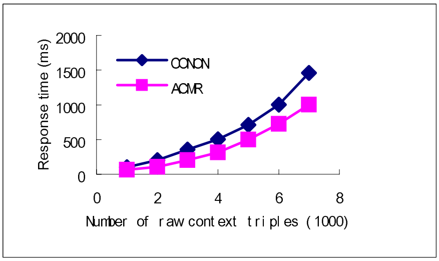


Fig. 7. Performance comparison between CONON and ACMR

the response time is almost halved. To achieve the tradeoff between the successful rate and response time, we configure the K as follow: Firstly, K is set to 2L to save the response time. If the policy engine cannot make a suitable policy successfully, the K is reset to 0 to insure the system against failure.

The second experiment compares the performance of ACMR system with the previous works in the context reasoning. Wang et.al. [10] [9] developed an ontology-based context reasoning mechanism without context filter technology. We call this mechanism as CONON because it is used to name the ontology they develop. In this experiment, K is set as described in the previous paragraph and the performance of CONON is got from [10]. Fig.7 shows the performance comparison between CONON and ACMR. In this figure, we can see while the number of raw context triples increase, the advantage of ACMR over CONON becomes larger and larger. It is easily to be explained in theory. As the proportion of useful contextual information is static, the more context triples are sent to the context filter, the more trashy contexts are discarded and the shorter the response.

6 Conclusions

In this paper, we present ACMR, a new context filtering system, in which Application-Oriented Context Reasoning (AOCF) technology is designed to upgrades the performance of system. Improvements are still needed in this system, such as adapting to the uncertain contexts in physical world. However, the results of experiments have shown the good performance of ACMR system.

References

1. Waiser, M.: The computer for the Twenty-first Century. Scienctific Am. Vol. 265, no. 3, (1991) 94-104
2. Dey, A.:Providing Architectural Support for Building Context-Aware Applications. PhD thesis, Georgia Institute of Technology (2000)
3. Ranganathan, Al-Muhtadi, J., Campbell, R.: Reasoning about Uncertain Contexts in Pervasive Computing Enviroments. Pervasive Computing, IEEE Vol. 3, no. 2, (2004) 62 - 70
4. http://www.uk.research.att.com/ab.html
5. Dey, A. et al.: A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. Human-Computer Interaction Journal, Vol. 16 (2001) 97-166
6. Kindberg, T., et al.: People, Places, Things: Web Presence for The Real World. Technical Report HPL-2000-16, HP Labs (2000)
7. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210
8. http://www.hpl.hp.com/semweb/jena2.htm
9. Wang, X.H., Dong, J., Chin, C.Y., Hettiarachchi, S.R., Zhang, D.Q.: Semantic Space: An Infrastructure for Smart Spaces. Pervasive Computing, IEEE Vol. 3, no. 3, (2004) 32 – 39
10. Wang, X.H., Zhang, D.Q., Pung, H.K.: Ontology Based Context Modeling and Reasoning using OWL. Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communication Workshops (2004)

Mobile Agent Based Wireless Sensor Network for Intelligent Maintenance

Xue Wang, Aiguo Jiang, and Sheng Wang

State Key Laboratory of Precision Measurement Technology and Instruments,
Department of Precision Instruments, Tsinghua University, Beijing 100084, P.R.China
wangxue@mail.tsinghua.edu.cn
jiangkai@tsinghua.org.cn
wang_sheng00@mails.tsinghua.edu.cn

Abstract. With the development of e-manufacturing, the flexible intelligent maintenance is necessary. Wireless sensor network is competent for a flexible maintenance system. Client/Server model adopted in traditional intelligent maintenance which requires transmitting vast data via network is unsuitable in wireless sensor network with limited bandwidth and unstable connection. Emerging mobile agent technology can reduce the network traffic and overcome the network latency, and is an eligible substitute for Client/Server model. This paper presents a mobile agent based wireless sensor network for intelligent maintenance. The flexible maintenance system integrates soft-computing, wireless sensor network and mobile agent technology. The system is used to maintain Mori Seiki Co. SV500 numerical control machining center and the results illustrate the effectiveness and efficiency of the system.

1 Introduction

In a flexible manufacture system, the deployment of manufacture equipments varies frequently along with the changes of tasks. More efficient and more flexible maintenance for equipments is required and traditional maintenance systems which lack of flexibility is incompetent [1].

Intelligent soft-computing algorithms have been applied to boost the validity of maintenance [2]. These algorithms fuse data from different sensors and form an authentic description of the sensed object. Multi-source information can gain more exact features and more robusticity with less time and less cost, which is impossible for solo-source information.

Wireless sensor network (*WSN*) is an emerging technology, which is flexible in deployment. *WSN* can be applied potentially in military, spatial, industrial and consumptive fields [3]. Spatial coverage and structural diversity of *WSN* make it fit for large-range and multiple-object monitoring. In *WSN*, however, the bandwidth is limited and the connection is easily affected by the surrounding environments. Client/Server model in traditional intelligent maintenance requires transmitting vast data, which consumes high bandwidth and needs continuous and steady connection. So intelligent maintenance based on Client/Server model is challenged. *Mobile agent technology* provides an effective method to overcome the challenge. Mobile agents

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 316 – 325, 2005.

process data locally, which reduces the capability of program codes and decreases the demand for network bandwidth. Uninterrupted steady connections between *sensor nodes* (*SNs*) are not required and only short-time connection is needed when mobile agents are transferring [4], [5]. *Mobile agent based WSN for intelligent maintenance* integrates *intelligent soft-computing algorithms*, *WSN* and *mobile agent technology* and is a promising flexible maintenance system.

The rest of this paper is organized as follows: Section 2 gives an overview of the mobile agent principal and the reasons why mobile agents may play an important role in a *WSN* for intelligent maintenance. Section 3 describes an architecture of *mobile agent based WSN for intelligent maintenance*, which is the core section of the paper. Section 4 provides an experiment where the system is used to maintain Mori Seiki Co. SV500 numerical control machining center. Finally, Section 5 presents the conclusions of this paper.

2 Mobile Agent Principle

2.1 Overview of Mobile Agent Principle

A mobile agent is a piece of software entity which travels from host to host and performs activities on user's behalf. It is also autonomous, proactive, and reactive, and exhibits capabilities to learn, cooperate, and move, as depicted in Fig. 1 [6].

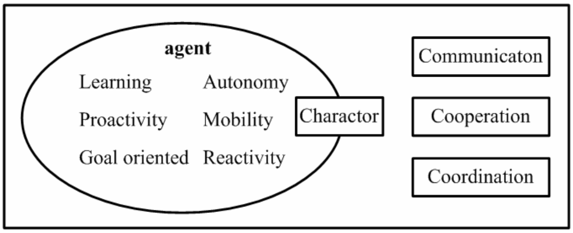
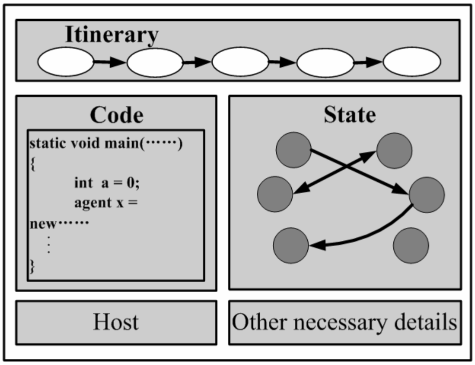


Fig. 1. The characteristics of a mobile agent



```
static void main(……)
{
        int  a = 0;
        agent x =
new……
   ⋮
}
```

Fig. 2. The basic structure of a mobile agent

A mobile agent generally consists of the following components [7], as illustrated in Fig. 2: *Itinerary* records the route and current position of the agent; *Code* stores fragments of the program; *State* records agent status and *Host* stores the server position; *Other necessary details* store other information related to the agent, so that operators will know what the agent does and who the owner is.

Mobile agents can reduce network traffic and provide an effective means of overcoming network latency. Mobile agents have the ability to interact with the environment where they execute. Once created, mobile agents operate asynchronously and autonomously in pursuit of their own goals and no external information or control inputs are required.

2.2 Mobile Agent Based WSN

Efficient maintenance is crucial for an orderly and large-scale manufacturing. Firstly, the maintenance system, which is nearly relevant to the operation of equipments and the quality of production, should respond to the fault/failure of equipments as soon as possible. Some researchers have applied *intelligent soft-computing algorithms* to the maintenance, such as neural networks etc. Secondly, in a flexible manufacturing system, the deployment of equipments transforms frequently along with the changes of tasks, so the traditional fixed maintenance system is unavailable. *WSN*, which consists of lots of *SNs*, is competent for a flexible maintenance system. The *SNs* can process raw data locally and self-control, and communicate with neighbor *SNs* or end-user by wireless connection, and cooperate with each other.

When *WSN* and *intelligent soft-computing algorithms* are introduced into the traditional maintenance system, a novel flexible intelligent maintenance system appears. The current data processing paradigm tends to use Client/Server model. In this model, however, there exists an inherent contradiction: *WSN* has limited bandwidth and is seriously affected by the surrounding environment; meanwhile, the *intelligent soft-computing algorithms* require transmitting vast data by the wireless connection for a precise result.

The contradiction has deferred the application of *WSN* for intelligent maintenance. Mobile agent paradigm is proposed to respond to the contradiction. In this paradigm, data stay at the local node, while the integration process is moved to the data nodes. By transmitting the computation engine instead of data, *mobile agent based WSN* offers the following important benefits [4], [8]: *Network bandwidth requirement* is

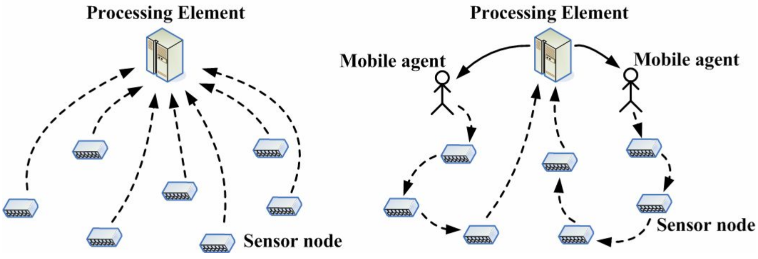


Fig. 3. Comparison of architectures between the canonical WSN and the mobile agent based WSN: (a) Canonical WSN (b) Mobile agent based WSN

reduced; *Network scalability* is better and the performance of the network is not affected when the number of *SNs* is increased; *Extensibility* is boosted and mobile agents can be programmed to carry task-adaptive processes which extends the capability of the system; *Stability* is improved and mobile agents can be sent when the network connection is alive and return results when the connection is re-established.

Fig. 3 provides a comparison of architectures between the canonical *WSN* and the *mobile agent based WSN*.

3 Mobile Agent Based Wireless Sensor Network for Intelligent Maintenance

3.1 Network Infrastructure

The *WSN* proposed in this paper adopts a two-layer hierarchical architecture, as depicted in Fig.4. In the lower layer, *SNs* communicate by wireless network, and equipped with a processing element and a local database. A processing element and its associated *SNs* are referred to as a patch network. In the upper layer, all processing elements are connected to the central-server (*CS*) accompanied by a global database (*GD*) [9].

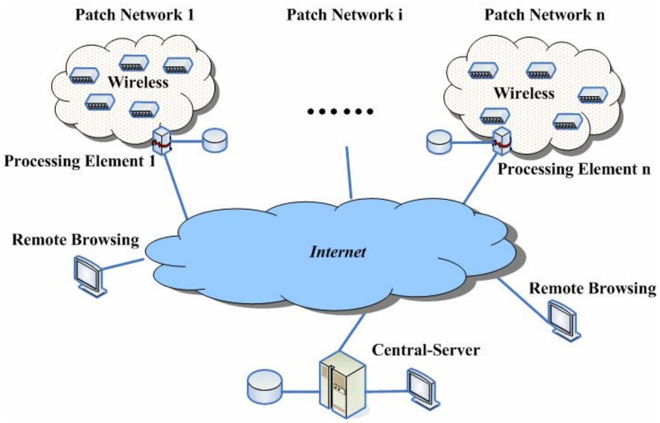


Fig. 4. Two-layer hierarchical architecture of WSN

Three different levels of agents are presented: *Low-Level Sensors* (*LLSensors*), *Soft-Computing Agents* (*SCAgents*) and *High-Level Decision Agents* (*HLDAgents*). They have different purposes, different features and different fields of action. A *Directory Service* is presented for registering some agent services. The architecture of the *mobile agent based WSN for intelligent maintenance* is shown in Fig.5.

LLSensors consist of data acquisition agents and preprocess agents. They are static agents and specific for the type of data source. *SCAgents* carry out local data fusions with the data from *LLSensors*. They are mobile agents that specialized with *intelligent soft-computing algorithms*. *HLDAgents* aggregate the fusion results of *SCAgents* and make final decisions. Additionally, *SCAManager* deals with the registration and management of *SCAgents* and *HLDAgents*; *Directory Service* stores the information of *SCAgents* and *HLDAgents*; *Performance Data Repository* and *Agent Cache Repository* are also presented to boost the performance of the network.

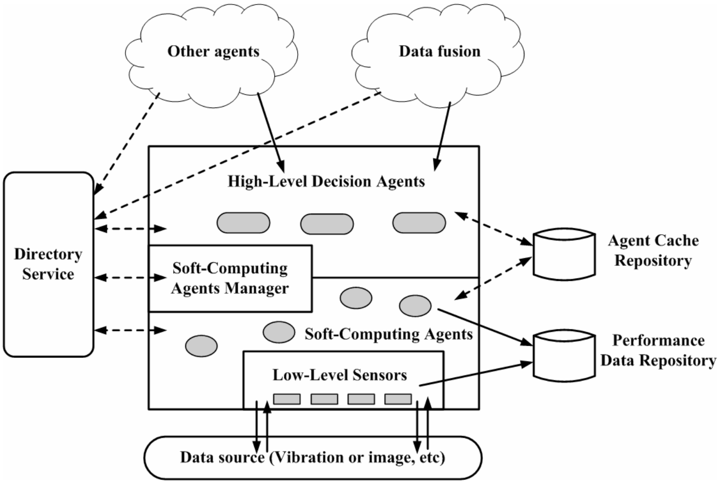

Other agents
Data fusion
Directory Service
High-Level Decision Agents
Soft-Computing Agents Manager
Soft-Computing Agents
Low-Level Sensors
Agent Cache Repository
Performance Data Repository
Data source (Vibration or image, etc)

Fig. 5. Architecture of the mobile agent based WSN for intelligent maintenance

3.2 Network Components

Mobile agent based WSN for intelligent maintenance consists of two primary parts: hardware system and software utilities. Hardware system mainly includes:

- *Sensor Node*: the place that acquires data and embeds a mobile agent context, a small local database and wireless communication module.
- *Central Server*: the place that embeds a mobile agent context and a wireless communication module, and dispatches the agents, and displays the real-time operating status.
- *Global Database*: the place that stores the fault/failure data of equipments. It is connected with central server by wired Ethernet.
- *Wireless Network* (*WN*): the network environment that adopts IEEE 802.11b, and provides the communication connection between *CS* and *SNs*.

Software utilities comprise mobile agent platform and mobile agent intelligent maintenance system. The former provide an operation context for mobile agents, where mobile agents can create, replicate, dispatch to or recall from *SNs*, sleep or waken, and eliminate, etc, as depicted in Fig.6; the latter consists of 5 kinds of agents:

- *Mobile Agent Platform* (*MAP*): the operation context for mobile agents.
- *Main Controller Agent* (*MCAgent*): the manager agent that offers a user-interface to accept the inputs, and display the results, and manages other mobile agents.
- *Data Acquisition Agent* (*DAAgent*): the static agents that acquire real-time data in *SNs*.
- *Alarm Agent* (*AAgent*): the mobile agents that monitor the fault/failure symbols of equipments and send alarms to *MCAgent*.

- *Intelligent Dispatching Agent* (*IDAgent*): the static agent that adopts Radial Basis Function neural network to decide the type of *intelligent analysis agents* (*IAAgents*) that will be dispatched to the appointed *SN*.
- *Mobile Soft-computing Library* (*MSL*): an aggregation of various types of *IAAgents*, detailed description in section 3.3.

Here *MAP* is IBM's Aglet Workbench and *GD* adopts Microsoft Access. *CS* and *SNs* are embedded by IEEE 802.11b communication modules.

3.3 Mobile Soft-Computing Library

The *MSL* integrated *intelligent soft-computing algorithms* and *mobile agent technology*, which endow mobile agents with intelligent analysis ability and realize the mobility of intelligent analysis algorithms. *MSL* is modularized into various types of *IAAgents* and every agent embodies a kind of *intelligent soft-computing algorithm*. *IAAgents* are divided into three classes by functions:

- Data acquisition class: Web Camera Agent and Database Agent;
- Data analysis class: Wavelet Agent, FFT Agent, Mahalanobis Distance Agent and Gabor Agent;
- Intelligent diagnosis class: Neural Network Agent, Taguchi Optimization Agent and Wavelet Probability Neural Network (WPNN) Agent, etc.

According to the tasks, *IAAgents* are dispatched by *MCAgent* to specific *SNs*, and carry out the appointed processions, and then return the results to *MCAgent*. If a task is too complex, several agents should cooperate to finish it. For example, the real-time image acquisition requires Gabor Agent and Mahalanobis Distance Agent. Firstly Gabor Agent acquires the Gabor features of multiple images; secondly Mahalanobis Distance Agent calculates the distance d of the Gabor features and the standard feature space; lastly compare d and the threshold T, if $d \le T$, then normal, if $d > T$, then fault or failure.

3.4 System Operation Process

3.4.1 Data Acquisition Agents Acquire Real-Time Data

After parameters setup, *DAAgent* is dispatched to the *SNs* appointed by IP and port. *DAAgent* acquires real-time signals, such as vibration, current or temperature, and make an elementary analysis. If normal, the data will be discarded; if failure/fault, the data and the failure/fault symbols will be stored in the local database of the *SN* (①).

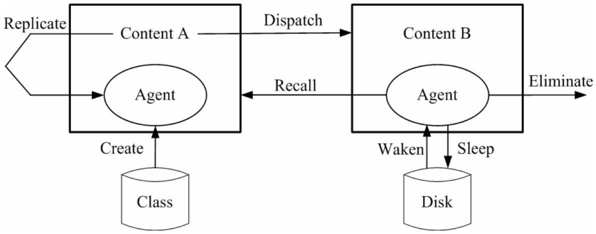


Fig. 6. Basic actions of mobile agents on the mobile agent platform

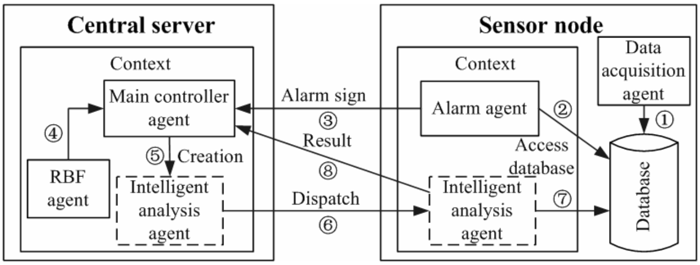


Fig. 7. The operation process of the mobile agent based WSN for intelligent maintenance

3.4.2 Alarm Agent Returns Symbols of Fault/failure

MCAgent dispatches *AAgents* to the specific *SNs* appointed by IP and port. *AAgents* resident in *SNs* and check the failure/fault symbols in the local database (②). If a symbol exceeds the threshold, *AAgent* reads the data related to the symbol, makes a farther analysis and then sends a message, whose parameters are the analysis results, to *MCAgent* (③).

3.4.3 Main Controller Agent Dispatches Intelligent Analysis Agents

The dispatch of *IAAgents* is controlled by *IDAgent*, whose inputs are the parameters of the messages from *AAgents* and whose outputs are a vector of 0 and 1. Every bit of the vector represents an *IAAgent*; if 1, then dispatch the *IAAgent*, if 0, then not. Once *MCAgent* is created, it is self-controlled and no external instructions are needed. *MCAgent* receives the message from *AAgents* and inputs the parameters of the message into *IDAgent* to decide which of *IAAgents* will be dispatched (④, ⑤). After the *IAAgent* is sent to the *SN* appointed by IP and port (⑥), it reads the related data and makes an intelligent soft-computing analysis (⑦), and then returns the final result to *MCAgent* (⑧). *MCAgent* displays the result in the main window. The operation process of the maintenance system can be depicted as Fig. 7.

4 Experiments and Results

The system is applied to the maintenance of ten Mori Seiki Co. SV500 numerical control machining centers, every one of which is installed with seven *SNs* , separately

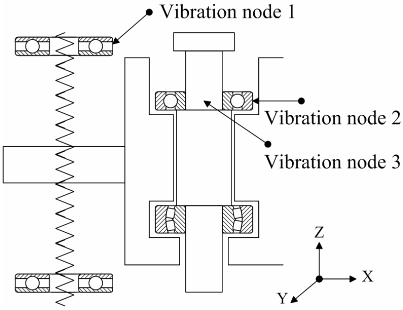
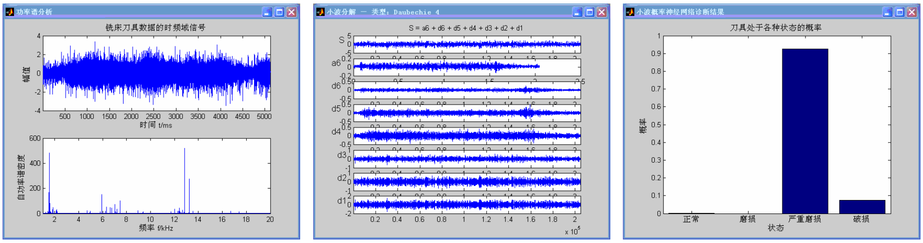
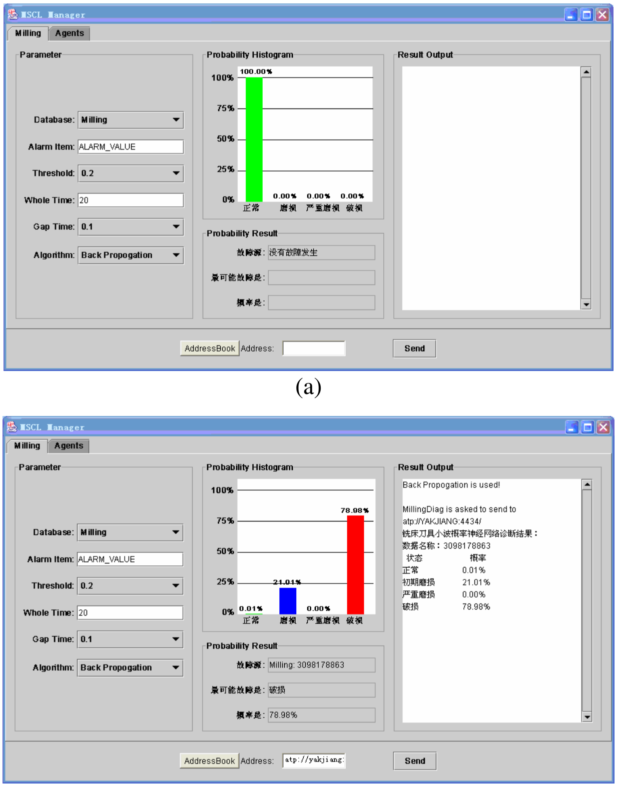


Fig. 8. Deploy positions of three SNs for vibration signals



(a) (b) (c)

Fig. 9. Intelligent analysis agents for SV500 maintenance: (a)FFT Agent (b)Wavelet Agent (c)WPNN Agent



Database: Milling

Alarm Item: ALARM_VALUE

Threshold: 0.2

Whole Time: 20

Gap Time: 0.1

Algorithm: Back Propogation

故障源: 没有故障发生

最可能故障是:

概率是:

Address:

(a)

Database: Milling

Alarm Item: ALARM_VALUE

Threshold: 0.2

Whole Time: 20

Gap Time: 0.1

Algorithm: Back Propogation

故障源: Milling: 3098178863

最可能故障是: 破损

概率是: 78.98%

```
Back Propogation is used!

MillingDiag is asked to send to
atp://YAKJIANG:4434/
铣床刀具小波概率神经网络诊断结果：
数据名称：3098178863
 状态        概率
正常        0.01%
初期磨损    21.01%
严重磨损    0.00%
破损        78.98%
```

Address: atp://yakjiang:

(b)

Fig. 10. The analysis processes and final decision results: (a)Normal (b)Failure

acquiring three channels of vibrations, three channels of currents and one channel of temperature. Fig.8 shows the positions of three *SNs* for vibration signals.

Three specially designed *IAAgents* are used to analyze the operation status of SV500. *MCAgent* inputs the parameters of the messages from *AAgents* into the *IDAgent* and decides which of the three agents will be dispatched to the appointed *SN*. Three *IAAgents* are: (1) *FFT Agent* which is used in a low probability of fault/failure and makes a power spectrum analysis of time domain signal and obtains the frequency domain features. (2) *Wavelet Agent* which is used in a medium probability of fault/failure and makes a Symlet4 7-level wavelet analysis of time domain signal; (3) *Wavelet Probability Neural Network* (*WPNN*) *Agent*, which is used in a high probability of fault/failure, and where 21 wavelet energy features are inputted into *WPNN* and a decision analysis is made to obtain the probability of 4 statuses: *Normal*, *Dull*, *Bad dull* and *Failure*, as depicted in Fig. 9.

The *IAAgents* analyze the data of seven *SNs* to get local decisions, and then return them to *MCAgent* which fuses the local decisions and get a precise evaluation of the operation status of SV500. The analysis processes and final decision results are shown in the main window of *MCAgent*, as depicted in Fig.10. Fig.10 (a) indicates a *Normal* status with the probability of 100%; Fig.10 (b) indicates a *Normal* status of the probability of 0.01%, a *Dull* status of the probability of 21.01% and a *Failure* status of the probability of 78.98%, so the most possible status is *Failure* and the SV500 must be powered off. So the maintenance system can guarantee the operation of equipments and decrease the unqualified productions efficiently.

5 Conclusions

This paper describes the use of the mobile agent paradigm to design an improved architecture of *WSN*. This paradigm saves network bandwidth and provides effective means for overcoming network latency. The promising maintenance system, *mobile agent based WSN for intelligent maintenance* integrates *intelligent soft-computing algorithms*, *WSN* and *mobile agent technology*. The system is applied for the automatic intelligent maintenance of SV500 and the experiment results indicate that mobile agent paradigm is an effective approach for *WSN*, especially when large amount of data transfer is involved, which is the case in intelligent maintenance.

Acknowledgement

This paper is sponsored by National Natural Science Foundation of China (No. 50175056; No. 60373014).

References

1. Kezunovic, M., Xu, X., Wong, D.: Improving Circuit Breaker Maintenance Management Tasks by Applying Mobile Agent Software Technology. Proc. of Transmission and Distribution Conference and Exhibition 2002: Asia Pacific (2002) 782–787
2. Xue, W., Aiguo, J., Sheng, W.: Distributed Sensor Networks for Multi-sensor Data Fusion in Intelligent Maintenance. Proc. of the 3rd International Symposium on Instrumentation Science and Technology. Harbin Institute of Technology Press, Harbin (2004) 587–592
3. Luo, R., C., Chih-Chen, Y., Kuo, L., S.: Multisensor Fusion and Integration: Approaches, Applications, and Future Research Directions. IEEE Sensors Journal 2 (2002) 107–119
4. Qi, H., Iyengar, S., S., Chakrabarty, K.: Distributed Sensor Networks——a Review of Recent Research. Journal of the Franklin Institute 6 (2001) 655–668

5. Horling, B., Vincent, R.: Distributed Sensor Network for Real Time Tracking. Proc. of the 5th International Conference on Autonomous agents (2001) 417–424
6. Yunyong, Z., Jingde, L.: Mobile Agent Technology. Tsinghua University Press, Beijing (2003)
7. Chuang, M., Chang, W.: Performance Monitoring Web Applications via a Mobile-agent Approach. Tunghai Science 5 (2003) 21–41
8. Puliafito, A., Tomarchio, O., Vita, L.: MAP: Design and Implementation of a Mobile Agent Platform. Journal of System Architecture. 2 (2000) 145–162
9. Marques, P., Simoes, P.: Providing Applications with Mobile Agent Technology. Proc. of Open Architectures and Network Programming (2001) 129–136

The Application of TSM Control to Integrated Guidance/Autopilot Design for Missiles

Jinyong Yu[1], Daquan Tang[2], Wen-jin Gu[1], and Qingjiu Xu[1]

[1] Department of Automatic Control, Naval Aeronautical Engineering Academy,
Yantai 264001, P.R. China
ihateujuer@163.com
[2] Department of Automatic Control, Beijing University of Aeronautics and Astronautics ,
Beijing 100083, P.R. China
tothelast@163.com

Abstract. A scheme for integrated guidance/autopilot design for missiles based on terminal sliding-mode control is proposed. Firstly, the terminal sliding mode control is introduced , based on which and the backstepping idea the guidance/control law is designed when an integrated guidance/autopilot model of the yaw plane is formulated. Secondly, an estimating method is given for the unavailable information of the maneuvering target, and an auxiliary control based on a sliding mode estimator is used to offset the estimation error. Finally, a simulation of some missile against high maneuvering targets on the yaw plane was made to verify the effectiveness and rightness of the scheme, and the simulation results have shown that high accuracy of hitting target can be got when the scheme is adopted.

1 Introduction

The dynamic performance of a sliding mode control system is mostly determined by the prescribed switching manifolds. In general, swiching manifolds are chosen to be linear hyperplanes. Such hyperplanes guarantee the asymptotic stability of the sliding mode. That is, the system state will reach the equilibrium in infinite time or tracking error will asymptoticly converge to zero. The convergent rate can be arbitrarily regulated according to parameter matrix of the switching manifolds. Nevertheless, tracking error will not converge to zero in a finite time in any case. Nonlinear switching manifolds such as TSM can improve the transient performance substantially. The main reason of using TSM in control system is that the TSM can make the state of system converge to zero in a finite time[1]. TSM control has been used successfully in control system designs[2-3]. In this paper, TSM is implicated to the integrated guidance/autopilot design.

On the guidance aspect, as the development of avionic meter and sensor, the acceleration of the targets could be got, thus the APN and PGL appeared. In the past ten years, the three dimensional guidance law and some factors are considered in the designing process such as induced drag, time-varying velocity and the internal dynamics [4].The traditional guidance law can not satisfy the modern need any more, so that some researchers proposed some new guidance algorithms, such as the

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 326 – 335, 2005.

guidance law based on the Lyapunov stability theorem[5],and variable structure control theory[6-7].However, there are some problems to be considered. First is that the measurements in an end game are nonlinear in Cartesian coordinates, as a consequent, there is linearization in the filtering update process. The measurement updates are linear in polar coordinate based state space, the propagation between the measurement updates in this case leads to nonlinear equations, thus the state used in guidance law are suboptimal. The second problem is that the guidance law is formulated assuming that the separability of the guidance and control law and the estimators which do not hold indeed. The third one is that the autopilot is usually designed independent of the estimator and the guidance law. Inspired by the idea of [8-10], a design integrated scheme is brought out from a new stand point.

The contribution of this paper is that based on the backstepping idea and terminal sliding mode control theory, a solution of the command of acceleration and fin deflection are deduced for the integrated guidance/autopilot scheme. Besides, a sliding mode estimator is designed for the auxiliary control.

The construction of this paper is as follows, in the second section, an integrated guidance/autopilot model is formulated, in the third section, the control-guidance law is deduced and by using an auxiliary control based on a sliding mode estimator, the control will be more accurate and the robust performance will be enhanced. A simulation is made and the simulation results are presented in the fourth section, finally, a conclusion is made.

2 The Model of Integrated Guidance/Autopilot

In Fig 1, a standard two-dimensional scenario is shown.

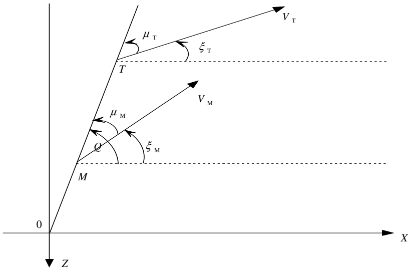


Fig. 1. Intercept scenario

Where R -the distance between missile and target; ξ_M, ξ_T -flight-path angle of missile and target; V_M, V_T - velocity of missile and target; Q -LOS.

The kinematical equations between missile and target can be shown as (1)

$$\dot{R} = V_T \cos \mu_T - V_M \cos \mu_M \tag{1}$$

$$\dot{R} = V_{\mathrm{T}} \cos \mu_{\mathrm{T}} - V_{\mathrm{M}} \cos \mu_{\mathrm{M}} \tag{2}$$

$$Q = \xi_{\mathrm{T}} + \mu_{\mathrm{T}} \tag{3}$$

$$Q = \xi_{\mathrm{M}} + \mu_{\mathrm{M}} \tag{4}$$

Equations (1) and (2) can be changed into the following form of (5-6)

$$\dot{R} = V_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q) - V_{\mathrm{M}} \cos(\xi_{\mathrm{M}} - Q) \tag{5}$$

$$R\dot{Q} = V_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) - V_{\mathrm{M}} \sin(\xi_{\mathrm{M}} - Q) \tag{6}$$

derivate (6), and substitute (5) into it, after some mathematical manipulations, it can obtain that

$$\begin{aligned} R\ddot{Q} + 2\dot{R}\dot{Q} &= -\dot{V}_{\mathrm{M}} \sin(\xi_{\mathrm{M}} - Q) + \dot{V}_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) \\ &+ V_{\mathrm{T}}\dot{\xi}_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q) - V_{\mathrm{M}}\dot{\xi}_{\mathrm{M}} \cos(\xi_{\mathrm{M}} - Q) \end{aligned} \tag{7}$$

because the overloads of missile and target can be written as $n_{\mathrm{T}} = -V_{\mathrm{T}}\dot{\xi}_{\mathrm{T}}$, $n_{\mathrm{M}} = -V_{\mathrm{M}}\dot{\xi}_{\mathrm{M}}$, so(7)can be written as

$$\begin{aligned} R\ddot{Q} + 2\dot{R}\dot{Q} &= -\dot{V}_{\mathrm{M}} \sin(\xi_{\mathrm{M}} - Q) + \dot{V}_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) \\ &- n_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q) + n_{\mathrm{M}} \cos(\xi_{\mathrm{M}} - Q) \end{aligned} \tag{8}$$

Let $R\dot{Q} = U$, then

$$\begin{aligned} \dot{U} &= -\dot{V}_{\mathrm{M}} \sin(\xi_{\mathrm{M}} - Q) + \dot{V}_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) - n_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q) \quad + n_{\mathrm{M}} \cos(\xi_{\mathrm{M}} - Q) - \dot{R}\dot{Q} \\ &= -\dot{V}_{\mathrm{M}} \sin(\xi_{\mathrm{M}} - Q) + \dot{V}_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) - n_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q) \quad + n_{\mathrm{M}} \cos(\xi_{\mathrm{M}} - Q) - \frac{\dot{R}}{R} U \\ &= f(\dot{V}_{\mathrm{M}}, \dot{V}_{\mathrm{T}}, n_{\mathrm{M}}, n_{\mathrm{T}}, \xi_{\mathrm{M}}, \xi_{\mathrm{T}}, Q) - \frac{\dot{R}}{R} U \end{aligned} \tag{9}$$

In addition , the following equations of missile shown in (6-8) can be got in[11].

$$n_M = \frac{qs}{mg}[(-C_x \sin\beta + 57.3\left(C_z^{\beta} + C_z^{\delta_y}\delta_y\right)\cos\beta] \tag{10}$$

$$\begin{aligned} \dot{\beta} &= \frac{1}{mV_M}(57.3qSC_z^{\beta}\beta - P\cos\alpha\sin\beta) + 57.3\frac{qS}{mV_M}C_z^{\delta_y}\delta_y \\ &+ \omega_x \sin\alpha + \omega_y \cos\alpha \end{aligned} \tag{11}$$

$$\dot{\omega}_y = \frac{qSL^2}{J_yV_M} m_y^{\bar{\omega}_y}\omega_y + \frac{57.3qsL}{J_y} m_y^\beta \beta + \frac{57.3qsL}{J_y} m_y^{\delta_y}\delta_y + \frac{J_z - J_x}{J_y}\omega_x\omega_z \tag{12}$$

Because the scheme researched is on the yaw plane , in order to simplify the form of expression and facilitate the design, some hypothesis are introduced.

1. α and β are small enough, so that the following relations can hold, $\sin\alpha \approx \alpha$, $\sin\beta \approx \beta$, $\cos\alpha = \cos\beta \approx 1$.
2. There is no coupling among the three channels.
3. The influence of fin deflection on aerodynamic force can be neglected comparing the other terms.

Thus we have the following integrated guidance/autopilot model of missile on the yaw plane shown in (13-16).

$$\dot{U} = f(\dot{V}_{\mathrm{M}}, \dot{V}_{\mathrm{T}}, n_{\mathrm{M}}, n_{\mathrm{T}}, \xi_{\mathrm{M}}, \xi_{\mathrm{T}}, Q) - \frac{\dot{R}}{R}U \tag{13}$$

$$n_M = \frac{qs}{mg}[(-C_x\beta + 57.3\left(C_z^\beta + C_z^{\delta_y}\delta_y\right)] \tag{14}$$

$$\dot{\beta} = \frac{1}{mV_M}(57.3qSC_z^\beta - P)\beta + \omega_y \tag{15}$$

$$\dot{\omega}_y = \frac{qSL^2}{J_yV_M} m_y^{\bar{\omega}_y}\omega_y + \frac{57.3qsL}{J_y} m_y^\beta \beta + \frac{57.3qsL}{J_y} m_y^{\delta_y}\delta_y \tag{16}$$

3 Integrated Control-Guidance Law Design

By using the constructive approach of backstepping and TSM, the control-guidance law is deduced, the process of which can be concluded as follows.

Step 1:

Choose the Lyapunov function candidate as (17)

$$V_1 = \frac{1}{2}U^2 \tag{17}$$

Derivate it and substitute (9) into it, it will obtain that

$$\begin{aligned} \dot{V}_1 = U\dot{U} &= U[-\dot{V}_{\mathrm{M}}\sin(\xi_{\mathrm{M}} - Q) + \dot{V}_{\mathrm{T}}\sin(\xi_{\mathrm{T}} - Q) \\ &- n_{\mathrm{T}}\cos(\xi_{\mathrm{T}} - Q) + n_{\mathrm{M}}\cos(\xi_{\mathrm{M}} - Q) - \frac{\dot{R}}{R}U] \end{aligned} \tag{18}$$

Let

$$\dot{V}_1 = -\omega V_1^{\chi} \tag{19}$$

The reaching time of the terminal attractor will be

$$t_r = \frac{(V(0))^{1-\chi}}{\omega(1-\chi)}, \forall \chi \in (0,1) \tag{20}$$

substituting (17) into (19) ,we have

$$\dot{U} = -2^{-\chi} \omega U^{2\chi-1} = -2^{-\chi} \omega |U|^{2\chi-1} sign(U) \tag{21}$$

if the control law is chosen as

$$\begin{aligned} n_{\mathrm{Mc1}} = -\frac{1}{\cos(\xi_{\mathrm{M}} - Q)}[-\frac{\dot{R}}{R}U - \dot{V}_{\mathrm{M}} \sin(\xi_{\mathrm{M}} - Q) + \dot{V}_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) \\ - n_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q) + 2^{-\chi} \omega |U|^{2\chi-1} sign(U)] \end{aligned} \tag{22}$$

then we have

$$\dot{U} = -2^{-\chi} \omega |U|^{2\chi-1} sign(U) \tag{23}$$

So that the desired dynamics of (19) can be got. As long as the missile is launching towards the target, then $\cos(\xi_{\mathrm{M}} - Q) \neq 0$,and (22) is nonsingular, such that $U = R\dot{Q}$ convergent to zero, and the missile can intercept the target successfully.

Step 2:

According to command of overload shown in (22) and output overload shown in (14), the command of side slip angle can be got through the following filter

$$\ddot{\beta}_d + \eta_2 \dot{\beta}_d = \eta_1 (n_{Mc} - n_M) \tag{24}$$

where $\eta_1 = 0.5, \eta_2 = 9.5$.

Step 3:

Let $\tilde{\beta} = \beta - \beta_d$,$\tilde{\omega}_y = \omega_y - \omega_{yd}$,where ω_y can be considered as a virtual control in (15),then we have

$$\begin{aligned} \dot{\tilde{\beta}} &= \frac{1}{mV_M}(57.3qSC_z^{\beta} - P)\beta + \omega_y - \dot{\beta}_d \\ &= \frac{1}{mV_M}(57.3qSC_z^{\beta} - P)\beta + \omega_y - \omega_{yd} - \dot{\beta}_d + \omega_{yd} \\ &= \frac{1}{mV_M}(57.3qSC_z^{\beta} - P)\beta + \tilde{\omega}_y - \dot{\beta}_d + \omega_{yd} \end{aligned} \tag{25}$$

Let

$$\omega_{yd} = -\frac{1}{mV_M}(57.3qSC_z^{\beta} - P)\beta + \dot{\beta}_d - k_3\,\tilde{\beta} \tag{26}$$

Then it can obtain that

$$\dot{\tilde{\beta}} = \tilde{\omega}_y - k_3\,\tilde{\beta} \tag{27}$$

Choose the Lyapunov function candidate as

$$V_2 = \frac{1}{2}\tilde{\beta}^2 + \frac{1}{2}\tilde{\omega}_y{}^2 \tag{28}$$

Derivate it , it can obtain that

$$\begin{aligned}\dot{V}_2 &= \tilde{\beta}\dot{\tilde{\beta}} + \tilde{\omega}_y\dot{\tilde{\omega}}_y = \tilde{\beta}(\tilde{\omega}_y - k_3\,\tilde{\beta}) + \tilde{\omega}_y(\frac{qSL^2}{J_yV_M}m_y^{\bar{\omega}_y}\omega_y \\ &+ \frac{57.3qsL}{J_y}m_y^{\beta}\beta + \frac{57.3qsL}{J_y}m_y^{\delta_y}\delta_y - \dot{\omega}_{yd}) \\ &= \tilde{\beta}\tilde{\omega}_y - k_3\,\tilde{\beta}^2 + \tilde{\omega}_y(\frac{qSL^2}{J_yV_M}m_y^{\bar{\omega}_y}\omega_y + \frac{57.3qsL}{J_y}m_y^{\beta}\beta \\ &+ \frac{57.3qsL}{J_y}m_y^{\delta_y}\delta_y - \dot{\omega}_{yd})\end{aligned} \tag{29}$$

Let

$$\delta_y = -\frac{J_y}{57.3qsLm_y^{\delta_y}}[\tilde{\beta} + (\frac{qSL^2}{J_yV_M}m_y^{\bar{\omega}_y}\omega_y + \frac{57.3qsL}{J_y}m_y^{\beta}\beta - \dot{\omega}_{yd}) + k_4\tilde{\omega}_y] \tag{30}$$

Then it can obtain that

$$\dot{V} = -k_4\tilde{\omega}_y{}^2 - k_3\,\tilde{\beta}^2 \le 2bV \tag{31}$$

where $b = \min(k_3, k_4)$,from (31) we can see that the exponential stability can be got.

In fact, the value of $\dot{V}_{\mathrm{T}}$, n_{T} and ψ_{T} in (13) can not be measured directly, so it is necessary to estimate them, and the estimating method will be given.

Refer to Figure 1, there are the following relations

$$x = x_{\mathrm{T}} - x_{\mathrm{M}} \tag{32}$$

$$z = z_{\mathrm{T}} - z_{\mathrm{M}} \tag{33}$$

$$R = \sqrt{x^2 + z^2} \tag{34}$$

$$\dot{R} = (x\dot{x} + z\dot{z})/R \tag{35}$$

$$Q = \arctan(-\frac{z}{x}) \tag{36}$$

$$\dot{Q} = \frac{z\dot{x} - \dot{z}x}{R^2} \tag{37}$$

After some mathematical manipulations, it can obtain that

$$\dot{x}_{\mathrm{T}} = \dot{R}\cos Q - R\dot{Q}\sin Q + \dot{x}_{\mathrm{M}} \tag{38}$$

$$\dot{z}_{\mathrm{T}} = -\dot{R}\sin Q - R\dot{Q}\cos Q + \dot{z}_{\mathrm{M}} \tag{39}$$

then the estimation of the velocity of the target can be expressed as

$$\hat{V}_{\mathrm{T}} = (\dot{x}_{\mathrm{T}}^2 + \dot{z}_{\mathrm{T}}^2)^{1/2} \tag{40}$$

and we can get the estimation of ψ_{T}

$$\hat{\xi}_{\mathrm{T}} = \arctan(-\frac{\dot{z}_{\mathrm{T}}}{\dot{x}_{\mathrm{T}}}) \tag{41}$$

By using a low pass filter, $\dot{\hat{\xi}}_{\mathrm{T}}$ and $\dot{\hat{V}}_{\mathrm{T}}$ can be got. Then it can obtain that

$$\hat{n}_{\mathrm{T}} = -\frac{\hat{V}_{\mathrm{T}}}{g}\dot{\hat{\xi}}_{\mathrm{T}} \tag{42}$$

so that (22) can be rewritten as

$$\begin{aligned} n_{\mathrm{Mc1}} = -\frac{1}{\cos(\xi_{\mathrm{M}} - Q)}[-\frac{\dot{R}}{R}A - \dot{V}_{\mathrm{M}}\sin(\xi_{\mathrm{M}} - Q) \\ + \dot{\hat{V}}_{\mathrm{T}}\sin(\hat{\xi}_{\mathrm{T}} - Q) - \hat{n}_{\mathrm{T}}\cos(\hat{\xi}_{\mathrm{T}} - Q) + 2^{-\chi}\omega\left|A\right|^{2\chi-1} sign(A)] \end{aligned} \tag{43}$$

In practice, the estimation error of the motion information of the target can not be neglected, now an auxiliary control based on a sliding mode estimator is used.

Let

$$\begin{aligned} D = \dot{\hat{V}}_{\mathrm{T}} \sin(\hat{\xi}_{\mathrm{T}} - Q) - \dot{V}_{\mathrm{T}} \sin(\xi_{\mathrm{T}} - Q) \\ - [\hat{n}_{\mathrm{T}} \cos(\hat{\xi}_{\mathrm{T}} - Q) - n_{\mathrm{T}} \cos(\xi_{\mathrm{T}} - Q)] \end{aligned} \tag{44}$$

Define the control based on the estimator as

$$-\hat{D} / \cos(\xi_{\mathrm{M}} - Q) = n_{Ms} \tag{45}$$

where $\hat{D}$ is the estimation of (44),define a sliding quantity

$$\sigma = A - z \tag{46}$$

where σ is an auxiliary sliding variable, which can be got by

$$\dot{z} = D - k_1 sign(\sigma) \tag{47}$$

derivate (46) and substitute (47) into it, we have

$$\dot{\sigma} = D - k_1 sign(\sigma) \tag{48}$$

As long as $k_1 > |D|$, σ will convergent to zero in finite time.

Then it can obtain that

$$D = \dot{\sigma} + k_1 sign(\sigma) \tag{49}$$

By using a low pass filter, we have

$$\hat{D} = \frac{1}{k_2 s + 1} [\dot{\sigma} + k_1 sign(\sigma)] \tag{50}$$

then the control part based on the sliding mode estimator can be expressed as

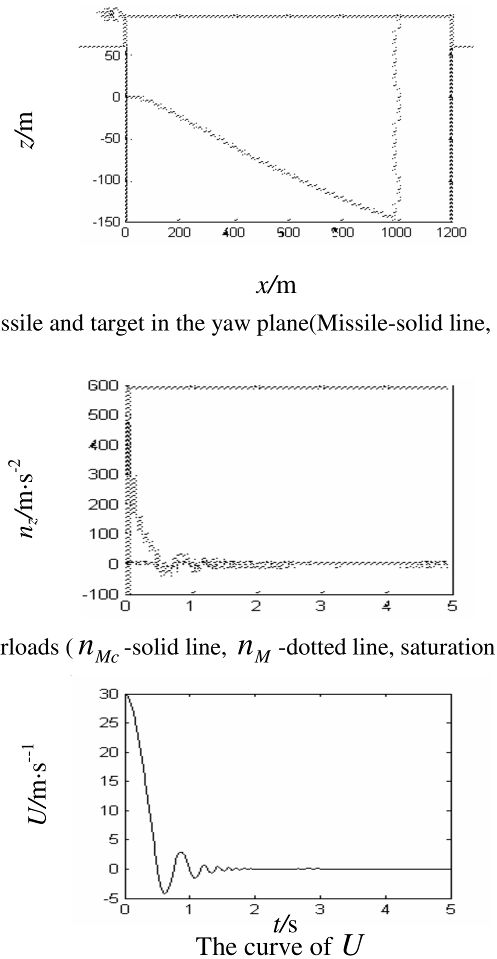
$$n_{Ms} = -\hat{D} / \cos(\xi_{\mathrm{M}} - Q) \tag{51}$$

then the total control can be written as

$$n_{Mc} = n_{Mc1} + n_{Ms} \tag{52}$$

4 Simulation

Some anti-vessel missile is taken as an example, when the simulation parameters are chosen as follows: simulation step is 0.002s, $V_{\mathrm{M}} = 300\mathrm{m/s}$, $V_{\mathrm{T}} = 50 + 5\sin(t)\mathrm{m/s}$, the initial value of ξ_{M} is set to be $\xi_{\mathrm{M}}(0) = 0^\circ$,and $\xi_{\mathrm{T}} = 90 + 50\sin(3t)$, the initial position coordinates are set to be as $(x_{\mathrm{M}}, z_{\mathrm{M}})$ =(0,0)(m), $(x_{\mathrm{T}}, z_{\mathrm{T}})$=(1000, 100)(m). The saturation value of the command overload is set to be 15*g*. Seen from the simulation results in Figure 2, the missile can hit the target accurately, and the missing distance is only 0.22m.



The curve of missile and target in the yaw plane(Missile-solid line, Target-dotted line)

The curves of overloads (n_{Mc} -solid line, n_M -dotted line, saturation of n_{Mc} -dashed line)

The curve of U

Fig. 2. Simulation results

5 Conclusion

In this paper, a scheme of integrated guidance/autopilot design for missiles on the yaw plane is considered. When a integrated guidance/autopilot model of the yaw plane is formulated , based on the backstepping idea and terminal sliding mode control, the guidance/control law is designed. In orderto verify the effectiveness and rightness of the integrated design scheme, a simulation of some missile against maneuvering targets have been made, the simulation results have shown high accuracy of hitting target have been got.

References

[1] Yu, H.Y., Gu, W.J. and Yang, Z.Y.: Novel Fast Adaptive Terminal Sliding Mode Control for Cross Beam System, Proceedings of the 5th Congress on Intelligent Control and Automation, June 15-19,(2004)1208–1211.

[2] Yu, X. and Man, Z.: Model Reference Adative Control Systems with Terminal Sliding Modes, Int. J.Control, Vol. 64, No.6,(1996)1165–1176.

[3] Man, Z. and Yu, X.:Terminal Sliding Mode Control of MIMO Systems, IEEE Trans,Circuits System, Vol.44, No1,(1997)1065-1070.

[4] Yaesh, I., and Ben-Asher, J. Z.,: Optimal Guidance with a Single Uncertain Time Lag, Journal of Guidance, Control, and Dynamics, Vol. 18,No. 5, (1995) 981–988.

[5] Song, S. H., and Ha, I. J.: A Lyapunov-Like Approach to Performance Analysis of 3-Dimensional Pure PNG Laws, IEEE Transactions on Aerospace and Electric Systems, Vol. 30, No. 1, (1994) 238–247.

[6] Zhou, D., Mu, C., and Xu, W.: Adaptive Sliding-Mode Guidance of a Homing Missile, Journal of Guidance, Control, and Dynamics, Vol. 22, No. 4,(1999) 589–594.

[7] Babu, K. R., Sarma, I. G., and Swmy, K. N.: Switched Bias Proportional Navigation for Homing Guidance Against Highly Maneuvering Target, Journal of Guidance, Control, and Dynamics, Vol. 17, No. 6, (1994) 1357–1363.

[8] Menon,.P.K., Ohlmeyer, E.J.:Integrated Design of Agile Missile Guidance and Autopilot Systems, Control Engineering Practice 9 (2001)1095-1106.

[9] Neil, F.P., Todd, D..J.n,:Integrated Missile Guidance and Control: a State Dependent Riccati Differential Equation Approach , Proceedings of the 1999 IEEE,International Conference on Control Applications,Kohala Coast-Island of Hawai'I,Hawai'I,USA, 8(1999)22-27

[10] Ernest, O.,J and Steve, M.: Optimal Design of Integrated Missile Guidance and Control, American Institute of Aeronautics and Astronautics, Inc.

[11] E.Devaud, Harcaut, J.P. and Siguerdidjane.H.: Various Strategies to Design a Three-axes Missile Autopilot, AIAA (99)187-195.

The Robust Control for Unmanned Blimps Using Sliding Mode Control Techniques

Guoqing Xia and Benkun Yang

School of Power and Nuclear Engineering, Harbin Engineering University,
Harbin 150001, P R China
xiaguoqing@hrbeu.edu.cn

Abstract. The control system design problem for unmanned blimps has not only theoretical significance but also practical signification because the moving behaviors of blimps are complex nonlinear with coupling, unstructured, imprecise models and disturbance. In this paper the sliding mode control method is applied to control the blimp maneuvering due to its strong robustness. The simulation results show that sliding mode control method is suitable for unmanned blimps through comparing to the simulation curves of sliding mode controllers and PD controllers.

1 Introduction

The applying research of blimps has received considerable attention in recent years due to the useful properties of blimps. Blimps possess easy hovering, wide altitude, low noise and low power consumption characteristics, so they are mainly used as low-speed, low-altitude platforms for exploration, monitoring, transportation and research purposes. Our current aims in the aerial robot field are designed the autonomous flight control system for unmanned blimps using onboard control and navigation systems for further research. The long goals are developing the cooperative control for multiple unmanned blimps to execute searching and exploring mission. If we want to bring blimp's properties into full play the control methods are required to adapt different ways. Azinheira et al applied H_∞ control approach to blimps [1]; de Paiva de al used PID controller to control a semi-autonomous blimp [2]; Hygounenc et al did backstepping techniques research for blimps [3]; Wells et al do research of remotely piloted airships [4]. In this paper, we consider the sliding mode control method because it has strong robustness, thus it can satisfy the requirement of complex nonlinear behaviors of blimps. At present, the sliding mode control method has been applied to a lot of fields, and the followings are some successful examples. Huang et al used sliding mode control method to design control systems of aircrafts [5]; Jafarov et al applied sliding mode control to the model F-18 [6]; Healey et al employed sliding mode control method for autonomous diving and steering of unmanned underwater vehicles [7]. Their research shows that the sliding mode control method has excellent properties.

The paper is organized as follows: section two describes the mathematic models of the blimp. Section three discusses the sliding mode control method. Section four gives the simulation results and section five provides our conclusions.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 336 – 345, 2005.

2 Blimp Modeling

2.1 Kinetic Model of Blimps

The complex nonlinear behaviors of blimps need us to design control approaches and evaluate the effects before they fly on the real world. The modeling and simulation is an important tool for research of new methods, so we need to build the models of the blimp. In order to describe the dynamic models that can reflect primary characteristics of the blimp, two types of coordinate systems are used to represent blimp motion. The one is an inertial frame system called earth-fixed frame ($E - X_e Y_e Z_e$), and the other one is called body-fixed frame ($O - xyz$) which fixed to blimp whose origin is at the blimp body. The coordinate system is given in figure 1.

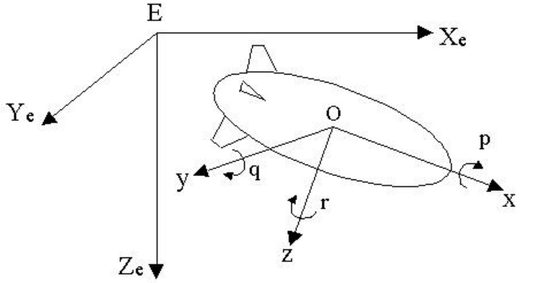


Fig. 1. Defining coordinate systems of a blimp, the earth-fixed reference frame is used to integrate dynamic equations of blimp motion; the body-fixed frame is used for calculating the forces and moments acting on the blimp

In the coordinates the complex nonlinear equations of the blimp motion can be expressed in six degrees of freedom (6DOF) model [8] as follow:

$$M\dot{x} = F_d(x) + E(x) + G + P \ . \tag{1}$$

where x-dot is the velocity vector (6× 1) including the linear speed u, v, w (surge, sway and heave) and angular speed p, q, r (roll, pitch and yaw), these speeds are described in the body-fixed frame, M is mass matrix (6×6) including added mass, F_d is dynamic forces and moments vector (6× 1), E is aerodynamic forces and moments vector (6× 1), G is gravity and buoyancy vector (6× 1), P is propulsion forces and moments vector (6× 1). For more details about the equation, see reference [8][9].

In computer simulation, the position and velocity of blimp need to be transformed from body-fixed frame to earth-fixed frame. The linear velocities and angular rates are $\{\dot{X}_e, \dot{Y}_e, \dot{Z}_e\}$ and $\{\dot{\phi}, \dot{\theta}, \dot{\psi}\}$ in earth-fixed frame, and the linear velocities and angular rates are $\{u, v, w\}$ and $\{p, q, r\}$ in body-fixed frame. The relationship between the linear and angular speeds in the body-fixed frame and earth-fixed frame is given by equations (2) and (3).

$$\begin{bmatrix} \dot{X}_e \\ \dot{Y}_e \\ \dot{Z}_e \end{bmatrix} = \begin{bmatrix} c\psi c\theta & -s\psi c\phi + c\psi s\theta s\phi & s\psi s\phi + c\psi s\theta c\phi \\ s\psi c\theta & c\psi c\phi + s\psi s\theta s\phi & -c\psi s\phi + s\psi s\theta c\phi \\ -s\phi & c\theta s\phi & c\phi c\theta \end{bmatrix} \begin{bmatrix} u \\ v \\ w \end{bmatrix} . \tag{2}$$

$$\begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -s\theta \\ 0 & c\phi & c\phi s\phi \\ 0 & -s\phi & c\phi c\theta \end{bmatrix}^{-1} \begin{bmatrix} p \\ q \\ r \end{bmatrix} . \tag{3}$$

where:

$c\phi = \cos\phi, c\theta = \cos\theta, c\psi = \cos\psi$ and $s\phi = \sin\phi, s\theta = \sin\theta, s\psi = \sin\psi$;

{ u, v, w } are linear velocities along Ox, Oy and Oz axes respectively;

{ p, q, r } are angular rates about Ox, Oy and Oz axes respectively;

{ X_e, Y_e, Z_e } are positions in earth-fixed frame;

{ ϕ, θ, ψ } are Euler angles, called roll angle, pitch angle and yaw angle respectively.

2.2 Mathematical Model for Designing Controller

For designing controller, it is very convenient to use a linear state space description of the system. Consider the roll motion is often induced by side-slip motions of the blimp, and because roll motion does not strongly influence sway and yaw and has no actuators to operate, so ignoring the roll effects. The lateral motions and the longitudinal equations are as respectively:

1) Lateral Motion Subsystem: The lateral motion model is given in the following equation (4), neglecting the interaction from longitudinal motion of the blimp and the effect of roll.

$$\begin{bmatrix} \dot{\psi} \\ \dot{v} \\ \dot{r} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & a_1 & a_2 \\ 0 & a_3 & a_4 \end{bmatrix} \times \begin{bmatrix} \psi \\ v \\ r \end{bmatrix} + \begin{bmatrix} 0 \\ b_1 \\ b_2 \end{bmatrix} \delta_r + D_r \ . \tag{4}$$

where D_r is disturbance term.

2) Longitudinal Motion Subsystem: The longitudinal motion model is expressed as equation (5), neglecting the interaction from lateral motion of the blimp.

$$\begin{bmatrix} \dot{\theta} \\ \dot{z} \\ \dot{w} \\ \dot{q} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ a_5 & 0 & 0 & 0 \\ a_6 & 0 & a_7 & a_8 \\ a_9 & 0 & a_{10} & a_{11} \end{bmatrix} \times \begin{bmatrix} \theta \\ z \\ w \\ q \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ b_3 \\ b_4 \end{bmatrix} \delta_e + D_e \ . \tag{5}$$

where D_e is disturbance term.

Equations (4) and (5) are used as designing the sliding mode control.

2.3 Navigation Systems

The position sensor of blimp is DGPS, and heading sensor of blimp is compass.

3 Control Strategy

The use of sliding mode technique has been gradually popular in the field of controller design for diverse systems because sliding mode control has strong robust ability during sliding motion. This paper will present the blimp flight control via the sliding mode control techniques. For linear system in the following equation (6), the sliding mode control is discussed [10].

$$\dot{X}(t) = AX(t) + Bu(t)\ . \tag{6}$$

where $A \in R^{n\times n}$, $B \in R^{n\times m}$, $X \in R^{n\times 1}$, $u \in R^{m\times 1}$ with $1 \le m < n$, matrix B can be assumed that it is full rank, that is rank(B)=m, without loss of generality. For more details about the method, see reference [10].

Define the sliding surface as: $\sigma(X) = SX$, where $S \in R^{m\times n}$, to determine the function, Lyapunov function is taken as: $V(X) = \frac{1}{2}(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_m^2)$, then $\dot{V}(X) = \sigma_1\dot{\sigma}_1 + \sigma_2\dot{\sigma}_2 + \cdots + \sigma_m\dot{\sigma}_m$.

To make asymptotic system stability is guaranteed then require $\dot{V}(X) < 0$ is satisfied, thus only need $\sigma_i\dot{\sigma}_i < 0 \quad for\ i = 1\ to\ m$.

Define $\dot{\sigma}_i = -\beta_i\sqrt{\sigma_i}\,\mathrm{sgn}(\sigma_i), \beta_i > 0 \quad for\ i = 1\ to\ m$, then that is $\sigma_i\dot{\sigma}_i < 0$ $for\ i = 1\ to\ m$.

For $\sigma(X) = SX$, then $\dot{\sigma}(X) = S\dot{X} = S(AX + Bu) = \Gamma$.

$$u = -(SB)^{-1}SAX + (SB)^{-1}\Gamma\ . \tag{7}$$

is the desired control law.

Where $\Gamma = \begin{bmatrix} -\beta_1\sqrt{\sigma_1}\,\mathrm{sgn}(\sigma_1) \\ \vdots \\ -\beta_m\sqrt{\sigma_m}\,\mathrm{sgn}(\sigma_m) \end{bmatrix}$

In order to solve S values, equation (6) needs to be expressed in following form

$$\dot{z}_1(t) = A_{11}z_1(t) + A_{12}z_2(t)\ . \tag{8}$$

$$\dot{z}_2(t) = A_{21}z_1(t) + A_{22}z_2(t) + B_2u(t)\ . \tag{9}$$

with the sliding surface is written as

$$\sigma(z) = S_1z_1(t) + S_2z_2(t)\ . \tag{10}$$

where the change of coordinates is defined by an orthogonal matrix T_r , so that $z(t) = T_r X(t)$, then $T_r A T_r^T = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$, $T_r B = \begin{bmatrix} 0 \\ B_2 \end{bmatrix}$, $ST_r^T = \begin{bmatrix} S_1 & S_2 \end{bmatrix}$.

During sliding motion, equation (10) should satisfy $S_1 z_1(t) + S_2 z_2(t) = 0$, consequently,

$$z_2(t) = -S_2^{-1} S_1 z_1(t) = -M z_1(t) \ . \tag{11}$$

With assumption the matrix S_2 is nonsingular, where $M \in R^{m \times (n-m)}$ is defined as: $M = S_2^{-1} S_1$, using equation (11) replaces equation (8) can be given

$$\dot{z}_1(t) = (A_{11} - A_{12} M) z_1(t) \ . \tag{12}$$

The design requirement is system asymptotically stable, that is $z_1 \to 0$ as $t \to \infty$ during sliding motion. Now it becomes that of fixing $M = S_2^{-1} S_1$ to give $(n-m)$ negative poles to equation (12). Let $S_2 = I_m$, S will be determined as: $ST_r^T = \begin{bmatrix} M & I_m \end{bmatrix}$.

Several approaches have been proposed for design S including poles placement, quadratic minimization and eigenvalue assignment methods. We will discuss quadratic minimization method here.

Define a cost function: $J = \frac{1}{2} \int_{\tau}^{\infty} x(t)^T Q x(t) dt$

where Q is both symmetric and positive definite, the aim is to minimize it. Using transformation matrix T_r , then $T_r Q T_r^T = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$, where $Q_{21} = Q_{12}^T$.

Thus

$$J = \frac{1}{2} \int_{\tau}^{\infty} (z_1^T Q_{11} z_1 + 2 z_1^T Q_{12} z_2 + z_2^T Q_{22} z_2) dt \ . \tag{13}$$

In order to solve this problem it is desirable to express it in the form of the standard LQR. For the last two terms in equation (13) can be replaced to yield $2z_1^T Q_{12} z_2 + z_2^T Q_{22} z_2 = (z_2 + Q_{22}^{-1} Q_{21} z_1)^T \times \ Q_{22}(z_2 + Q_{22}^{-1} Q_{21} z_1) - z_1^T Q_{21}^T Q_{22}^{-1} Q_{21} z_1$, so equation (13) can be written as

$$J = \frac{1}{2} \int_{\tau}^{\infty} [z_1^T (Q_{11} - Q_{12} Q_{22}^{-1} Q_{21}) z_1 + \ (z_2 + Q_{22}^{-1} Q_{21} z)^T Q_{22} (z_2 + Q_{22}^{-1} Q_{21} z_1)] dt \ . \tag{14}$$

Define

$$\hat{Q} = Q_{11} - Q_{12} Q_{22}^{-1} Q_{21} \ . \tag{15}$$

$$P = z_2 + Q_{22}^{-1} Q_{21} z_1 \ . \tag{16}$$

Equation (14) can be written as

$$J = \frac{1}{2}\int_{\tau}^{\infty} (z_1^T \hat{Q} z_1 + P^T Q_{22} P) dt \ . \tag{17}$$

Using equation (16) replaces equation (8) to eliminate the z_2, it becomes

$$\dot{z}_1(t) = \hat{A} z_1(t) + A_{12} P \ . \tag{18}$$

where $\hat{A} = A_{11} - A_{12} Q_{22}^{-1} Q_{21}$

Now the problem becomes that of minimizing the function (17) subject to the equation (18). Defining a positive definite matrix P_1, and it must satisfy the algebraic Riccati equation: $P_1\hat{A} + \hat{A}^T P_1 - P_1 A_{12} Q_{22}^{-1} A_{12}^T P_1 + \hat{Q} = 0$, then the minimizing equation (17) is given by $P = -Q_{22}^{-1} A_{12}^T P_1 z_1$, to replace the expression into equation (16) yields

$$z_2 = -Q_{22}^{-1}(A_{12}^T P_1 + Q_{21}) z_1 \ . \tag{19}$$

Comparing this one with equation (11) yields the expression

$$M = Q_{22}^{-1}(A_{12}^T P_1 + Q_{21}) . \tag{20}$$

$$S = [M \quad I_m] T_r . \tag{21}$$

Thus $\sigma(X) = SX = [M \quad I_m] T_r X$, then the final control law is:

$$u = K_l X + K_s \beta \sqrt{\sigma}\, \mathrm{sgn}(\sigma) \ . \tag{22}$$

where $K_l = -(SB)^{-1} SA$, $K_s = -(SB)^{-1}$.

Let $u_L = K_l X$ is linear feedback control. $u_N = K_s \beta \sqrt{\sigma}\, \mathrm{sgn}(\sigma)$ is nonlinear feedback control, namely sliding mode control, thus

$$u = u_L + u_N \tag{23}$$

4 Simulation Results

The blimp trajectory control system is given in figure 2.

When the sliding mode controllers are applied one undesirable aspect is the chatter induced by the nonlinear switching term near the sliding surface. One way to reduce the chatter is to use the concept of a boundary layer around the sliding surface and use tanh function replace sgn function [7].

Renew define sliding surface functions as:

$$\dot{\sigma}_i = -\beta_i \sqrt{\sigma_i / \alpha_i}\, \tanh(\sigma_i / \alpha_i) \qquad \alpha_i, \beta_i > 0 \quad for \quad i = 1 \ \ to \ \ m$$

where α is boundary layer thickness. The method can reduce the chattering effect when the system is near the sliding surface.

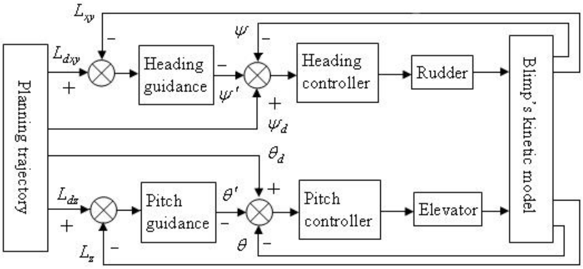
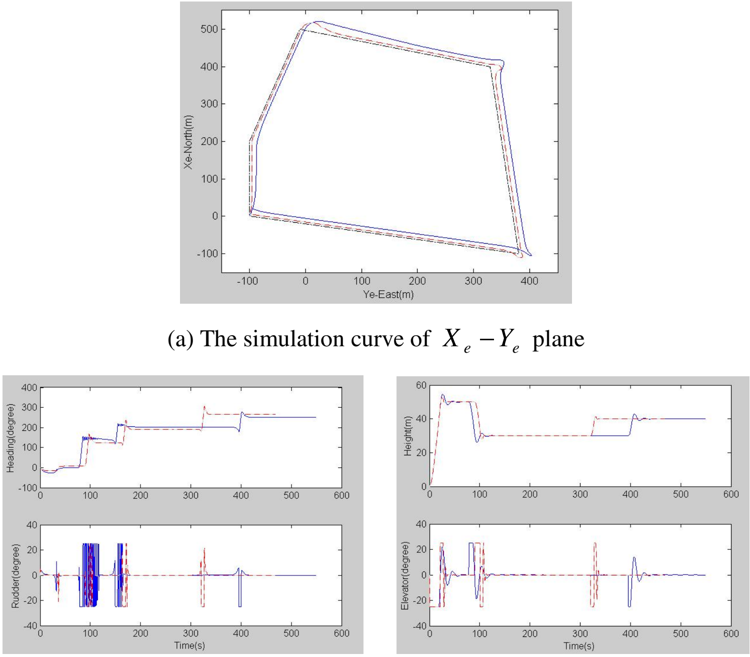


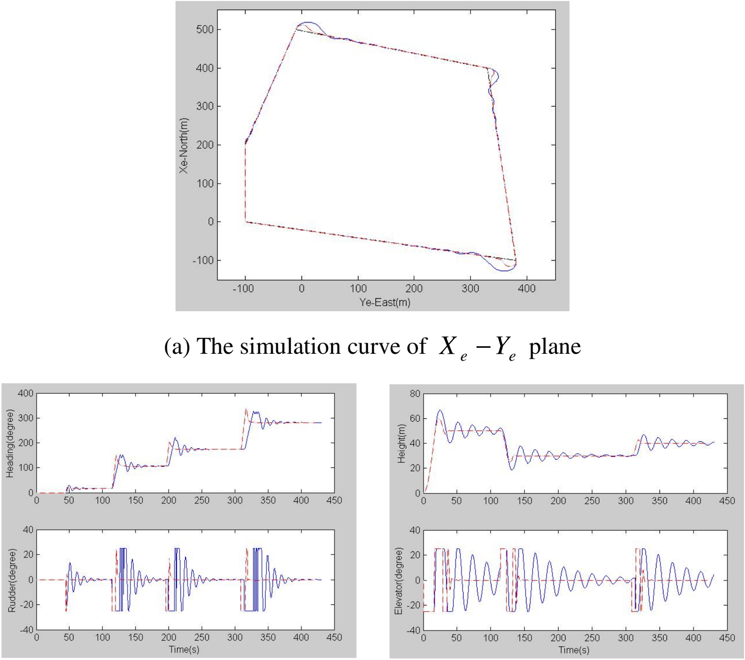
Fig. 2. Block diagram of blimp trajectory control system. It includes horizontal plane control loop and vertical plane control loop.



(a) The simulation curve of $X_e - Y_e$ plane

(b) The curve of heading and rudder (c) The curve of height and elevator

Fig. 3. The simulation results are with constant wind at 45 degrees to the X_e axis. (a) is projected to horizontal plane curve in earth-fixed reference frame; (b) is heading curve and controlled degrees of rudder; (c) is height curve and controlled degrees of elevator.



(a) The simulation curve of $X_e - Y_e$ plane

(b) The curve of heading and rudder

(c) The curve of height and elevator

Fig. 4. The simulation results are that all aerodynamic coefficients are increased by 100% from designed conditions. (a) is projected to horizontal plane curve in earth-fixed reference frame; (b) is heading curve and controlled degrees of rudder; (c) is height curve and controlled degrees of elevator.

The control law of lateral motions is:

(1) Sliding mode controller:

$$\sigma = 0.9866v - 1.1403r + 2.7085(\psi_{com} - \psi) + 0.11 tranking\_error$$

$$\delta_r = -0.5115v - 0.557r - 0.032\sqrt{2\sigma}\tanh(2\sigma)$$

(2) PD controller:

$$\delta_r = -1.3(\psi_{com} - \psi) + 4.3r + 0.07 tracking\_error$$

The control law of longitudinal motions is:

(1) Sliding mode controller:

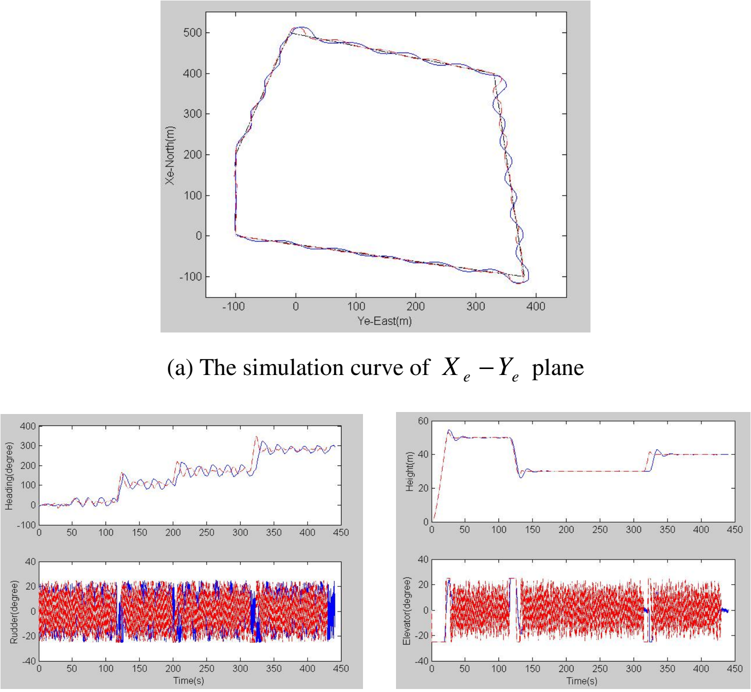
$$\sigma = 1.3148w - 1.1872q - 1.75595\theta + 0.7142(height_{com} - height)$$

$$\delta_e = -0.8532w + 0.6684q - 0.0101\theta + 0.0322\sqrt{2\sigma}\tanh(2\sigma)$$

(2) PD controller:

$$\delta_e = -0.04(height_{com} - height) + 0.09\theta + 4.4q$$

We have done a lot of simulation research on this issue. The followings are four simulation cases on blimp's trajectory. The case 1 is under designed conditions of PD control law; simulation results indicate the two types of controllers have similar perfo-



(a) The simulation curve of $X_e - Y_e$ plane

(b) The curve of heading and rudder (c) The curve of height and elevator

Fig. 5. The simulation results are with white noise of mean zero and variance 0.5. (a) is projected to horizontal plane curve in earth-fixed reference frame; (b) is heading curve and controlled degrees of rudder; (c) is height curve and controlled degrees of elevator.

rmance. The simulation results of case 2, case 3 and case 4 are shown in figure 3, figure 4 and figure 5, respectively. In figures, the red dashed lines denote the results by sliding mode controller; the blue solid lines show the results by PD controller; the black dash-dot lines is desired trajectory.

Figure 3 illustrates the simulation results with constant wind at 45 degrees to the X_e axis. The results indicate that the sliding mode controller has quite better robustness and precision than PD controller.

Figure 4 illustrates the simulation results that all aerodynamic coefficients are increased by 100% from designed conditions. The results indicate that the sliding mode controller has far better robustness and precision than PD controller.

Figure 5 illustrates the simulation results with white noise of mean zero and variance 0.5. The results indicate that the sliding mode controller has quite better robustness than PD controller.

5 Conclusions

In this paper, we discussed sliding mode controllers and PD controllers in controlling an unmanned blimp. For complex nonlinear movement of blimps we can design its sliding mode controllers through linearized models. The simulation results indicate that the sliding mode controllers can provide stronger robust performance to blimps maneuvering than PD controllers via comparing to the simulation curves of two types of controllers. Computer simulations have validated that sliding mode control method is to be suitable for the moving situations of nonlinear dynamic and uncertain models of blimps. The sliding mode controllers will be employed in further developing.

Acknowledgements

This work has been studied since I was in Advanced Intelligent Laboratory, School of Computer and Information Science, University of South Australia, in Mawson Lakes Campus as a visiting scholar, in 2003. The author wishes to acknowledge the Guidance of Dr Dan R Corbett. The author is also indebted to staffs of the School for the great help on my research work.

References

1. Azinheira, J. R., de Paiva, Ramos,E. C. J. J. G. and Bueno,S. S. :Mission Path Following for an Autonomous Unmanned Airship. IEEE International Conference on Robotics and Automation, pp. 1269-1275, 2000
2. de Paiva, E. C., Bueno, S. S., Gomes,S. B. Ramos,V. J. J. G. M. Bergerman:A Control System Development Environment for AURORA's Semi-Autonomous Robotic Airship. IEEE International Conference on Robotics and Automation, pp. 2328-2335, May 1999
3. Hygounenc, E. and Soueres, P.:Automatic airship control involving backstepping techniques.IEEE International Conference on System, Man and Cybernetics, Vol. 6, pp 1-6, 2002
4. Wells, N. :Practical Operation of Remotely Piloted airships.11th AIAA Lighter-than-Air Systems Technology Conference, 1995
5. Huang, Y. J. :Sliding Mode Control Design for Aircraft Control Systems.The First IEEE Regional Conference on Aerospace Control Systems Proceedings, pp309-313, 1993
6. Jafarov, E. M.,Tasaltin, R.:Robust sliding-mode control for the uncertain MIMO aircraft model F-18, IEEE Transactions on Aerospace and Electronic Systems, Vol 36, pp1127-1141, 2000
7. Healey, A. J. and Lienard, D. :Multivariable Sliding-Mode Control for Autonomous Diving and Steering of Unmanned Underwater Vehicles", IEEE Journal of Oceanic Engineering, vol.18,n.3, ,pp.327-339, 1993
8. Gomes and Ramos, J. Jr. G.:Airship Dynamic Modeling for Autononous Operation. IEEE International Conference on Robotics and Automation, pp 3462-3467, 1998
9. Ethin, B. :Dynamics of Flight —Stability and Control.John Wiley & Sons, 1982
10. Edwards, C. and Apurgeon, S. K: Sliding Mode Control Theory and Applications.Taylor & Francis Ltd, 1998

Real-Time Implementation of High-Performance Spectral Estimation and Display on a Portable Doppler Device

Yufeng Zhang[1,2], Jianhua Chen[1], Xinling Shi[1], and Zhenyu Guo[2]

[1] Department of Electronic Engineering, Information School, Yunnan University, Kunming, Yunnan Province, 650091, The People's Republic of China
yfengzhang@yahoo.com, {chenjh, shixl}@ynu.edu.cn
[2] Department of Electrical and Computer Engineering, The George Washington University, Washington, DC 20052, USA
zyguo@seas.gwu.edu

Abstract. In present study, the complex real-time autoregressive (AR) modeling based on the Levision-Durbin recursive algorithm and the maximum entropy spectral estimation is developed to calculate the spectrograms of Doppler blood flow signals for producing more comprehensive information about the components of the blood flow profile and increasing the display quality. A portable Doppler blood flow analysis device, which is based on a digital signal processor (DSP) TMS320VC549 (Texas Instruments) and contains a $240 * 320$ LCD color graphic display module (Hantronix), has been used to implement the spectral estimation algorithm and display spectrograms on the LCD in real-time. Directional Doppler spectrograms are computed directly from the in-phase and quardrature components of the Doppler signal. The results are applied to the development of a compact, economic, versatile bi-directional and battery- or line- operated Doppler device, which can be conveniently used to perform basic vascular testing.

1 Introduction

The Doppler ultrasound technique has been a widely used tool in clinical applications since the 1980s. Although Doppler ultrasound may be used for the study of various types of motion within the body, its major use remains the detection and quantification of the flow in arteries and veins. Diagnostic information is usually extracted from the Doppler spectrogram computed using the STFT. [1,2,3]. However the STFT has a main shortcoming in the trade-off between time and frequency resolution. The STFT algorithm requires that the signal being analyzed is stationary during a short time interval. By compromising between frequency resolution of the spectrum and the blood ejection rise time during systole, a 10 ms window is often used in practice. To increase the frequency resolution, a longer time interval is required. Thus, the stationary assumption may not be valid. In addition, the spectral components occurring in a large interval will be smeared in the time domain, resulting in decreased time resolu-

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 346 – 355, 2005.

tion. Thus, the STFT is not necessarily the best tool for analyzing Doppler blood flow signals [4]. To partly solve this problem, the AR modeling has been used as an alternative technique. Kitney and Giddens [5] stressed the best performance on spectral tracking and spectral resolution of autoregressive spectral estimation when short frames were used. Kaluzynski [6] reported the advantages of using the AR modeling for the analysis of pulsed Doppler signals, especially for short data lengths. Vaitkus *et al.* [7,8] addressed the good spectral matching ability of the AR modeling approach using a simulated stochastic stationary Doppler signal with a known theoretical spectrum as a reference to test spectral estimation techniques.

Conventionally, a Doppler blood flow analysis device is a piece of large-sized equipment. Its applications were limited due to the considerable weight/size of the table-top device placed on the wheel-table and the dependency on the wall power plugs. Therefore, a battery- operated portable Doppler system, which produces crisp, clear Doppler sound complemented with a display, is desired for the evaluation of the peripheral vasculature.

In the present study, we have developed a portable Doppler blood flow analysis device utilizing the complex real-time autoregressive modeling based on the Levision-Durbin recursive algorithm and the maximum entropy spectral estimation to calculate the spectrograms of Doppler blood flow signals. The objectives of our study are:

1) to design a small sized device based on a digital signal processor and a LCD graphic display module, make it suitable for flexible applications;

2) to implement the AR modeling algorithm in real-time for producing more comprehensive information about the components of the blood flow profile.

3) to display Doppler spectrograms in real-time on the LCD in 256 colors for increasing the display quality.

In the next section, we first describe briefly the mathematical background of the AR modeling algorithm used in this study. Then the rest of sections deal with the implementation of the hardware for the spectral estimation and display, the development of the Doppler signal processing algorithm, and the method of displaying the Doppler spectrograms on the LCD, followed by results, discussions and conclusions.

2 Mathematical Background of the AR Modeling Algorithm

Two audio signals, both containing the Doppler information, but shifted by $\pm 90^{o}$ to each other, are called the in-phase $x_i(n)$ and the quadrature $x_q(n)$ Doppler signals respectively [9]. The directional Doppler blood flow signal is a complex-valued signal expressed as

$$x\ (n) = x_i(n) + jx_q(n). \tag{1}$$

According to Guo *et al.* [10], it is possible to compute the forward and the reserve blood flow components directly from the complex Doppler signal using the complex AR modeling.

The complex AR modeling is expressed as

$$x(n) = \sum_{m=1}^{p} a_{p,m} x(n-m) + e(n). \tag{2}$$

where *x(n)* is the complex Doppler signal, *p* is the order of the model, a_{pm} is the complex coefficients, and *e(n)* is the modeling error. The Yule-Walker Equations together with the Levinson-Durbin algorithm [11] are used to compute the complex AR coefficients a_{pm} and the modeling error variance c_{p} . This algorithm proceeds recursively to compute the parameter sets { $a_{1,1}$, σ_1^2 }, { $a_{2,1}, a_{2,2}$, σ_2^2 }, ... , { $a_{p,1}$, $a_{p,2}$, ... , $a_{p,p}$, σ_p^2 }, as follows:

$$a_{1,1} = -R_{xx}(1) / R_{xx}(0). \tag{3}$$

$$\sigma_1^2 = (1 - \left|a_{1,1}\right|^2) R_{xx}(0). \tag{4}$$

The recursion for *m*= 2,3, ... , *p* are given by

$$a_{m,m} = -\left[R_{xx}(m) + \sum_{k=1}^{m-1} a_{m-1,k} R_{xx}(m-k) \right] \Big/ \sigma_{m-1}^2 . \tag{5}$$

$$a_{m,i} = a_{m-1,i} + a_{m,m} a_{m-1,m-i}^{*} \quad (i=1, \ldots, m-1). \tag{6}$$

$$\sigma_m^2 = (1 - \left|a_{m,m}\right|^2) \sigma_{m-1}^2 . \tag{7}$$

where $R_{xx}(m)$ is the complex auto-correlation function of *x(n)* computed by

$$R_{xx}(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n+m) x^{*}(n). \tag{8}$$

$x^{*}(n)$ is the complex conjugate of *x(n)*. *N* is the number of data points. Supposing that *p*+1 values of the auto-correlation function $R_{xx}(0)$ to $R_{xx}(p)$ are computed using (8), the maximum entropy extrapolation of the auto-correlation function is

$$R_{xx}(n) = -\sum_{k=1}^{p} a_{p,k} R_{xx}(n-k) \quad \text{for } |n| > p. \tag{9}$$

and then, the maximum entropy spectral estimation based upon the extrapolation of the known values of the auto-correlation function is given by[11]:

$$S(k) = \sum_{n=-M}^{M-1} R_{xx}(n) e^{-j\frac{2\pi k}{N}n} . \tag{10}$$

where *M* can be chosen as a power of two, in order to enable the computation be performed using a fast Fourier transform. We have chosen to compute 128 (*M*=128) frequency bins, largely as a matter of compatibility with the previous studies [5-10].

3 Hardware

The estimation and the display of the autoregressive spectrograms in real-time are implemented on a digital board based on a digital signal processor TMS320VC549 [12]. The configuration of the digital board for the spectral estimation and display is shown in Fig. 1. The hardware accommodates a TMS320VC549 DSP chip, address interface, SRAMs, FLASH ROMs, ADC circuit, the LCD display interface circuit and the JTAG emulator interface. This board includes 256K words of one wait-state data SRAM, and a total of 256K words of FLASH ROM. The generic array devices, two chips of GAL20V8, select the SRAM, FLASH ROM, or on board peripherals. In this configuration, the DSP can work at 80 MHz with two wait states when access the onboard SRAM.

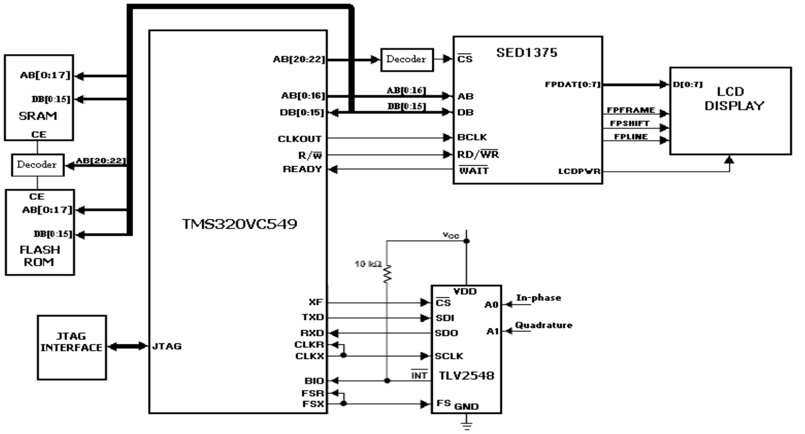


Fig. 1. The configuration of the digital board for spectral estimation and display

The TMS320VC549 synchronous serial port, whose interface consists of transmit and receive clocks, frame-sync and data lines, is used to access the onboard ADC TLV2548. The analog channels A0 and A1 are used for the in-phase and quadrature Doppler signal inputs.

Implemented by using an EPSON SED1375 [13] Embedded Memory Color LCD Controller, the Hantronix HDM3224C $240*320$ LCD color graphic display module with CCFL backlight is used in this portable device. The SED1375 is a color/monochrome LCD graphics controller with an embedded 80K Byte SRAM display buffer. The high integration of the SED1375 provides a low cost, low power, single chip solution to meet the requirements of the portable embedded device application.

A 14 pin standard interface, which is used by the JTAG emulator to interface to TMS320VC549 digital processor's On-Chip Scan-Based Emulation Logic (IEEE Std 1149.1† (JTAG) Boundary Scan Logic) [12], simplifies the code development and shortens the debugging time. Debuggers providing the assembly language and the high level C language debug are available with the JTAG emulator. Data processed by the DSP can be loaded to the host machine through this JTAG interface and stored as a data file.

4 Implementations

4.1 Spectral Estimation Implementation

The in-phase and the quadrature Doppler signals are digitized by the 12-bit ADC TLV 2548 with the sampling frequencies varied among 5 kHz, 10 kHz, 15 kHz, 20 kHz and 25 kHz according to the velocity measured. The TLV2548 device is configured as: single-shot conversion mode, conversion clock = SCLK, short sample period (12 SCLKs), 4V internal reference voltage. The single-shot conversion mode is a simple method, that is, convert/sample, and immediately read the data out of the data converter. Because of preventing the DSP from being used by other devices at the end of conversion, this mode is able to keep the sample frequency accuracy. The TLV2548 is programmed to generate an interrupt when the FIFO is filled. The digitized signals are then received by the TMS320VC549 digital signal processor to estimate the spectrum.

In order to simplify the code development and shorten the development time, a special data format (Q.15) and several assembly routines in the TI C54x DSPLIB [14] are used to perform the signal processing. The TI C54x DSPLIB is an optimized DSP function library for C programmers on TMS320C54x DSP devices. It includes over 50 C-callable assembly-optimized general-purpose signal processing routines. These routines are typically used in computationally intensive real-time applications where optimal execution speed is critical. The routines used in this study include:

cbrev (DATA *x, DATA *r, ushort n) --- complex bit reverse
cfft (DATA x, nx, short scale) --- forward complex FFT
acorr (DATA *x, DATA *r, ushort nx, ushort nr, type) --- auto-correlation.

The function *cbrev* bit-reverses the position of elements in the complex input signal *x* into the output vector *r*. Use this function in conjunction with the *cfft* routine to provide the correct format for the FFT input or output data. The function *cfft* com putes a radix-2 complex forward FFT of the *nx* complex elements stored in the vector *x* in bit-reversed order. The original content of the vector *x* is destroyed in the process. The *nx* complex elements of the result are stored in the vector *x* in normal-order. The function *acorr* computes the first *nr* points of the positive-side of the auto-correlation of the real vector *x* and stores the results in the real output vector *r*. The full-length autocorrelation of vector *x* will have 2**nx*–1 points with even symmetry around the lag 0 point (r[0]). The complex auto-correlation of the directional Doppler signal in (8) can be achieved by using the function *acorr*.

In this study, a special data format *DATA*, which is called Q.15, is used by the signal processing routines. Q.15 format places the sign bit at the leftmost binary digit, and the next 15 leftmost bits contain the fractional component. The approximate allowable range of numbers in Q.15 representation is (–1, 1) and the finest fractional resolution is $2^{-15} = 3.05 * 10^{-5}$.

The software, which is programmed using Texas Instruments' TI C54x assembly language [15] and TI C54x C language [16] can be divided into two main parts: the signal processing and the user interface. The main process is implemented as a main task loop. Interrupt-driven data acquisition is performed as a background task triggered by a sampling pulse generated by an embedded timer. Equation (10) is used to estimate the AR-based spectrum with a frequency increment $\Delta f = {}^{f_s}\!/_{128}$ Hz (f_s is the sampling frequency). The auto-correlation values $R_{xx}(0)$ to $R_{xx}(p)$ and the autoregressive coefficients { $a_{p,1}$, $a_{p,2}$, ..., $a_{p,p}$, σ_p^2 } are estimated using 64 sample values for each 10 ms time segment. Within one-second period, 100 spectra are computed.

In order to calculate the Doppler spectrogram in real-time, two data frame buffers are used to store the digitized Doppler audio signals. Each frame length is $2*256$ words. One frame buffer is used to store the current digitized signal acquired when the acquisition background task is triggered by the sampling pulse. As soon as this frame buffer is full, a new spectral estimation for the data in this frame will begin. At the same time, the signal sampling is still continuing and the digitized signal will be stored in another frame buffer.

4.2 Spectral Display Implementation

The SED1375 implements the 16-bit interface to the TMS320VC549 processor, which may operate in Chip Select, plus individual Read Enable/Write Enable for each word. The SED1375 is initialized to control the panel with the specifications: 320x240 color single passive LCD panel at 70Hz, Color Format 2, 8-bit data interface, 8 bit-per-pixel (256 colors) and 6 MHz input clock (CLKI).

The SED1375 Look-Up Table (LUT) consists of 256 indexed red/green/blue entries. Each LUT entry consists of a red, green, and blue component. Each component consists of four bits, or sixteen intensity levels. Any LUT element can be selected from a palette of 4096 ($16 * 16 * 16$) colors. The SED1375 works at Eight Bit-Per-

Pixel (256 colors) mode. In this mode, one byte of the display buffer represents one pixel on the display. When using the color panel, each byte of the display memory acts as and index to one element of the LUT. For example, a display memory byte with a value of 00h will display the color contained in the first LUT entry, while a display memory byte of FFh will display a color formed by the 256th LUT entry. In this way, the displayed color is obtained by taking the display memory value as an index into the LUT.

As soon as a frame of the spectrogram computed by the DSP using the AR modeling algorithm is completed, a subroutine service will be called immediately to display it on the LCD in 256 colors. In this subroutine, the DSP finds the square root of the power spectrum to reduce its dynamic range, and normalizes its value range from 0 to 255 corresponding to the 256 color elements in the LUT. The power spectral densities of the AR modeling results are sequenced on the timeline to plot a three dimensional sonogram on the LCD. The color scales of sonograms represent the power levels corresponding to the frequencies at each point of time. As the power level increases, the color tone of the sonogram goes into bright and as it diminishes, the color tone of the sonogram becomes dark. The display subroutine is developed in assembly language because of the tight speed requirements to deal with the spectra in real-time.

The Doppler spectral computation and display subroutines are called once a frame of data acquisition is completed in the main task loop. During the spectral computation and display period, the processor can still be interrupted by the programmable timer for the data acquisition of the next frame. Thus the time interval of the spectral computation and display performed by the DSP should be less than 10 ms, which is the window duration used to estimate the spectrum in this study because the signal is assumed to be stationary over this time segment.

5 Results and Discussion

Using this portable device, the complex AR modeling spectrograms based on the Levision-Durbin recursive algorithm and maximum entropy spectral estimation are computed and displayed on the color LCD in real-time. All programs are stored in the Flash ROM onboard, which is nonvolatile when the device is power-off. But the Flash ROM can not run at the DSP working frequency (80 MHz) because its access speed is too slow. The programs need to be moved to the high speed RAM in the digital signal processor chip from Flash ROM by an initial startup program when the device is powered on or reset.

Fig. 2 shows the typical Doppler spectrograms of a normal femoral artery displayed on the LCD (Here the 256 colors have been printed out as the 256 gray scale levels) based on the AR modeling with different model orders p. Comparing the spectrograms with different orders p=11, 15 and 19, the display quality has no significant difference.

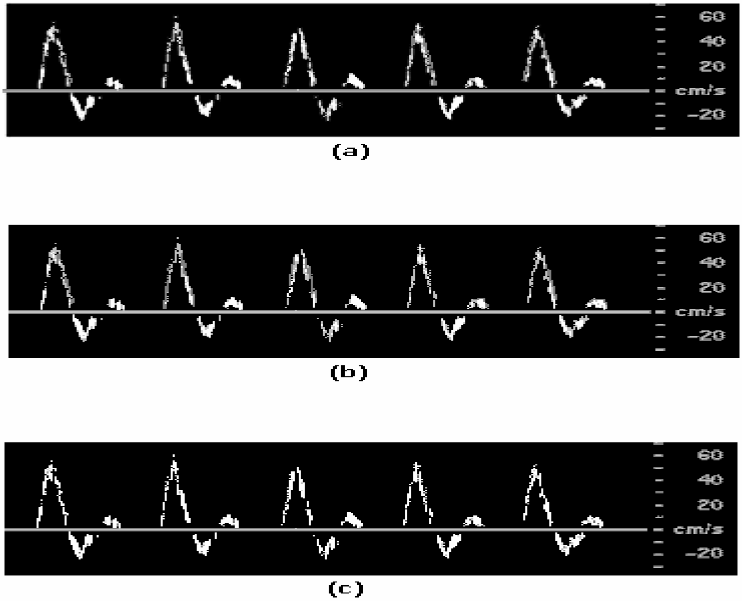


Fig. 2. The typical Doppler spectrograms of a normal femoral artery displayed on the LCD (The 256 colors have been printed out as the 256 gray scale levels here). They are estimated from the signal at sampling frequency 25 kHz based on AR modeling with (a) p=11. (b) p=15. (c) p=19.

The time interval for each interrupt service routine T_i (used to process the data acquisition) should be multiplied by the sample number N_s in a frame and added to the total processing time interval T_t when running in real-time. Thus

$$T_t = N_s * T_i + T_s + T_d. \tag{11}$$

where T_s is the spectral estimation time interval, and T_d is the time interval used to display the spectrogram with 128 frequency components in a frame. The total processing time interval T_t of individual frame (including the time intervals due to the subroutines for a frame of the data acquisition, the spectral estimation and display) based on the AR modeling with different model orders is listed in Table 1.

Table 1. The total processing time interval T_t of individual frame based on the AR modeling algorithm with different model orders

p	5	9	11	15	19
T_t (ms)	4.767	4.877	4.947	5.147	5.447

Corresponding to the sampling frequency range of up to 25 kHz, this portable Doppler blood flow analysis device can process the Doppler shift frequency range of up to 12.5 kHz. With the TMS320VC549 digital signal processor, it is possible to implement in real-time the Doppler spectrum computation based on the AR modeling of order up to 19 and display the results in 256 colors on the LCD module without losing data. In present implementation, when using the AR modeling to estimate the spectrum, 128 frequency components in a spectrum are calculated from 64 samples in a frame. This means that a four to one improvement on the STFT –based spectral estimation algorithm as far as stationarity of the data is concerned. This advantage means that it permits accurate estimation of frequency content from a much shorter segment of data and therefore the trade-off between temporal and spectral resolution is not as limiting as the STFT algorithm, and signals with rapid changes are more easily interpreted.

In this application, the TMS320VC549 digital signal processor produces and displays the spectra in real-time at a rate of 100 frames per second. However this DSP is capable of running the present application at a rate of about 180 frames per second when using the AR modeling algorithm with the order p=19 at sampling frequency 25 kHz. If a better temporal resolution is desired, more frames should be calculated per second. In this case, another digital signal processor chip with higher processing speed should be considered.

6 Conclusion

The complex real-time autoregressive modeling based on the Levision-Durbin recursive algorithm and the maximum entropy spectral estimation has been developed to calculate the spectrograms of Doppler blood flow signals. A portable Doppler blood flow analysis device, which is based on a digital signal processor TMS320VC549 and contains a $240 * 320$ LCD color graphic display module, has been used to implement the spectral estimation algorithm and display spectrograms on the LCD in real-time. The special data format (Q.15) and several assembly routines in the TI C54x DSPLIB are used to perform the signal processing to simplify the code development. The results have been applied to the development of a portable Doppler device. This battery- or line- operated portable Doppler system, which produces crisp, clear Doppler sound complemented with a display and print-out to make documentation for reimbursement fast and convenient, is easy and simple to use for the evaluation of the peripheral vasculature. It has the advantage of being high accurate spectral estimation, low-cost, and will be useful at different conditions, including bed-site hospitals and clinical offices.

References

1. Caro, C.E., Parker, K.H., Fish, P.J., Lever, M.J.: Blood Blow near the Arterial Wall and Arterial Disease, Clin Hemorheology. 5 (1985) 849–871
2. Cannon, S.R., Richards, K.L., Rollwitz, W.T.: Digital Fourier Techniques in the Diagnosis and Quantification of Aortic Stenosis with Pulse-Doppler Echocardiography, J. Clin. Ultrasound 10 (1982) 101–107

3. Kalman,P.G., Johnston,K.W., Zuech,P., Kassam,M., Poots,K.: In vitro Comparison of Alternative Methods for Quantifying the Severity of Doppler Spectral Broadening for the Diagnosis of Carotid Arterial Occlusive Disease, Ultrasound Med. Biol. 11 (1985) 435–440
4. Guo, Z., Durand, L.G., Lee, H.C.: Comparison of Time-frequency Distribution Techniques for Analysis of Simulated Doppler Ultrasound Signals of the Femoral Artery, IEEE Trans. Biomed. Eng. 41(1994) 332-342
5. Kitney, R.I., Giddens, D.P.: Analysis of Blood Velocity Waveforms by Phase Shift Averaging and Autoregressive Spectral Estimation, J. Biomech. Eng. 105 (1983) 398–401
6. Kaluzynski, K.: Analysis of Application Possibilities of Autoregressive Modeling to Doppler Blood Flow Signal Spectral Analysis, Med. Biol. Eng. Comput. 25 (1987) 373–376
7. Vaitkus, P.J., Cobbold, R.S.C.: A Comparative Study and Assessment of Doppler Ultrasound Spectral Estimation Techniques. Part I: Estimation methods, Ultrasound Med. Biol. 14 (1988) 661–672
8. Vaitkus, P.J., Cobbold, R.S.C.: A Comparative Study and Assessment of Doppler ultrasound Spectral Estimation Techniques. Part II: methods and results, Ultrasound Med. Biol. 14 (1988) 673–688
9. Evans, D.H., McDicken, W.N.: Doppler Ultrasound: Physics, Instrumentation and Signal Processing. Second Edition, JOHN WILEY&SONS, LTD (2000)
10. Guo, Z., Durand, L.G., Allard, L., Cloutier, G., Lee, H.C., Langlois, Y.E.: Cardiac Doppler Blood-flow Signal Analysis: Part 2 Time frequency representation based on autoregressive modeling, Med. & Biol. Eng. & Comput. 31 (1993) 242-248
11. Kay, S.M., Marple, S.L.: Spectrum Analysis—a Modern Perspective, Proc. IEEE, 69 (1981) 1380-1419
12. TMS320VC549 Datasheet, Literature Number: SPRS078F, Texas Instruments (2000)
13. SED1375 Technical Manual, Document No. X27A-Q-001-01, Epson (2000)
14. TMS320C54x DSPLIB User’s Guide, The Staff of the Texas Instruments C5000 DSP Software Application Group (2001)
15. TMS320C54x Assembly Language Tools User’s Guide, Literature Number: SPRU102C, Texas Instruments (1998)
16. TMS320C54x Optimizing C/C++ Compiler User’s Guide, Literature Number: SPRU103G, Texas Instruments (2002)

Application of Wavelet Transform in Improving Resolution of Two-Dimensional Infrared Correlation Spectroscopy

Daqi Zhan and Suqin Sun*

Department of Chemistry, Tsinghua University,
Beijing, 100084, China
sunsq@mail.tsinghua.edu.cn

Abstract. FTIR is a great improvement in IR spectroscopy, and two-dimensional infrared (2D IR) correlation spectroscopy well advances its capabilities. But for complicated mixture systems, such as traditional Chinese medicines, the spectra are rather similar and these methods fall short. To improve the resolution of 2D IR spectrum, and make it possible to distinguish complicated mixture systems, the application of wavelet transform to 2D IR was explored in this paper. After performing decomposition, processing, and reconstruction of the set of dynamic spectra, the resolution of the synchronous 2D IR correlation spectrum was improved obviously. More peaks appeared and the peaks became quite clear and separate. Four *Coptis* samples of aged 1-4 could be distinguished with this approach. Using wavelet transform, 2D IR would become more powerful in analysis and discrimination.

1 Introduction

Wavelet analysis has been a powerful analytical tool in many scientific and practical fields, such as quantum mechanics, signal analysis, image processing and so on. The main reason is that wavelet analysis has many particular characteristics. For example, it can analyze signals at different scales or resolutions, which is called multiresolution. And it also can localize signals in both frequency and time domains. Compared to Fourier analysis, which uses endless basic functions sine and cosine, wavelet analysis uses basic functions that are only nonzero in limited space or time. Therefore, wavelets can well express the signals that have non-stationary variations [1]. In Fourier translation infrared (FTIR) spectra, sharp peaks always exist, so in some ways, it would be better to process with wavelets than other methods.

FTIR really enhances the capabilities of IR spectroscopy. Signal noise ratio (SNR) is improved, so are the resolution and the speed of scan. However, no matter what have been done, the end point of a technique could never be reached. Some things are always waiting for study. For FTIR, the resolution is still expected to be enhanced so that more information could be obtained from the spectra. In recent years, a new tech-

* Corresponding author.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 356 – 365, 2005.

nology called two-dimensional (2D) correlation spectroscopy [2] attracts more and more people. Here, the 2D concept is some different from the one in nuclear magnetic resonance (NMR). It is called generalized two-dimension. It can base on many kinds of perturbations, like temperature, concentration, pressure, magnetic field, electric field, etc. [3]. This technology has been use in various fields, such as FTIR spectroscopy, fluorescence spectroscopy, and Raman spectroscopy and so on [4][5]. When used in FTIR, it is called 2D IR. Though 2D IR well improves the resolution of the IR spectrum, and was used to discriminate rather similar complicated mixture systems like traditional Chinese medicine [4], that is not enough. Sometimes the IR spectra of different samples are almost the same as each other, even in high resolution 2D correlation spectra. To distinguish these kinds of samples, it still requires enhancing the resolution.

As we know, *Coptis* is a common traditional Chinese medicine, and it is a kind of broadspectrum antimicrobials. Its effective components are Coptisine, Berberine, Epiberberine, Palmatine, and some other alkaloids. Different ages of *Coptis* have discrepancies in their functions. So it is very important to distinguish them. But the spectra of different ages *Coptis* are rather similar, it's difficult to distinguish them directly.

This paper is concerned with the application of wavelet transform to 2D IR, wish to enhance the performances of 2D correlation spectrum, and explore the possibility to distinguish the much similar complex compounds like *Coptis*.

1.1 Wavelet Transform

Wavelet transform is the development of Fourier transform with the following general definition:

$$W_{a,b}(f(t),\Psi)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\Psi(\frac{t-b}{a})dt\ . \tag{1}$$

Where a, $b \in R$ with $a \neq 0$, f is the signal function, ψ is the basic wavelet function. The parameter a is the dilation factor while b is the translation factor. An intuitive physical explanation of equation (1) is very simple: W is the 'energy' of $f(t)$ of scale a at $t=b$ [5]. Because of localized vibration, ψ looks just like a window function. This makes it possible to decompose a signal in given time domain or space domain. Hence, wavelet transform in some sense is the same as the linear combination of wavelets of different scales and different locations.

Using wavelet transform, signals can be decomposed into several levels. At each level, two sets of coefficients are obtained. One is for approximation coefficients and the other is for detail coefficients. The decomposition can be show in Fig.1. Original signal *S0* is decomposed in to two parts, one is the approximation *A1*, the other is the detail *D1*. Then, *A1* can be taken to decompose. In this way, original can be decomposed into specified levels.

After decomposition, the approximation and detail coefficients can be processed according to special need, such as signal denoising, image compression and so on. At last, to obtain the processed signal, reconstruction with the coefficients is needed. This is also called inverse wavelet transform [7], see equation (2).

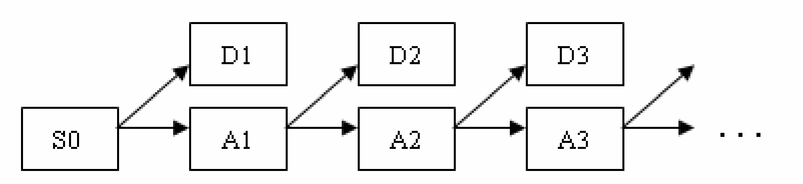


Fig. 1. Signal decomposition, *S0* is the original signal. *A*, *D* denotes approximation and detail, respectively.

$$f(x)=\frac{1}{C_{\Psi}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|a|^{-1/2}\,\Psi(\frac{x-b}{a})W_f(a,b)\frac{dbda}{a^2}\ . \tag{2}$$

Where,

$$C_{\Psi}=2\pi\int_{-\infty}^{\infty}\frac{|\hat{\Psi}(\lambda)|^2}{|\lambda|}d\lambda\ . \tag{3}$$

1.2 2D IR Correlation Spectrum

2D IR correlation spectroscopy was developed from generalized 2D correlation spectroscopy, whose concept was described in 1986 by Noda [8]. In common use, we should obtain a set of dynamic spectra first, then use statistical correlation to analyze the spectra. To get the synchronous and asynchronous 2D spectrum, we can use the following equations (4) and (5).

$$\Phi(\nu_1,\nu_2)=\frac{1}{m}\sum_{j=1}^{m}y_j(\nu_1)\bullet y_j(\nu_2)\ . \tag{4}$$

$$\Psi(\nu_1,\nu_2)=\frac{1}{m}\sum_{j=1}^{m}y_j(\nu_1)\sum_{k=1}^{m}N(j,k)y_k(\nu_2)\ . \tag{5}$$

Where $N(j, k)$ is the element of the Hilbert-Noda transformation matrix, when $j=k$, $N(j, k)=0$, otherwise, $N(j, k)=1/\pi(k-j)$. In equations (4) and (5), v is wavenumber. $y_j(v)= y(v, t_j)$ is set of dynamic spectra measured at m different equally spaced points along the external perturbation variable t [2].

Many interesting information can be obtained from the correlation maps. In synchronous spectrum, the peaks indicate the relative similarity of variation behavior of spectral intensities measured at two separate wavenumbers. In asynchronous spectrum, the peaks denote the dissimilarity of the intensity variation behavior [2]. This

2D spectrum can be used to study the mechanism of chemical reactions, distinguish complicated mixture system, and so on.

In this paper, the wavelet transform was operated in Matlab which is a widely used tool in scientific computing, and 2D IR spectrum was calculated by the software composed by author Zhan.

2 Experiment

2.1 Apparatus and Accessories

Spectrum 2000 GX FT-IR spectrometer (Perkin Elmer), equipped with a DTGS detector, in the 400-4000 cm^{-1} range with a resolution of 4 cm^{-1}. Spectra were calculated from a total of 16 scans.

Love Control Corporation's portable programmable temperature controller (Model 50-886). Range: 50-120℃.

2.2 Samples

Coptis of 1, 2, 3, 4-years old, all were identified and provided by Institution of Traditional Chinese Medicine of Hubei Province, China.

2.3 Procedure

First, all the *Coptis* were purified, comminuted, and desiccated. Then each sample of *Coptis* powder was blended with KBr powder, ground again, and pressed into a tablet. Put the sample tablet in the sample pool of temperature controller. A pre-established program controlled the whole process of increasing temperature. During the process of increasing temperature from 50 to 120 ℃, the spectra were collected at intervals of 10 ℃. The full temperature scan took a total time of 40 min.

2D IR correlation spectra were obtained by analyzing the set of temperature-dependent dynamic spectra with our 2D IR correlation analysis software.

3 Results and Discussions

3.1 Comparison and Discussion of Spectra Without Wavelet Transform

The 4 *Coptis* IR spectra, taken at room temperature, are shown in Fig. 2. The IR spectra belong to four *Coptis* samples of 1-4 years of age, respectively, arranged from the top down. Fig. 2 shows that the four spectra are rather similar. There are only a few slight differences between these spectra, like at 1077 cm^{-1}, two-year sample has a little peak, and the others have no obvious peak. Another difference is at 1622 cm^{-1}. But all this differences are not obvious, especially for people who are not professional in IR spectrum. In a word, it's difficult to distinguish them directly.

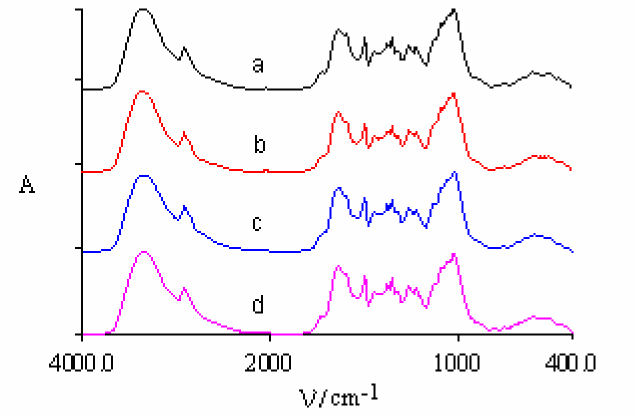


Fig. 2. IR spectra of four *Coptis* samples of different ages at room temperature. (*a*) one-year old, (*b*) two-year old, (*c*) three-year old, (*d*) four-year old.

Second derivative spectra can enhance the apparent resolution and amplify tiny differences of IR spectrum, but for these *Coptis* of different ages, second derivative spectra still looks deficient, see Fig. 3. These derivative spectra also have no obvious differences.

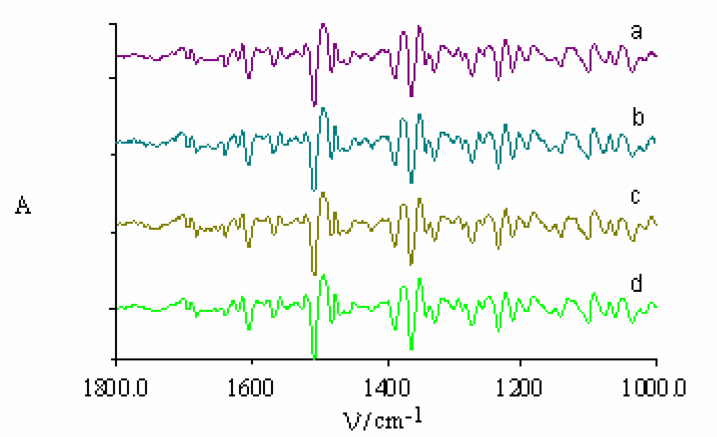


Fig. 3. Second derivative IR spectra of four *Coptis* samples of different ages. (*a*) one-year old, (*b*) two-year old, (*c*) three-year old, (*d*) four-year old.

Although 2D IR has been applied to distinguish similar complicated compounds like Panax and Fritillary [4]. Because it enhances the resolution of FTIR and can show the relative similarity and dissimilarity of variation behavior of spectral intensities measured at two separate wavenumbers. Here, when directly using 2D IR technique

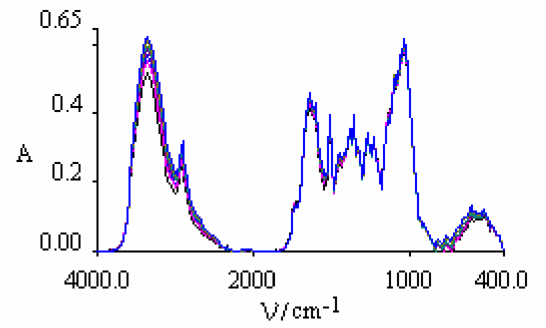


Fig. 4. Temperature-dependent dynamic FTIR spectra of one-year old *Coptis*

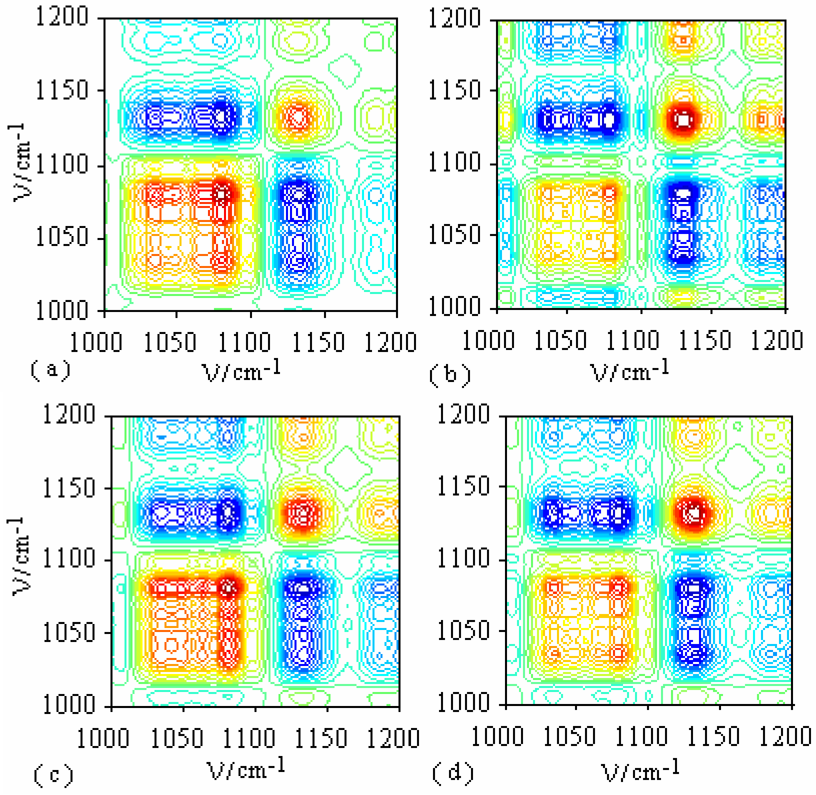


Fig. 5. Synchronous 2D IR correlation spectra of the *Coptis* samples. (*a*) one-year old, (*b*) two-year old, (*c*) three-year old, (*d*) four-year old.

to these four samples, we still can not get a satisfying result. Fig. 4 shows the set of temperature-dependent dynamic FTIR spectra of one-year old *Coptis*. The other samples have similar dynamic spectra. Fig. 5 shows the synchronous 2D IR correlation spectra of these four samples in region of 1000-1200 cm^{-1}.

Looking into Fig. 5, we can find some differences between these four *Coptis* samples. For example, at 1035 cm^{-1}, (*c*) has an obvious and a little strong autopeak (the peak on the diagonal). The other samples each have an autopeak at this position too, but the one of (*b*) is weak and not evident. Even so, the differences are so slight to distinguish these four *Coptis* samples by the methods mentioned above.

3.2 Comparison and Discussion of Spectra with Wavelet Transform

In recent years wavelet transform is widely studied and used. The marriages of wavelet transform and other sciences produce many exciting products. Though wavelets can well express the signals that have non-stationary variations, sometimes different wavelet basic functions result quite differently in the same matter.

In this paper, all operations of wavelet transform were performed under Matlab. After comparison with other wavelets, Symlet wavelets with order 8 (sym8 for short in Matlab) was chosen.

To examine the capability of decomposition and reconstruction, first we use sym8 to decompose a spectrum signal at level 5, and then reconstruct it directly, then observe the correlation between the original signal and reconstructed signal, we found that these two signals is perfectly equivalent, the correlation factor is 1.0000. Thus it can be seen that wavelet sym8 can well express IR spectrum.

In IR spectrum, all information comes from the position and intensity of the peaks. To enlarge the differences between the *Coptis*, we applied wavelet transform to each spectrum. Firstly, the spectrum signal was decompose into level 5, secondly, the detail and approximation coefficients at each level were extracted, and thirdly, these coefficients were processed. In this paper, the detail coefficients at level 3, level 4 and level 5 were multiplied by 2.5, which was to make the peaks in IR spectrum clear, obvious, and separate. See Fig. 6. At last, signal was reconstructed using the processed coefficients. The main program commands dealing with wavelet transform are as follows.

```
[c,l] = wavedec(s0,5,'sym8');
a5 = wrcoef('a',c,l,'sym8',5);
d1 = wrcoef('d',c,l,'sym8',1);
d2 = wrcoef('d',c,l,'sym8',2);
d3 = wrcoef('d',c,l,'sym8',3);
d4 = wrcoef('d',c,l,'sym8',4);
d5 = wrcoef('d',c,l,'sym8',5);
s  = d1+d2+d3*2.5+d4*2.5+d5*2.5+a5;
```

After wavelet transform process, as shown in Fig. 6, the peaks in the IR spectrum are much obvious and stood out. The weak peaks were enlarged, and the peaks joined together were separated noticeably.

For synchronous spectra, wavelet transform bring them many changes, which are shown in Fig. 7. Comparing these two maps, it's apparent that the resolution is enhanced. The autopeaks at 1035cm^{-1}, 1081 cm^{-1}, 1133 cm^{-1} become more legible. The off-diagonal peaks called cross peaks are also exciting. For example, the ones at

(1100, 1035), (1100, 1080), (1100, 1132), and (1185, 1132) are very difficult to identify in the (*a*) map, but in (*b*) map, these peaks are very clear and legible. As we known, in spectrum analysis, most of the information comes from peaks. Therefore, the more peaks appear, the more information could be obtained. This makes it possible to distinguish these four *Coptis* samples.

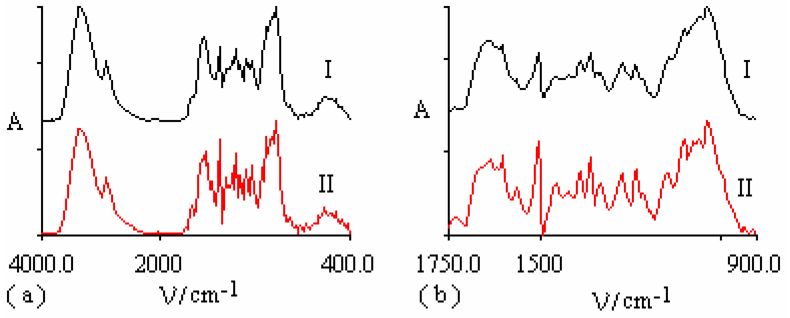


Fig. 6. Comparison of original IR spectrum and the one with wavelet transform processing. (*a*) region of 4000-400cm^{-1}, (*b*) region of 1750-900 cm^{-1}. (*I*) is original spectrum, (*II*) is processed spectrum.

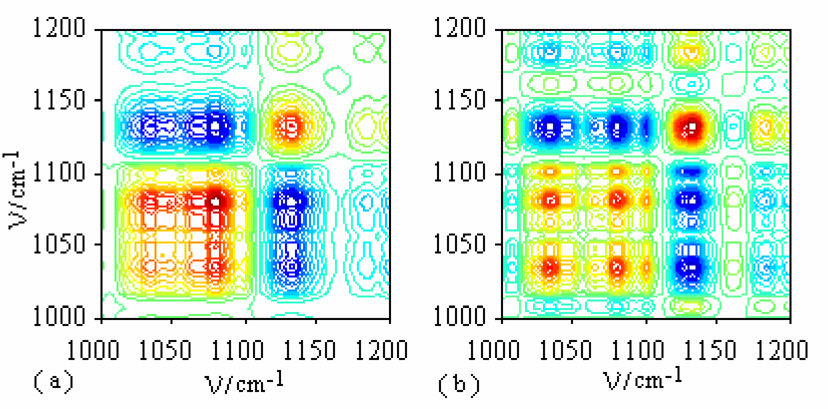


Fig. 7. Comparison of synchronous 2D IR spectra of one-year old *Coptis* without and with wavelet transform processing

Fig. 8 (a-d) are the synchronous spectra of the four samples with wavelet transform processing. It's obvious that all these maps are in higher resolutions than those in Fig. 5. More peaks appear in the spectra. The peaks become apparent and well separated from other peaks.

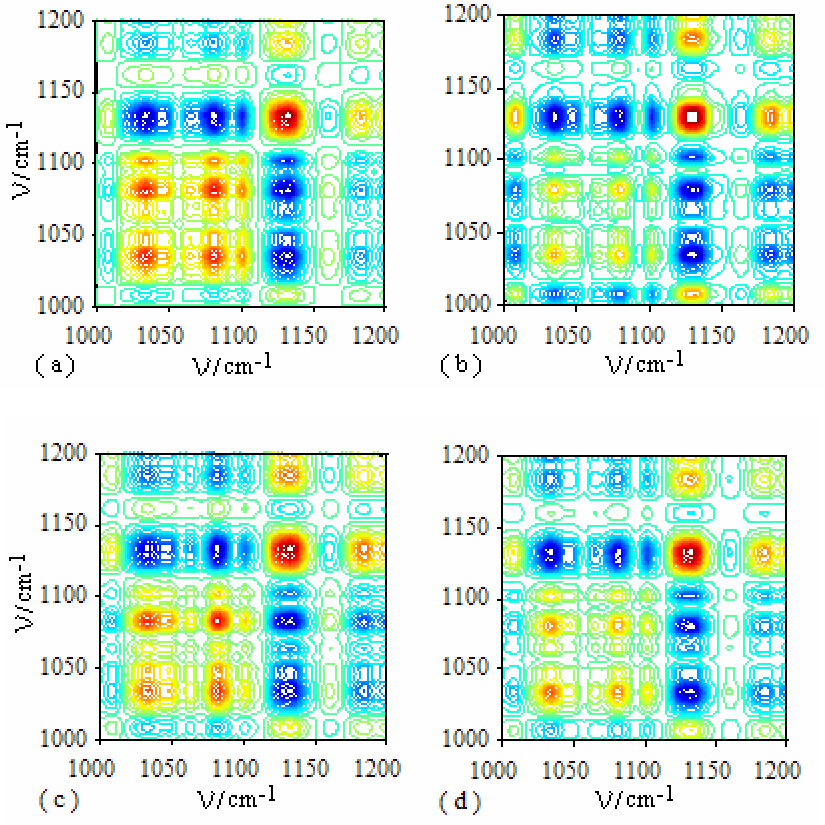


Fig. 8. Synchronous 2D IR correlation spectra of the *Coptis* samples with wavelet transform processing. (*a*) one-year old, (*b*) two-year old, (*c*) three-year old, (*d*) four-year old

It's quite easy to distinguish spectrum (*b*) from others, because only the autopeak at 1032 cm^{-1} is strong, the other autopeaks in spectrum (*b*) are quite weak. This is different with the other three spectra. In spectrum (*a*), the autopeaks at 1035 cm^{-1} and 1080 cm^{-1} are quite strong, and the cross peaks at (1100, 1035), (1100, 1080), and (1100, 1132) are obvious, but in spectra (*c*) and (*d*), these cross peaks do not exist, or very weak. These make spectrum (*a*) separated with spectra (*c*) and (*d*). And in spectrum (*c*), the autopeak at 1080 cm^{-1} is quite strong, but the one at 1035 cm^{-1} is much weaker than the autopeak at 1080 cm^{-1}. And in spectrum (*c*), there is a cross peak at (1048, 1033), which doesn't exist in spectrum (*d*). And in spectrum (*d*), the autopeak at 1035 cm^{-1} is stronger than the autopeak at 1080 cm^{-1}, that is opposite to spectrum (*c*). So spectrum (*c*) and spectrum (*d*) could be differentiated.

In a word, with the help of wavelet transform, these four similar *Coptis* of different ages were distinguished from each other.

4 Conclusions

This study showed that after applying wavelet transform to 2D IR correlation spectroscopy, the resolution of the synchronous spectrum was improved noticeably. More peaks appeared in the spectrum, and the peaks became clear and separate. In this way, the different ages *Coptis* samples could be distinguished while the spectra of conventional IR, second derivative IR and 2D IR without wavelet transform for these samples were rather similar and hardly to be differentiated. Therefore, with this method, the IR technique could be better used in analysis and discrimination of complicated mixture systems.

Acknowledgements

The work described in this paper was substantially supported by a grant from the Ministry of Science and Technology of the People's Republic of China (2002BA906A29-4).

References

1. Albert Boggess, Francis J. Narcowich: A Fist Course in Wavelets with Fourier Analysis. Publishing House of Electronics Industry (2004) 6-10
2. Noda,I.: Advance in Two-Dimensional Correlation Spectroscopy. Vibrational Spectroscopy 36 (2004) 143-165
3. Nada,I.: Appl. Spectrosc. 47 (1993) 1329
4. Rui Hua, Suqin Sun, Qun Zhou, et al.: Discrimination of Fritillary According to Geographical Origin with Fourier Transform Infrared Spectroscopy and Two-Dimensional Correlation IR Spectroscopy. Journal of Pharmaceutical and Biomedical Analysis 33 (2003) 199-209
5. Ma A, Stratt RM: Selecting the Information Content of Two-Dimensional Raman Spectra in Liquids. Journal of Chemical Physics 119 (2003) 8500-8510
6. Jianping Li, Shang-An Yan, Yuanyan Tang: The Application of Wavelet Analysis Method to Civil Infrastructure Health Monitoring. WAA 2001, LNCS 2251 (2001)393-397
7. Stephane Mallat: A Wavelet Tour of Signal Processing. China Machine Press (2003) 79-82
8. Noda,I., Bull: Am. Phys. Soc. 21 (1986) 520

Intelligent PID Controller Tuning of AVR System Using GA and PSO

Dong Hwa Kim and Jin Ill Park

Dept. of Instrumentation and Control Eng., Hanbat National University,
16-1 San Duckmyong-Dong Yuseong-Gu, Daejon City, Korea 305-719
kimdh@hanbat.ac.kr

Abstract. This paper deals with applying Euclidian data distance based GA-PSO (Genetic-Particle Swarm Optimization) algorithm (EU-GA-PSO) to PID controller tuning of AVR (Automatic Voltage Regulator). Through this approach, global and local optimal solution can simultaneously achieved for tuning of controller parameter.

1 Introduction

The Proportional-Integral-Derivative (PID) controller has been widely used owing to its simplicity and robustness in chemical process, power plant, and electrical systems. Its popularity is also due to easy implementation in hardware and software. However, with only the P, I, D parameters, it is very difficult to control a plant with complex dynamics, such as large dead time, inverse response, and a highly nonlinear characteristic in power plant. That is, since the PID controller is usually poorly tuned, a higher of degree of experience and technology are required for the tuning in the actual plant: [1], [2]. On the other hand, in the last decade, evolutionary based approaches have received the increasing attention of engineers dealing with problems not amenable to existing design theories. A typical task of a GA in one of artificial intelligence in this context is to find the best values of a predefined set of free parameters associated to either a process model or a control vector. A recent and thorough survey of evolutionary algorithms for evaluation of improved learning algorithm, control system engineering can also be found in [2], [3].

This paper focuses on dealing with an enhanced optimal tuning of PID controller for AVR (Automatic Voltage Regulator) using hybrid system consisted of PSO (Particle Swarm Optimization) and GA (Genetic Algorithm). To obtain an advanced learning structure, there are two processing steps in our operation. In the first step, Euclidean data distance is used to select global data on crossover and mutation operator to avoid local optimization, and obtain fast running time. That is, Euclidean data distance based mutation and crossover on GA's differentiation is operated. In the second step, in order to enhance learning efficiency of GA, PSO is applied. A PSO moves through the problem space, with the moving velocity of each particle represented by a velocity vector. Therefore, global and local optimal solution can

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 366 – 375, 2005.

simultaneously be achieved and the most appropriate parameter of PID controller can be selected for the given plant and system during operating.

2 Advanced GA Using Euclidean Distance and PSO

2.1 Euclidean Data Distance

When individuals in GA is differentiated to search optimal solution, since only near data can effect optimal solution, convergent time is fast but local optimization or suboptimal value can be obtained. In this paper, data which has the longest Euclidean distance as well as near data on data set (individuals) can also allow to affect crossover process to avoid this local optimization or suboptimal value. Fig. 1 is definition of Euclidean distance for GA and initial condition for learning of function is given as Table 1. In Table 1. ID and IT means the number of individuals, iteration, respectively. Table 2 is results obtained by Min and Max in test function

$$F_1(x) = \sum_{i=1}^{3} x_i^2$$

2.2 GA with Euclidean Data Distance

In this paper, GA with Euclidean distance is introduced into process of crossover and mutation of GA. Namely, it allow parent's individuals with the longest Euclidean distance to select in the processing of operation of GA. Using this method, because all data can have an effect on searching for optimal solution, we can avoid a localsolution or suboptimal and it is possible to obtain the global and exact solution for tuning of controller. Where, the distance between two points on *n* space is defined by

$$\text{distance}=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}\ . \tag{1}$$

Selection method of dada set is defined as

$$A(x_1, y_1) \oplus B(x_1, y_1) \Rightarrow A', B'\left(x_1{}'|_{\min(x_1,y_1)}^{\max(x_1,y_1)}, y_1{}'|_{\min(x_1,y_1)}^{\max(x_1,y_1)}\right). \tag{2}$$

For performance comparison of GA with Euclidean data distance, the Himmelblau function is used as test equation:

$$F(x) = \left(x_1^2 + x_2 - 11\right)^2 + \left(x_1 + x_2^2 - 7\right)^2 . \tag{3}$$

Fig. 2 illustrates relationship between objective function and generation by a GA. The bigger Euclidean distance, the faster divergence time in optimal value of objective function but the smaller Euclidean distance, the faster in the average value of objective function.

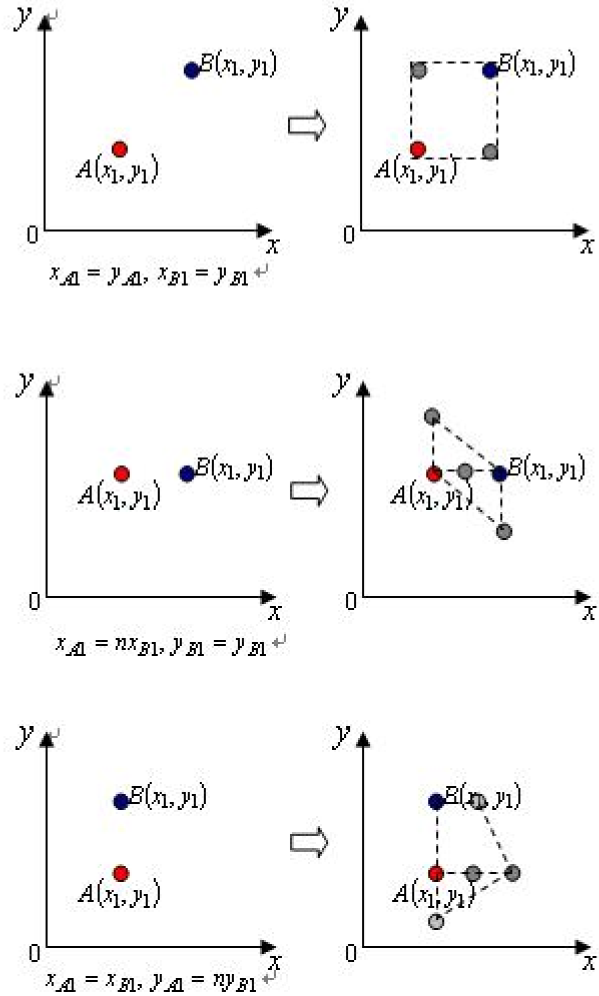


Fig. 1. Definition of Euclidean distance for GA

Table 1. Initial condition for performance test

Function	Definition		ID	IT
	$x_i^{(L)}$	$x_i^{(U)}$		
$F_1(x) = \sum_{i=1}^{3} x_i^2$	-5.12	5.11	60	100

Table 2. Result by Min-Max

	x1	x2	Optimal value of objective function	Average value of objective function
Max	1.0885e-009	7.1709e-010	1.6991e-018	3.5601e-013
Min	-2.2190e-011	1.0253e-009	1.0518e-018	3.7901e-013

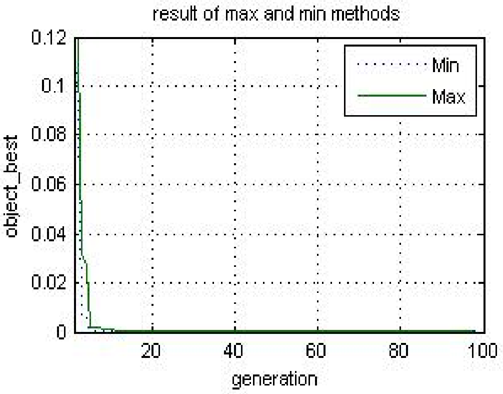


Fig. 2. Optimal value of objective

2.3 Overview of PSO

The PSO conducts searches using a population of particles which correspond to individuals in GA. A population of particles is randomly generated initially. Each particle represents a potential solution and has a position represented by a position vector. A swarm of particles moves through the problem space, with the moving velocity of each particle represented by a velocity vector. At each time step, a function representing a quality measure is calculated by using the results of crossover and mutation as input. Each particle keeps track of its own best position, which is associated with the best fitness it has achieved so far in a vector. At each time step, by using the individual best position, and global best position, the flexibility of PSO to control the balance between local and global exploration of the problem space helps to overcome premature convergence of elite strategy in GA, and also enhances searching ability. Of course, GA with Euclidean data distance is used for hybrid system with PSO.

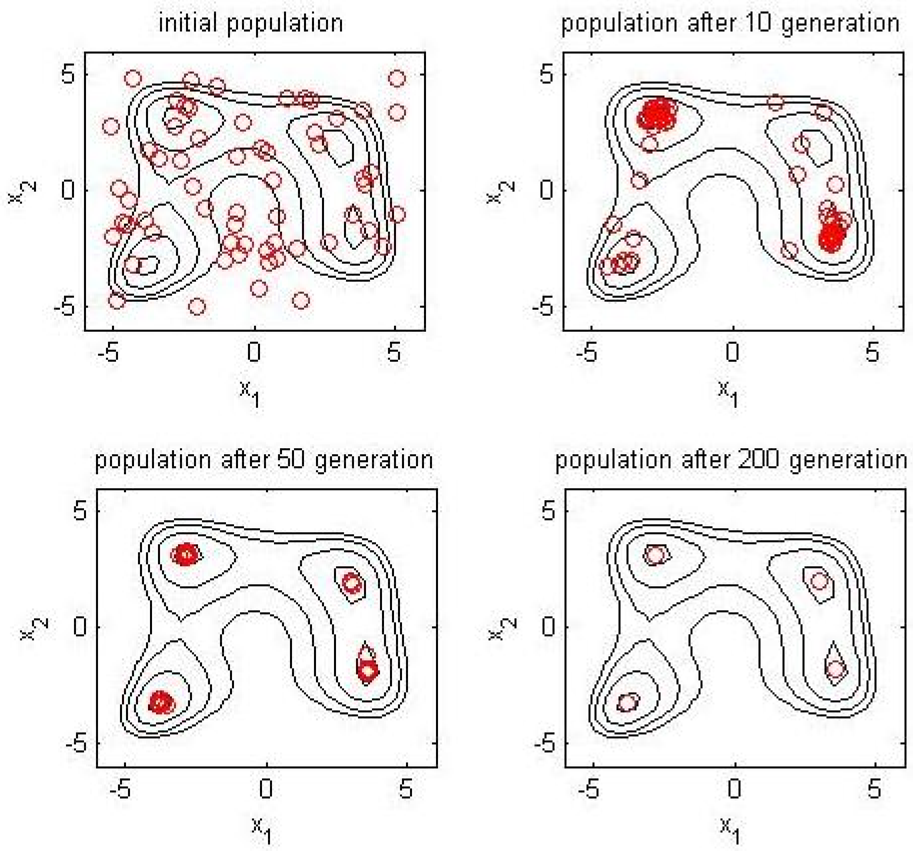


Fig. 3. Contour of GA obtained by Euclidean distance

2.4 Improvement of GA by PSO and Euclidean Distance (EU-GA-PSO)

The characteristic of hybrid system of PSO and GA have been studied [3-5]. This paper focuses on hybrid system using GA and PSO based on Euclidean distance. Position and speed vector of PSO is given by

$$\begin{aligned} v_{f,g}^{(t+1)} &= w\cdot v_j^{(t)} + c_1^* rand(\,)^* \left(pbest_{j,g} - k_{j,g}^{(t)}\right) + c_2^* Rand(\,)^* \left(gbest_g - k_{j,g}^{(t)}\right) \\ j &= 1,2,\ ...,n;\quad g = 1,2,...,m;\quad k_{j,g}^{(t+1)} = k_{j,g}^{(t)} + v_{j,g}^{(t+1)};\quad k_g^{\min} \le k_{j,g}^{(t+1)} \le k_g^{\max} \end{aligned} \qquad (4)$$

n : The number of agent in each group, m : The number of member in each group, t : Number of reproduction step, $v_{j,g}^{(t)}$: The speed vector of agent *j* in reproduction step of t^{th}. $V_g^{\min} \le v_{j,g}^{(t)} \le V_g^{\max}$, $k_{j,g}^{(t)}$: The position vector of agent *j* in reproduction step of t^{th}, w : Weighting factor, $c1, c2$: Acceleration constant, $rand(), Rand()$:

Random value between 0 and 1, $pbest_j$: Optimal position vector of agent j , $gbest$: Optimal position vector of group.

The value of position vector and speed vector is determined by acceleration constant $c1, c2$. If these values are large, each agent moves to target position with high speed and abruptly variation. If vice versa, agents wander about target place. As weighting factor w is for the searching balance of agent, the value for optimal searching is given by

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{iter_{\max}} \times iter . \tag{5}$$

where $w_{\max}$: Max mum value of w (0.9), $w_{\min}$: Minimum value of w (0.4), $iter_{\max}$: The number of iterative number, $iter$: The number of iterative at present.

The speed vector is limited by $V_g^{\min} \leq v_{j,g}^{(t)} \leq V_g^{\max}$. In this paper, the value of speed vector for each agent is limited with 1/2 to avoid abrupt variation of position vector. Computing process for each step is as the following step.

[Step 1] Initialize each variables of GA; [Step 2] Initialize each variables of PSO.;

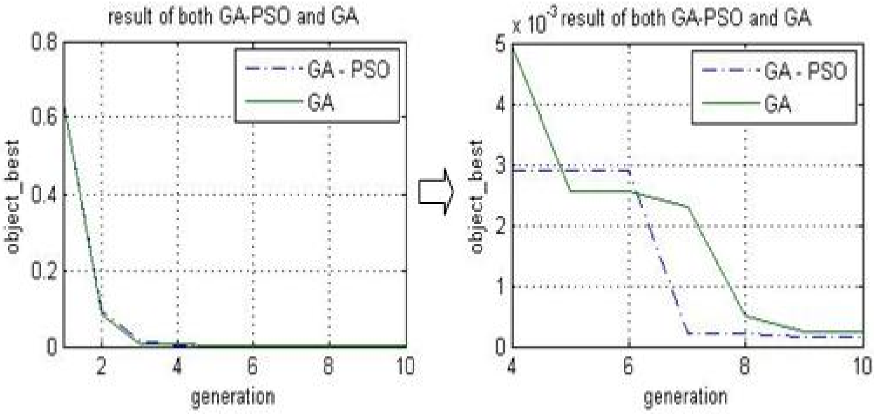


Fig. 4. Comparison between the conventional GA and the GA-PSO system

[Step 3] Calculate affinity of each agent for condition of optimal solution of GA. At this point, optimal position condition of PSO is introduced into GA; [Step 4] Arrange the group of PSO and agent in GA; [Step 5] Update position vector $pbest$ and speed vector $gbest$; [Step 6] Operate crossover in GA using Euclidian distance and position vector PSO; [Step 7] Operate mutation in GA; [Step 8] If condition of GA is satisfied with target condition (the number of iteration and target value), processing for reproduction stops. In this paper, at first time, position of individual on data set is calculated by Euclidean distance based method and then mutation and crossover is performed to improve running speed on optimal process. Therefore, it is able to obtain global optimal solution because all data effect solution. Fig. 3. is contour of GA obtained by Euclidean distance and Fig. 4 represents relationship between objective function and generation to the number of particle in PSO.

3 Simulation and Discussion

3.1 AVR System and PID Controller

The performance index of control response in AVR system of Fig. 5 is defined by

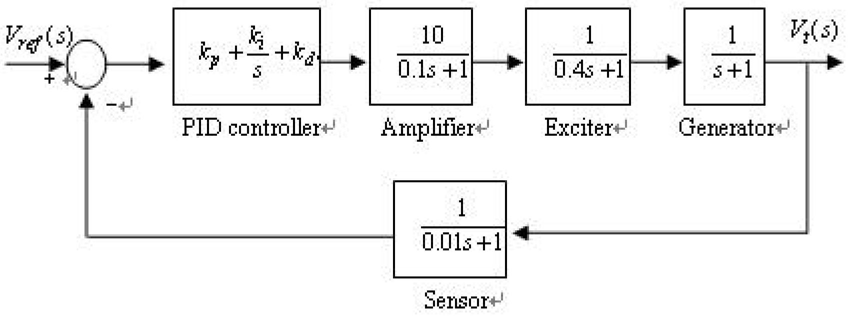


Fig. 5. Block diagram of an AVR system with a PID controller

$$\begin{aligned} \min F(k_p,k_i,k_d) &= \frac{e^{-\beta}\cdot t_s/\max(t)}{\left(1-e^{-\beta}\right)\cdot\left|1-t_r/\max(t)\right|} + e^{-\beta}\cdot Mo + ess \\ &= \frac{e^{-\beta}\cdot\left(t_s+\alpha_2\cdot\left|1-t_r/\max(t)\cdot Mo\right|\right)}{\left(1-e^{-\beta}\right)\cdot\left|1-t_r/\max(t)\right|} + ess \quad . \\ &= \frac{e^{-\beta}\cdot\left(t_s/\max(t)+\alpha\cdot Mo\right)}{\alpha} + ess \end{aligned} \tag{6}$$

$\alpha = \left(1-e^{-\beta}\right)\cdot\left|1-t_r/\max(t)\right|$, k_p, k_i, k_d : Parameter of PID controller, β : Weighting factor, Mo : Overshoot, t_s : Settling time (2%), ess : Steady-state error, t : Desired settling time.

In equation (3), if the weighting factor, β increases, rising time of response curve is small, and β decreases, rising time is big. Performance criterion is defined as $Mo = 50.61\%$, $ess = 0.0909$, $t_r = 0.2693(s)$, $t_s = 6.9834(s)$.

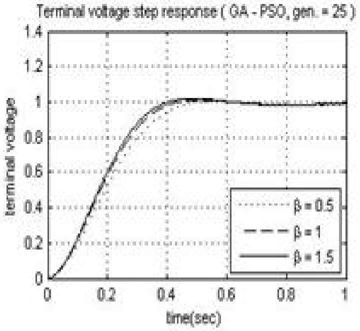


Fig. 6. Terminal voltage step response of an AVR system with the GA-PSO PID controller

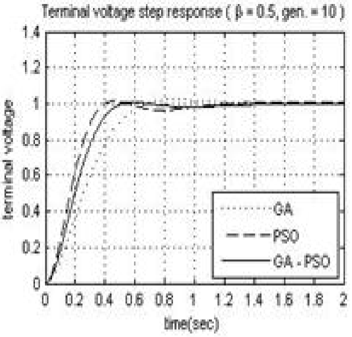


Fig. 7. Terminal voltage step response of an AVR system with controllers ($\beta = 0.5$, generations=10)

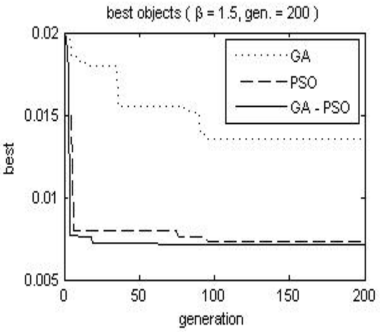


Fig. 8. Comparison of the best objects values of both methods($\beta = 1.5$, generations =200)

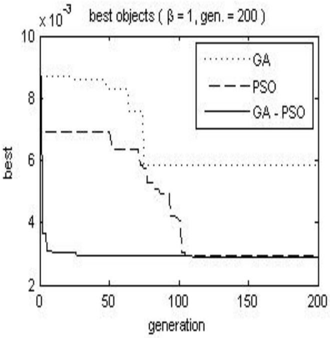


Fig. 9. Comparison of the best objects Values of both methods($\beta = 1$, generations =200)

3.2 Simulation Results

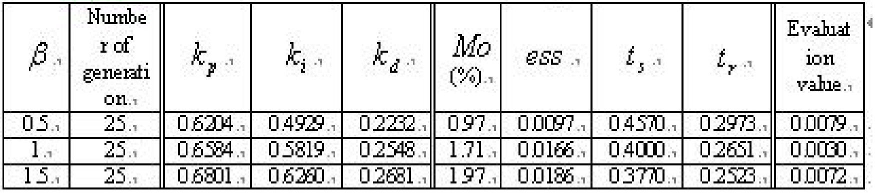
Simple crossover and dynamic mutation of GA is used and the number of individuals is 50, 200, and initial value of crossover and mutation are 0.6, 0.5, respectively. In PSO, parameters, k_p , k_i , k_d is defined as member of each agent and the number of each agent is 10. The number of group is 5. Here, weighting factor: $w_{\max} = 0.9, w_{\min} = 0.4$, restriction of velocity vector: $V_{k_p}^{\max} = k_p^{\max}/2, V_{k_i}^{\max} = k_i^{\max}/2, V_{k_d}^{\max} = k_d^{\max}/2, \quad V_{k_p,k_i,k_d}^{\min} = -V_{k_p,k_i,k_d}^{\max}$, acceleration constant. Terminal voltage step response of an AVR system with controller is given as Figs. 6 and 7. Fig. 8 is comparison of the best objects values of both methods ($\beta = 1.5$, generations=200) and Fig. 9 represents comparison of the best objects values of both methods ($\beta = 1$, generations=200).

4 Conclusion

Even though several control theories have been developed significantly, the proportional-integral-derivative (PID) controllers have been widely used owing to their simple structure, which can be easily understood and implemented for a wide range of process control, motor drives, flight control, and instrumentation.

Owing to their popularity in the industrial world, over the past 50 years, several approaches for determining PID controller parameters have been developed for stable processes that are suitable for auto-tuning and adaptive control and for single input single output (SISO) systems. In spite of the enormous amount of research work reported in the tuning approaches, many PID controllers are poorly tuned in practice. One of the reasons is that most of the tuning methods are derived for particular processes and situations, and therefore apply well only to their own areas. It is a common experience that we are not certain which tuning method should be chosen to provide good control to a given process. Intelligent controllers can have self-initialisation and recalibration features to cope with little a priori knowledge and significant changes in the process dynamics.

Table 3. Simulation results of PID controller in AVR system to β variation



β	Number of generation	k_p	k_i	k_d	Mo (%)	ess	t_s	t_r	Evaluation value
0.5	25	0.6204	0.4929	0.2232	0.97	0.0097	0.4570	0.2973	0.0079
1	25	0.6584	0.5819	0.2548	1.71	0.0166	0.4000	0.2651	0.0030
1.5	25	0.6801	0.6260	0.2681	1.97	0.0186	0.3770	0.2523	0.0072

On the other hand, the particle swarm is an algorithm for finding optimal regions of complex search spaces through interaction of individuals in a population of particles. Though the algorithm, which is based on a metaphor of social interaction, has been shown to perform well.

This paper deals with applying Euclidian data distance to controller tuning of AVR. In our algorithm, there are two steps, In the first step, Euclidean data distance is used to select global data. A function representing a quality measure is calculated by crossover to avoids local optimal, and obtain fast running time. In the second step, in order to enhance learning efficiency of GA, PSO is applied. A PSO moves through the problem space, with the moving velocity of each particle represented by a velocity vector and using the results of crossover and mutation as input.

At each time processing steps in this operation. Each particle also keeps track of its own best position, which is associated with the best fitness it has achieved so far in a vector. All graph of hybrid system by EU-GA-PSO show more satisfactory results than existing GA or PSO.

References

1. Matsummura,S.: Adaptive Control for the Steam Temperature of Thermal Power Plants,Proceedings the 1993 IEEE on Control applications (Sept. 1998) 1105-1109
2. Kennedy,J. and Eberhart,R.: Particle Swarm Optimization in Proc. IEEE int. Conf. Neural Networks, vol. IV, Perth, Australia (1995) 1942-1948
3. Angeline, P.J.: Using Selection to Improve Particle Swarm Optimization, IEEE int. Conf. 4-9 (May, 1998) 84-89
4. Chia-Feng Juang: A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design, Systems, Man and Cybernetics, Part B, IEEE Trans. Vol. 34 (2, April, 2004) 997-1006
5. Dong Hwa Kim:Comparison of PID Controller Tuning of Power Plant Using Immune and genetic algorithm. Measurements and Applications, Ligano, Switzerland (July 2003) 29-31
6. Dong Hwa Kim:Robust PID Controller Tuning Using Multiobjective Optimization based on Clonal Selection of Immune Algorithm,Proc. Int. Conf. Knowledge-based intelligent information and engineering systems. Springer-Verlag (2004)50-56
7. Yoshida,H., Kawata,K. and Fukuyama,Y.: A Particle Swarm Optimization for Reactive Power and Voltage Control Considering Voltage Security Assessment, IEEE Trans. Power Syst., Vol. 15 (Nov. 2000)1232-1239

Design and Implementation of Survivable Network Systems

Chao Wang[1], Jianfeng Ma[1,2], and Jianming Zhu[3]

[1] Key Laboratory of Computer Networks and Information Security,
(Ministry of Education), Xidian University, Xi'an 710071, China
xdkevin@gmail.cn
[2] School of Computing and Automatization,
TianJin Polytechnic University, Tianjin 300160, China
ejfma@hotmail.com
[3] School of Information, Central University of Finance and Economics,
Beijing 100081, China
tyzjm65@163.com

Abstract. As their scales and complexities increase, the computer-based network systems suffer from increasing probability of being intruded or crashed and decreasing dependability. Such a problem can be solved by extending the traditional security research to survivability research. This paper concentrates on the design and the implementation of the architecture and the applications of network systems to achieve its survivability. For the architecture, a double-**barrier** secure structure is constructed, i.e. the outer barrier defends and detects intrusions and the inner one tolerates intrusions and faults. The CORBA-based applications tolerate and detect intrusions and faults and recover from the adverse environment. Then the applications of the network system will be performed successfully despite the intrusions and faults, that is, the survivability of the network system will be achieved.

1 Introduction

The more complex topologies and applications the network systems have, the more vulnerabilities can be uncovered in the network systems and the security mechanisms adopted by them, which leads to more and more faults and attacks to the network systems. Extending the traditional security research to survivability research can reduce, even eliminate, the effects of the malicious attacks, system component faults and accidents on the network systems.

1.1 Survivability of Network Systems

At the earliest Neumann et al. defined the survivability of network systems [1]. Nowadays the most popular definition is made by Ellison et al. as following: Survivability is the capability of a system to fulfill its mission, in a timely manner, in the presence of attacks, failures, or accidents [2]. To maintain their capabilities to

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 376 – 385, 2005.

deliver essential services, survivable systems must exhibit the four key properties: resistance to attacks, recognition of attacks and the extent of damage, recovery of essential services after being attacked, and adaptation and evolution to reduce effectiveness of future attacks. In our view, the survivability researches should put emphasis on the protection of the capability to continue essential services. So the former three properties, i.e. resistance, recognition, and recovery, are fundamental ones and the last one, i.e. adaptation and evolution is a kind of survivability requirement of higher level.

1.2 CORBA Middleware

Presently, the heterogeneity of network systems is so significant that the qualities of services (QoSs) are impaired evidently. The heterogeneities emerge from several aspects, such as hardware and operating system, programming language, network protocols and so on. With the help of middleware, the system developers can neglect the heterogeneities and focus on designing the applications, which reduces the potential conflicts embedded in the network systems and ease the design and maintenance.

Middleware is a kind of software that lies between the application layer and network layer, which enables the applications independent of the underlying environment composed of heterogeneous operating systems, hardware platforms, network communication protocols, etc. Microsoft's DNA 2000 [3], SUN's J2EE [4], and OMG's CORBA [5, 6, 7] are the main middleware platforms for distributed heterogeneous applications. CORBA has the advantages of much better integratability, availability, and scalability over the others and attains the most extensive usage.

Though CORBA middleware cannot meet directly the survivability requirement of distributed systems, CORBA provides an efficient platform for systems to obtain high survivability and scalability.

1.3 Content and Organization

We focus on the design and the implementation of the architecture and the applications of network systems to achieve their survivability. The state-of-the-art security mechanisms are made full use of in the architecture and organized into a two-barrier security structure, which defends, recognizes, and tolerates malicious attacks and faults. CORBA-based applications themselves can detect, tolerate, and recover from attacks and faults. The services of the network system will survive and no longer rely on the dependability of peculiar system components. Meanwhile, the system can be extended easily, that is, the architecture has good scalability.

The rest of this paper is organized as follows. Section 2 introduces the architecture and the architecture-level survivability of network systems. Section 3 describes the application-level survivability. Section 4 discusses the related work. Finally, Section 5 presents some concluding remarks.

2 Architecture and Architecture-Level Survivability

2.1 Architecture

Fig. 1 shows the architecture of survivable network system. Proxy servers and COTS servers are the core components of the architecture and assumed to be vulnerable to attacks and faults. Proxy servers provide the public access points of the network system for clients, i.e., accept and make response to their requests. COTS servers process client requests and are shielded by the proxy servers.

The survivability of network systems is achieved hierarchically, that is, on the system architecture-level and the system application-level. Traditional security mechanisms such as firewall, authentication, access control, and IDS etc. compose the architecture-level protection and the CORBA-based applications compose the application-level protection.

2.2 Architecture-Level Survivability

In this paper, traditional security mechanisms are organized into a double-barrier security structure, i.e. the outer barrier defends and detects intrusions and the inner one tolerates intrusions and faults.

2.2.1 Outer Barrier

In Fig. 1, static security mechanisms such as firewall, authentication, and access control defend and IDS detects malicious attacks, which compose the outer security barrier.

The multiagent-based IDS [8] is a most important component of the architecture. The IDS is distributed across the whole system and its agents are located in every host to monitor the interesting events. Network traffic is analyzed and suspected actions are reported to the administrator.

The traditional security mechanisms, although defective, are the most important means to defend attacks, but they are powerless to system component faults.

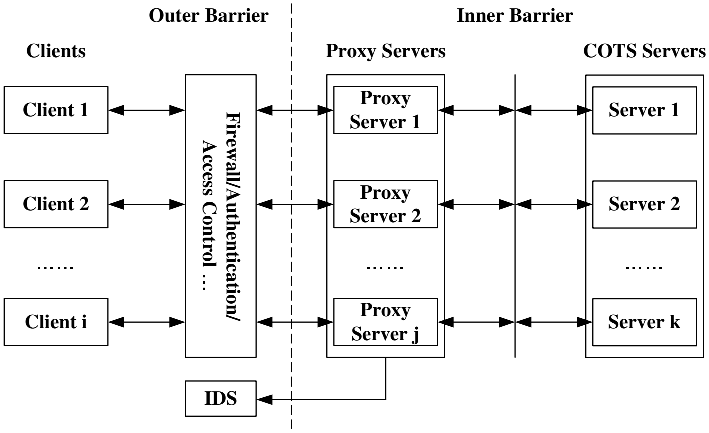


Fig. 1. Architecture of survivable network system

2.2.2 Inner Barrier

Proxy servers and COTS servers are diverse and redundant in software/hardware. Diversity eliminates the common-mode faults and identical vulnerabilities of the hosts. Redundancy shields the system component faults.

Proxy servers accept client requests and forward them to COTS servers. Proxy servers enforce the service policy specified by the operative intrusion strategy. The policy tells which COTS servers the request should be forwarded to, as well as how results from the COTS servers should be adjudicated to arrive at a final response. The policy also defines the criteria for external attack triggers that the proxy server generates and passes to the system administrator.

Each proxy server will have a unique physical IP address that is tied to distinct physical network interface; all the proxy servers share a pool of virtual IP addresses amongst themselves. Only the virtual IP addresses are visible to the clients, which allows easy migration of address from one machine to another in case of a fault or an intrusion. When under attacks, the service could be migrated to the other proxies to improve survivability.

Once IDS detects the existence of an attack or a fault, the administrator takes measures to reconfigurate the system such as the adjusting of the degree of redundancy and the level of access control and additional audit.

Although there exists a two-barrier security structure, the continuous applications of the network system cannot be guaranteed because: 1) the security mechanisms themselves are defective and trail attack techniques in time; 2) applications depend on peculiar system components. To survive, the applications should have the ability to defend themselves.

3 Application-Level Survivability

Diversity and redundancy brings up the problem of network heterogeneity (hardware, OS, network type, programming language, etc.), which gets worse the quality of services (QoSs) of system applications. The heterogeneities can be neglected with the help of middleware. Moreover, the applications based on CORBA middleware can protect themselves.

3.1 Application Model of Survivable Network System

Besides successful attacks, system component faults may occur, for example, processor faults, communication faults, etc. As to CORBA-based application systems, object faults also need consideration. Faulty processors and faulty object replicas may behave in an arbitrary or malicious manner, that is, sending error messages or not any message. Communication between processors may be unreliable and, thus, messages need to be retransmitted.

To eliminate the faults as mentioned above, the concept of object group is adopted. An object is replicated several times and its replicas are distributed into the whole system to form an object group. Logically, the object group equals the replicated object. All the correct group members are identical, that is, all of them are in the same state, put into the same operations, and return the same values. A voting is carried out based on the outputs of all the group members and its result is submitted to the application. Thus, a few faults can be shielded by the operations of the correct replicas.

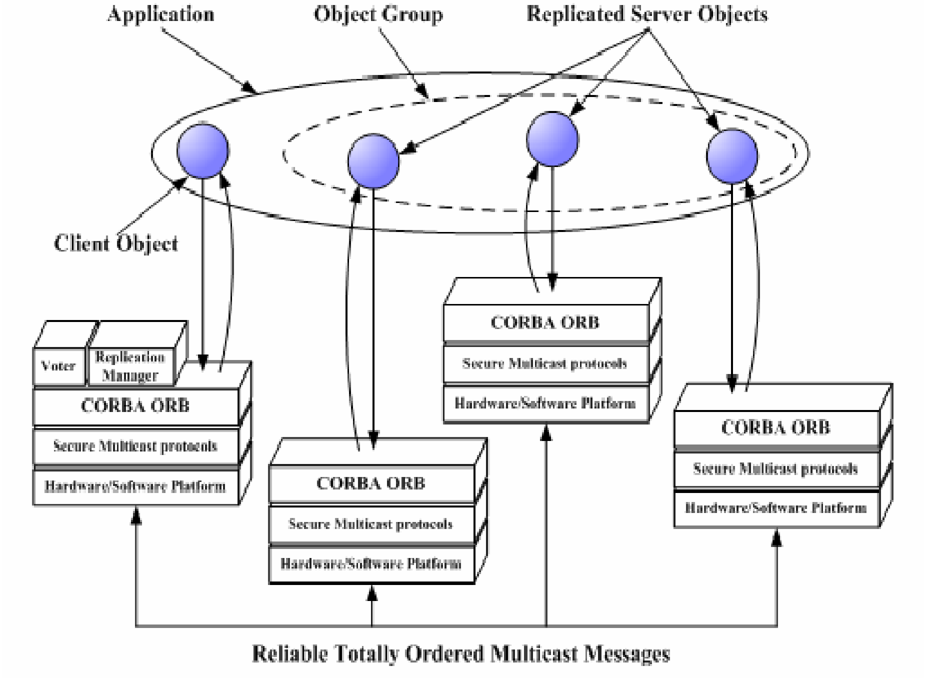

Application
Object Group
Replicated Server Objects
Client Object
Voter
Replication Manager
CORBA ORB
Secure Multicast protocols
Hardware/Software Platform
CORBA ORB
Secure Multicast protocols
Hardware/Software Platform
CORBA ORB
Secure Multicast protocols
Hardware/Software Platform
CORBA ORB
Secure Multicast protocols
Hardware/Software Platform
Reliable Totally Ordered Multicast Messages

Fig. 2. CORBA-based application model

In Fig. 2, Replication Manager is responsible for the replication of an object, the distribution of the replicas and the construction of the object group. To avoid the unnecessary loss of replicas due to the removal of a malicious processor, at most one replica of an object is allocated to a processor.

Although every object has its own object reference, the object group contacts with other objects through group reference. The object group interface allows other objects to transparently send an invocation or response to all of the group members by hiding the details of the underlying group communication protocols.

To provide dependable applications, some assumptions should be made for our research:

A1. Network partition doesn't occur;

A2. The difference in the fashions that ORB support fault-tolerate mechanism [9] is neglected;

A3. Reliable totally ordered multicast system (such as SecureRing [10]) exists;

A4. Not more than 1/3 of object group members break down at the sane time.

3.2 Creations of an Object Group and Its Members

The Replication Manager is the core component of the application model. Three interface components, i.e. Property Manager, Generic Factory, and Object Group Manager, perform the functions of the Replication Manager. The Property Manager specifies a set of fault-tolerate properties for the object replicas including replication style (Here, active replication is adopted because it can shield and recover from component faults transparently and rapidly), membership style, consistency style, factories, initial number of replicas, minimum number of replicas, etc. The Generic Factory interface is inherited by the Replication Manager to allow the application to invoke the Replication Manager directly to create and delete replicated objects in the same way; The Object Group Manager component allows applications to control directly the creation, deletion and location of individual replicas of an application object.

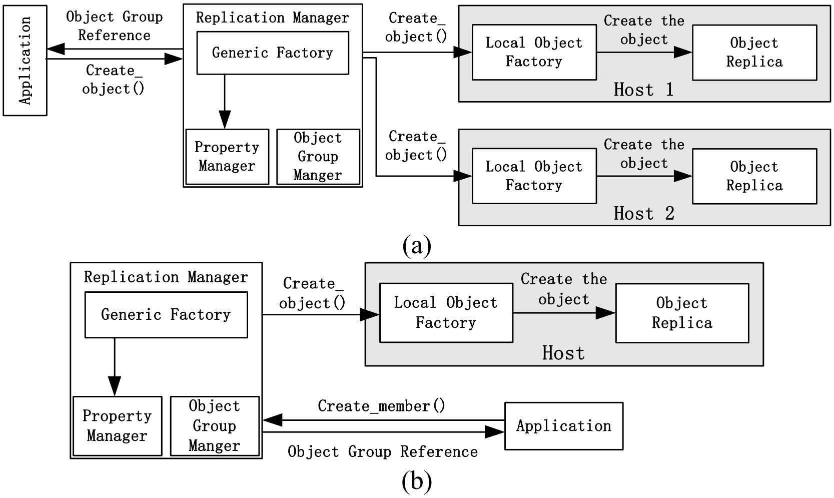


Fig. 3. (a) Creation of object group; (b) Creation of object group member

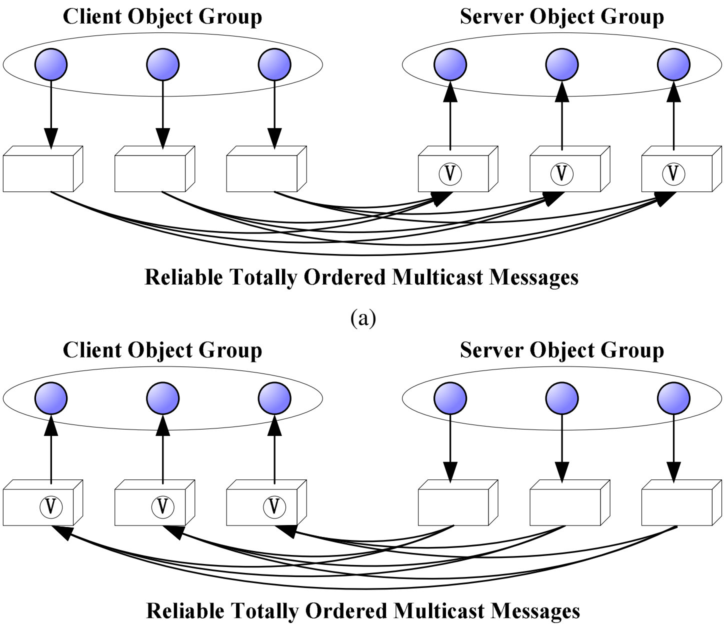
Fig. 3 (a) shows the creation process of an object group, which is described as following:

1. The application invokes the *resolve_initial_references()* operation to obtain a reference to the Replication Manager;
2. The application invokes the *create_object()* operation of the Generic Factory interface to obtain the *factory_creation_id* and records the return value convenient for the subsequent deletion of the object group;
3. The Replication Manager learns fault-tolerate properties from the Property Manager and location property from the Local Object Factory;
4. The Replication Manager invokes the *create_object()* operation of the Generic Factory interface to create the members on the corresponding locations and the operation returns the reference of the member and its *factory_creation_id* which

is recorded by the Local Object Factory and the Replication Manager convenient for the subsequent deletion of the member;

5. The Replication Manager determines the identifier of the object group, constructs the object group reference, activates the group members and returns the related information to the application.

The creation process of a group member is similar to that of the object group except for that the application invokes directly the *create_member()* operation of the Object Group Manger to create a member (See Fig. 3 (b)).


Client Object Group
Server Object Group
V
Reliable Totally Ordered Multicast Messages
Client Object Group
Server Object Group
V
Reliable Totally Ordered Multicast Messages

(a)

(b)

Fig. 4. (a) Majority voting on invocation; (b) Majority voting on response

3.3 Fault Detection and Toleration

All the components of the system may crash. So client objects need to be replicated to construct object groups as the server objects do. Voting is adopted to select correct messages from the object groups (See Fig. 4 (a) and (b)).

Majority voting is the most popular form to select the correct messages. To achieve majority voting, the voter needs several copies of an identical invocation (response) as the input and submits an output to the application.

The voter obtains the number of the group members, i.e., the group size (denoted by n) from the Property Manager. $\left\lceil {}^{n+1}\!/_{3} \right\rceil$ identical messages or $\left\lceil {}^{2n+1}\!/_{3} \right\rceil$ total messages are required to carry out the voting, which can guarantee the correctness of the applications according to the assumption A4. After the completion of the voting, the messages that arrive late will be discarded. The application will be rebooted if not any message can be selected out by the voter.

In an object group there exist several client object replicas. Their repeated invocations will corrupt the states of the server object replicas, which can be avoided by the use of a unique operation identifier (invocation identifier and response identifier). The invocation identifier uniquely identifies an invocation by the client object replicas. The Replication Manager of each server replica uses a response identifier when returns the results of the invocation. The response enables the Replication Manager of each client replica to associate the copies of the returned response with the invocation. In fact, the votings are based on the received messages with the same invocation (response) identifiers.

Obviously, votings can neglect a few component faults and find the faulty objects and processors that send malicious messages, that is, votings can tolerate and detect partial component faults.

3.4 Recovery

Once a fault is detected, the faulty component should be expelled from the system immediately, otherwise it will corrupt other components. To protect the system to the utmost extent, the host that runs the faulty object will be deleted as well as other objects located on it.

It is necessary to maintain a constant object group size to hold the application's fault-tolerate property. So an object replica will be created and added to the object group after a member being deleted. Section 3.2 has introduced the creation of a group member. It is fairly difficult to synchronize the state of the new replica with that of the other replicas of the object.

As to recovery, every object should be considered to have three kinds of state: application-level state, ORB/POA-level state, and infrastructure-level state. All the three kinds of states are required to be consistent when the state synchronization is sought. The process of state synchronization is described as following:

1. The Replication Manager invokes the *get_state()* operation of a correct replica;
2. The message returned by the correct replica is overlapped with its ORB/POA-level state and infrastructure-level state, which is used as the parameter of a *set_state()* invocation on the new replica;
3. The correct replica processes orderly the queued operations which arrived during processing the *get_state()* operation;
4. The new replica carries out the *set_state()* invocation that comes from the Replication Manager.
5. The new replica processes orderly the other queued messages that arrived during synchronizing its state.

Thus, the new replica and the correct replicas will be in the same state, i.e., a group member is recovered.

From the description as mentioned above, it can be seen that the system applications can detect, tolerate, and recover from partial component faults, that is, the system applications can protect the system and themselves. Moreover, the protection doesn't rely on certain system components any more.

3.5 Performance Analysis

For network systems, there exist some tradeoffs between their properties, including the tradeoff between survivability and other properties (performance, real-time, etc.). In this paper, some techniques are adopted to achieve dependable and continuous services, which will depress the performance of network systems. For example, all the objects of an application are replicated to shield unexpected faults and votings are required to elect results from the outputs of the replicas of the same objects, which will inevitably lead to additional overhead. But compared with non-replication systems, the systems presented are much more dependable and robust.

4 Related Work

In [11, 12], survivable architecture of network systems is presented, in which a control subsystem is introduced to manager the information system. In [13], a survivable model of database system is presented. In these works, the survivability of a system is heavily dependent on the monitor component (or IDS), which will be the bottleneck of network systems' survivability.

The construction of fault-tolerant CORBA-based system is introduced in [14] and state consistency and recovery are discussed in [15, 16]. But the fault-tolerant CORBA used by them does not support heterogeneous CORBA environments.

5 Conclusions

A survivable model of network system is presented in this paper. Based on the state-of-the-art security mechanisms, the system are defended from both the whole system architecture and its applications, which enables the applications independent of the peculiar system components, that is, the survivability of the network system is protected to the utmost extent.

However, similar to the secure research, it is an "Impossible Mission" to construct an absolutely survivable network system for the survivability research because of the insurmountable defects of every aspect of the system construction as well as their unavoidable permanent existences. So it is much advisable to meet the survivability requirement at a certain aspect, such as continuity, availability, reliability and so on.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant No.90204012), the National "863" High-tech Project of China (Grant No. 2002AA143021), the Excellent Young Teachers Program of MOE, China and the Key Project of Chinese Ministry of Education.

References

1. Neumann, P. G., Hollway, A., Barnes, A.: Survivable Computer-Communication Systems: The Problem and Working Group Recommendations. Technical Report VAL-CE-TR-92-22 (revision1), U.S. Army Research Laboratory, AMSRL-SL-E, White Sands Missile Range, NM 88002-5513, May 1993
2. Ellison, B., Fisher, D.A., Linger, R.C., et al.: Survivable Network Systems: An Emerging Discipline, Technical Report CMU/SEI-97-TR-013, Software Engineering Institute, Carnegie Mellon University, November 1997
3. http://www.microsoft.com/technet/archive/itsolutions/intranet/build/dna2k.mspx
4. http://java.sun.com/j2ee/overview.html
5. Object Management Group, Inc. The Common Object Request Broker (CORBA): Architecture and Specification, Y 2.0, November 1995
6. Vinoski, S.: CORBA: Integrating Diverse Applications within Distributed Heterogeneous Environments. IEEE Communications Magazine, 1997, vol.35, No.2, pp.46 –55
7. Vinoski, S.: New Features for CORBA 3.0, Communications of the ACM. vol. 41, pp. 44-52, October 1998
8. Spafford, E. H. and Zamboni, D.: Intrusion Detection Using Autonomous Agents. Computer Networks, vol.34, No.4, pp.547-570, October 2000
9. Felber, P., Narasimhan, P.: Experiences, Strategies, and Challenges in Building Fault-Tolerant CORBA Systems. IEEE Trans. Computers vol.53, no.5, pp.497-511, 2004
10. Kihlstrom, K.P., Moser, L.E., Melliar-Smith, P.M.: The SecureRing Group Communication System. ACM Transactions on Information and System Security, vol.4, no.4, pp. 371-406, November 2001
11. Knight, J. C., Sullivan, K. J., Elder, M. C. et al.: Survivable Architectures: Issues and Approaches. DARPA Information Survivability Conference and Exposition, January 2000
12. Knight, J. C., Heimbigner, D., Wolf, A. et al.: The Willow Architecture: Comprehensive Survivability for Large-Scale Distributed Applications. The International Conference on Dependable Systems and Networks, June 2002
13. Jianming Zhu, Jianfeng Ma: Intrusion-Tolerant Based Survivable Model of Database System. Chinese Journal of Electronics. (To be appeared)
14. Narasimhan, P.: Transparent Fault Tolerance for CORBA. PhD thesis, Department of Electrical and Computer Engineering, University of California, Santa Barbara, December 1999
15. Narasimhan, P., Kihlstrom, K. P., Moser, L. E., et al.: Providing Support for Survivable CORBA Applications with the Immune System. In Proceedings of the 19th IEEE International Conference on Distributed Computing Systems, pp. 507-516, May 1999
16. Narasimhan, P., Moser, L. E., Mellar-Smith, P. M.: Strongly Consistent Replication and Recovery of Fault-Tolerant CORBA Applications. Computer System Science and Engineering Journal, vol. 17, no. 2, pp. 103-114, March 2002

Optimal Placement of Active Members for Truss Structure Using Genetic Algorithm

Shaoze Yan[1], Kai Zheng[2], Qiang Zhao[1], and Lin Zhang[1]

[1] Department of Precision Instruments and Mechanology, Tsinghua University,
Beijing 100084, P.R. China
yansz@mail.tsinghua.edu.cn
[2] Mechanical Engineering School, Beijing University of Science and Technology,
Beijing 100083, P.R. China
zhengkai00@mails.tsinghua.edu.cn

Abstract. The objective of this work is to develop an optimization methodology to design adaptive truss structures with multiple optimally placed active members. The finite element model of truss structures with piezoelectric members is presented. The performance function is built for optimal design of active members at discrete locations in the output feedback control system by using the method proposed by K. Xu et al. Genetic algorithm (GA) is used to search the optimal locations of active members for vibration suppression of adaptive truss structure. A numerical example of the planar truss structure with two piezoelectric active members is performed, and the corresponding experimental set-up is designed for active vibration control. The experimental results demonstrate the effectiveness of optimal placement of active members for adaptive truss structures using genetic algorithm.

1 Introduction

The adaptive truss structure is a new kind of structures integrated with active members, which can automatically respond to internal or external disturbances. The technology on piezoelectric adaptive truss structure is expected to solve the problems which are caused by space nonlinear disturbances such as vibration, precision location and shape retention. Stringent vibration control methods are required for many space structures such as antenna reflectors and space stations. An effective approach to improve the performance of truss structures is to damp out vibration in orbit by using variable-length active members. The number of actuators is usually limited because of cost and weight, and so it is important to locate the actuators optimally. The optimal placement of actuators at discrete locations is an important problem that will have impact on the control of truss structures. Though algorithms exist for the placement of sensor/actuator systems on continuous structures, the placement of actuators on discrete truss structures is a very difficult problem. Because of the nature of truss structures, it is not possible to place sensors and actuators at any location in the structure. This usually creates a non-linear constrained mixed integer planning problem that can be very difficult to solve.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 386 – 395, 2005.

Genetic algorithm (GA) is becoming increasingly popular due to their ability to solve large complex optimization problems which other methods have difficulty solving [1]. The introduction of algorithms based on genetic search procedures should increase the rate of convergence and thus reduce the computational time for solving the difficult control problem. Genetic algorithm in conjunction with gradient-based optimization techniques was used for the simultaneous placement and design of an active civil structure [2]. A genetic algorithm procedure was adopted to solve the optimization problem for actuator and sensor placement of a truss structure with closely-spaced modal frequencies [3]. The improved simulated annealing approach and genetic algorithm were applied to improve the shape accuracy of truss structures [4, 5]. The optimization procedure of genetic algorithms was discussed in order to solve the optimization of active member placement for intelligent truss structures [6, 7, 8]. In this paper, the objective of this work is to develop an optimization methodology to design adaptive truss structures with multiple optimally placed active members. A dynamic finite element model and state equations of piezoelectric intelligent truss structure are presented in Section 2. Optimal placement method using genetic algorithm is described in Section 3. Numerical example and experimental validation are presented in Section 4. The final section is the conclusions.

2 Formulation of the Problem

2.1 Modeling of Truss Structures Integrated with Active Members

The finite element formulation of an active member is similar to that of a passive structural member except for the initial strain or elongation caused by the built-in actuator, e.g., stack of piezoelectric wafers [9]. An active member itself typically consists of some passive mechanical components in addition to the actuator element, as shown in Figure 1. The total strain in an active member consists of strain ε_m caused by mechanical loads and initial strain ε_a induced by the actuator, i.e., $\varepsilon = \varepsilon_m + \varepsilon_a$ [9, 10]. As in the linear finite element formulation, the axial displacement in the element is approximated by the nodal displacements and linear shape functions.

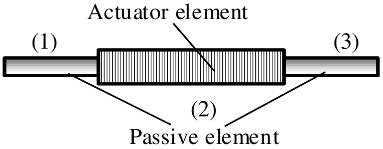


Fig. 1. Finite element representation of the active member

The total potential energy, including strain energy density and the kinetic energy expression of the active element, can be founded in some references [9]. Transforming element motion equation from the element coordinate to the global coordinate system, the active element equation is given by

$$[M^e]\{\ddot{x}^e\}+[C^e]\{\dot{x}^e\}+[K^e]\{x^e\}=[T]^{\mathrm{T}}\{F^e\}+[T]^{\mathrm{T}}\{F_v\}, \tag{1}$$

where $\{x^e\}$ is the nodal displacement vector of the active element, $\{x^e\}=\{u_i \quad v_i \quad u_j \quad v_j\}^T$, u_i and v_i are the horizontal and vertical displacements at the *i*th node, respectively. $[C^e]$ is the damping coefficient matrix of the active member. $[M^e]$ is the equivalent mass matrix,

$$[M^e]=\mathrm{diag}(m_i, m_i, m_j, m_j), \tag{2}$$

$$\begin{gathered} m_i=[3m_1^e(l_1+2l_2+2l_3)+3m_2^e(l_2+2l_3)+3m_3^e l_3]/(l_1+l_2+l_3), \\ m_j=[3m_1^e l_1+3m_2^e(2l_1+l_2)+3m_3^e(2l_1+2l_2+l_3)]/(l_1+l_2+l_3), \\ m_i^e=\rho_i A_i l_i/6,\ i=1,2,3; \end{gathered} \tag{3}$$

where ρ_i, A_i and l_i are the density, cross-section area, and length of the *i*th element of the active member, respectively. $[T]$ is the transformation matrix which consists of the active member's direction cosines,

$$[T]=\begin{bmatrix} \cos\alpha & \sin\alpha & 0 & 0 \\ -\sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & \cos\alpha & \sin\alpha \\ 0 & 0 & -\sin\alpha & \cos\alpha \end{bmatrix}. \tag{4}$$

where α is the angle between the global coordinate *X* and the local coordinate *x*. $[K^e]$ is the stiffness matrix of the active member,

$$[K^e]=k_e[T]^T\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}[T], \tag{5}$$

$$k_e=k_1^e k_2^e k_3^e/(k_1^e k_2^e+k_1^e k_3^e+k_2^e k_3^e),\ k_i^e=E_i A_i/l_i,\ i=1,2,3\,. \tag{6}$$

where E_i and k_i^e are the elastic modules and axial stiffness of the *i*th element of the active member. $\{F^e\}$ is the nodal force vector of the member. $\{F_v\}$ represents the actuating force vector caused by the active member,

$$\{F_v\}=-\{K_{uv}^e\}V_d,\ \{K_{uv}^e\}=K_{uve}\{-1\ \ 0\ \ 1\ \ 0\}^T, \tag{7}$$

$$K_{uve}=k_1^e k_3^e k_{uv}^e/(k_1^e k_2^e+k_1^e k_3^e+k_2^e k_3^e),\ k_{uv}=e_{33}A_2/l_2, \tag{8}$$

where e_{33} is the stress coefficient of piezoelectric materials.

The system equations of motion for entire truss referring to the global coordinate system can be obtained from Eq. (1) by the standard assemblage procedure [10]. Assembling all the elements accordingly yields the governing equation of motion for a n-degree of freedom adaptive truss structure,

$$[M]\{\ddot{x}\}+[C]\{\dot{x}\}+[K]\{x\}=\{F_e\}+[B_0]\{F_c\}, \tag{9}$$

where $\{x\}$ is displacement vector in the global coordinate system. $[M]$, $[C]$ and $[K]$ are the $n\times n$ mass, damping and stiffness matrices, respectively. $[B_0]$ is the $n\times m$ assembling matrix of active members and m is the number of active members. $\{F_e\}$ is the $n\times 1$ external nodal force vector. $\{F_c\}$ is the $m\times 1$ equivalent actuating force vector of active members.

By introducing the coordinate transformation, $\{x\}$ can be described as

$$\{x\}=[\varPhi]\{\eta\}, \tag{10}$$

where $[\varPhi]$ is the $n\times n$ modal matrix whose columns are the eigenvectors and $\{\eta\}$ is the vector of modal coordinates.

Using Eq. (10), Eq. (9) can be transformed into n uncoupled equations,

$$\{\ddot{\eta}\}+[D]\{\dot{\eta}\}+[\bar{K}]\{\eta\}=[\Phi]^T\{F_e\}+[\Phi]^T[B_0]\{F_c\}, \tag{11}$$

where $[D]$ and $[\bar{K}]$ are diagonal matrices, $[D]=[\Phi]^T[C][\Phi]=diag\left[2\xi_j\omega_j\right]$, $[\bar{K}]=[\Phi]^T[K][\Phi]=diag\left[\omega_j^2\right]$. The modal matrix $[\Phi]$ is normalized with respect to the mass matrix. ω_j and ξ_j are structural frequency and modal damping factor.

For active members, a constant output velocity feedback control law will be considered, i.e. [11, 12],

$$\{F_c\}=-[G]\{y\}, \tag{12}$$

where $[G]$ is the control gain matrix. $\{y\}$ is the output vector, $\{y\}=[H_0]\{\dot{\eta}\}$, and $[H_0]$ is the placement matrix of the sensors.

2.2 State Equations

In the following, it will be assumed that there is no external forces, $\{F_e\}=0$, without affecting generality of the results [5]. The control analysis is performed by reducing the second-order uncoupled Eq. (11) to a first-order equation. This can be achieved by using the transformation,

$$\{z(t)\}=\{\eta \quad \dot{\eta}\}^{\mathrm{T}}, \tag{13}$$

where $\{z(t)\}$ is the $2n\times 1$ state variable vector. It is convenient to express Eq. (11) as a first-order form, so the state equation and output vector can be written as

$$\{\dot{z}(t)\} = [A_0]\{z(t)\} + [B]\{u(t)\}\,,\ \{y(t)\} = [\mathrm{H}]\{z(t)\}\,, \tag{14}$$

where $[A_0]$ is a $2n\times 2n$ plant matrix and $[B]$ is a $2n\times m$ input matrix, which is related to placements of active members of the truss. $u(t)$ is the input vector. The plant and the output matrices are given by

$$[A_0] = \begin{bmatrix} [O] & [I] \\ [-\omega_i^2] & [-2\xi_i\omega_i] \end{bmatrix},\ [B] = \begin{bmatrix} [O] \\ [\Phi]^T[B_0] \end{bmatrix},\ [H] = \begin{bmatrix} [O] \\ [H_0] \end{bmatrix}. \tag{15}$$

The free vibration response of the closed-loop system for any initial condition $\{z(0)\}$ is [11, 12, 13]

$$\{z(t)\} = e^{At}\{z(0)\}\,, \tag{16}$$

where e^{At} is the fundamental transition matrix and

$$[A] = [A_0] - [B][G][H]\,. \tag{17}$$

2.3 Optimal Placement Problem

A performance function can be chosen that includes both the structural response and the control effort. The standard performance function is

$$\hat{J} = \frac{1}{2}\int_0^\infty \left(z^TQz + u^TRu\right)dt\,, \tag{18}$$

which can be expanded in terms of the fundamental transition matrix [11, 12],

$$\hat{J} = \{z(0)\}^T\left[\frac{1}{2}\int_0^\infty e^{A^Tt}\left([Q]+[H]^T[G]^T[R][G][H]\right)e^{At}dt\right]\{z(0)\}\,. \tag{19}$$

The dependence on $\{z(0)\}$ is eliminated by using an average performance function proposed by Levine et al [13]. The performance function can be described as [11, 12],

$$J_c = \frac{1}{2}\mathrm{trace}[P]\,, \tag{20}$$

where the matrix $[P]$ is related to the performance function,

$$[P] = \int_0^\infty e^{A^Tt}\left([Q]+[H]^T[G]^T[R][G][H]\right)e^{At}dt\,, \tag{21}$$

The matrix $P(t)$ is obtained by using the associated Lyapunov equation [11, 12],

$$[P][A] + [A]^T[P(t)] + [H]^T[G]^T[R][G][H] + [Q] = 0\,. \tag{22}$$

The (0,1) formulation of the optimal placement problem makes use of a variable g_i that indicates the presence (1) or absence (0) of an active element at a given location [14]. The objective function of optimal placement of active members is based on Eq. (20) as follows,

$$\underset{\{g_i,\, g_2,\ldots,g_p\}}{Minimize}\; J_c\,, \tag{23}$$

subject to the constraint

$$\sum_{i=1}^{p} g_i = m, \quad g_i \in (0,1), \;\; i = 1,2,\ldots,p\,, \tag{24}$$

where m is the number of the active members, and p is the total number of members in the truss structure.

3 Optimal Placement Method Using Genetic Algorithm

The number of possible configurations of a fixed number of active members may become very large for a given control system. Therefore, exhaustive search is ruled out and intelligent search techniques have to be employed. GA is a guided random search technique simulating natural evolution. It deals with a population, and the chromosome of each individual represents a candidate solution, i.e., an actuator placement configuration in the present problem. By dealing with this group of solutions, this approach tries to avoid being trapped at local optima.

3.1 Problem Parameter Encoding

The GA works with an encoding of the parameter set, not the parameters themselves. In the standard GA approach, the genes in the chromosome are binary. Assuming the active members can be placed at the p possible positions in the truss structure, the number p is selected as the coding length. Assuming $g_i (i = 1,2,\ldots,p)$ is the gene value of the *i*th bit of the string, $g_i = 1$ or 0 denotes the active member or passive member placed at that position. For example, let $p = 8$, $m = 2$, the coding (10100000) shows that two active members are placed at the positions 1 and 3.

3.2 Calculation of Fitness Value

After encoding, an initial chromosome population is randomly generated. Each individual in the initial population should satisfy the constraints given by the problem. Each individual from the population is evaluated and assigned a fitness value. The best performing individuals, based on their fitness value, are chosen from the population to produce the next generation of population through a genetic reproduction process. Define J_c in Eq. (20) as the objective function, so the fitness value of the *i*th individual can be described as

$$f_i(x) = 1/J_c,\ i = 1,2,...,m\,. \tag{25}$$

3.3 Genetic Operators

A simple genetic algorithm uses three operators, namely selection, crossover, and mutation [15].

Selection. This selection process is random by nature; however, the individuals with best fitness values would have the more chance to be chosen for reproduction. The fitness of the individuals are calculated, which are newly generated. The fitness of a solution is measured by its capacity to maximize or minimize the objective function. Probability of the i th individual is chosen as

$$P_s = f_i(x) / \sum_{i=1}^{m} f_i(x)\,. \tag{26}$$

Crossover. After selection operations, two parent individuals exchange information (genetic information) to produce a new offspring. This process is called crossover. Crossover is governed by a crossover probability, which tells the GA program the number of bit to crossover. The GA chooses the bits that are crossed-over randomly. In this operation, two individuals are randomly selected as parents from the pool of individuals formed by the selection procedure and cut at a randomly selected point. A crossover operation can thus yield better solutions by combining the good features of existing solutions. If no crossover is performed due to the crossover probability, the offspring is the exact copy of the parent solutions.

Mutation. Each offspring created from crossover is altered in a process called mutation. Mutation alters each offspring at random location in the solution string. Mutation is governed by a mutation probability that determines the number of bits to be mutated. In binary, mutation will change a bit equal to 0 into 1 and a bit equal to 1 into 0. Mutated offspring are then placed into the next generation of the population. The mutation operator randomly mutates or reverses the values of bits in a string. The effect of this operation is to create diversity in the solution population, thus the operation can avoid local minimum traps.

After the crossover and mutation operation, the offspring generated could violate the constraints of the problem, even if all individuals of the parents satisfy the constraints. For example, two individuals 10100000 and 01000010 for placements of two active members could become 10100010 and 01000000 after the crossover operation at the 5th bit of the strings. Of course, the mutation operation could also produce the individuals violating the constraints of the system. A constrained mutation operation is proposed to avoid constraint violations in the genetic algorithm [3]. Assuming there is m active members in the truss, and the number of $g_i = 1$ $(i = 1,2,...,p)$ for the certain individual is n . When $n > m$, $n - m$ bits of $g_i = 1$ are randomly chosen to change into 0. When $n < m$, $m - n$ bits of $g_i = 0$ are randomly selected to change into 1.

The process of selection, crossover and mutation are repeated until the new population is completed. The new population will replace the old population and the entire process will continue until an end condition is met, and the best solution is found.

4 Numerical Example and Experimental Validation

A planar truss structure is shown in Fig. 2, which is analyzed to investigate the effectiveness of the proposed scheme. The corner nodes 1 and 6 in the truss structure are fixed on the base, and the truss has 8 bars, in which there are two active members to be placed. The passive members and elements are made of aluminum alloy tube with diameter 20mm and thickness 1.5mm. The Young's module of aluminum alloy is $7.0\times10^{10}\ \mathrm{N/m^2}$, and density $2.74\times10^{3}\ \mathrm{kg/m^3}$. The lump masses of end nodes of passive and active members are 0.29kg and 0.5kg, respectively. The lengths of members are 500mm except for diagonal bars 3 and 5. The mass density and Young's module of piezoelectric stacks are $7.5\times10^{3}\ \mathrm{kg/m^3}$ and $6.3\times10^{10}\ \mathrm{N/m^2}$, respectively. The equivalent stress coefficient e_{33} is $18.62\,\mathrm{C/m^2}$. The piezoelectric stacks are made of circular piezoelectric patches with the diameter of 10mm and their corresponding length is 120 mm.

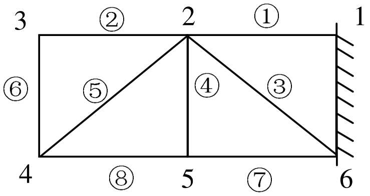


Fig. 2. A planar truss

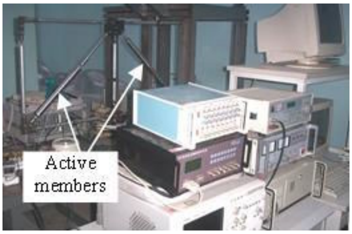


Fig. 3. Experimental setup

The control systems used for the present study are limited to collocated sensors and actuators with direct velocity feedback gains. The control gain matrix $[G]$ is selected as a constant gain matrix, $[G]=200[I]$. Because the system is collocated, let $[H]=[B]^T$. For this problem, the GA generates a string composed of bits that store information about the final design of the truss structure. This string is randomly generated in Matlab and put into a population composed of 20 strings generated by the GA. Each bit in a string represents an element in the truss structure. Eight bit numbers are used to describe all possible element scenarios. Weighting matrices $[Q]$ and $[R]$ are taken to be $5000[I]$ and $[I]$, respectively. Probability of crossover is 0.8, and probability of mutation 0.01. After the convergence is achieved, the optimal individual is obtained to be 00101000, which means that placement (3 and 5) is the optimal placement for this problem.

Table 1. The active damping ratios of the truss structure

Mode Number	Optimal placement (positions: 3 and 5)		Non-optimal Placement (positions: 1 and 6)	
	f(Hz)	ξ (%)	f(Hz)	ξ (%)
1	61.2	11.267	60.5	8.232
2	142.4	4.983	143.3	3.251
3	248.7	1.121	247.4	1.021
4	334.3	1.538	331.3	1.543
5	437.2	2.134	431.8	1.715

Figure 3 shows the photograph of the experimental setup for vibration control. The intelligent controller is applied to the truss vibration suppression, and control command signals are calculated to drive the active members. In order to compare the control effects between the optimal placement and the others, the modal tests are performed for the optimal placement (two active members are placed on placements 3 and 5) and the non-optimal placement (the two active members are placed on placements 1 and 6), respectively. The experimental results are shown in Table 1. As expected, the modal damping ratios increase significantly by using the closed-loop control of the optimal placement, and the first modal damping ratio increases about 30% especially. It can be seen that the controller of the optimal placement provides excellent damping.

5 Conclusions

In the design of adaptive truss structures, the determination of the actuator locations is a very important issue. This paper describes the optimal location selection of actuators for vibration control in the adaptive truss structures. A procedure is given for the placement of active members in the adaptive structures for output feedback type control. GA is applied to the active member placement optimization problem for vibration suppression of truss structures. The proposed optimal placement approach is applied to multiple input control of a planar truss structure with active members. An experimental set-up of piezoelectric adaptive truss structure is designed for vibration control, and the experimental results show that the optimal placement of active members is feasible and effective.

Acknowledgement

This work is supported by the National Science Foundation of China under grant 59905016 and 50275080.

References

1. Bland, S.M., Sheng, L.Z., Kapania, R.K.: Design of Complex Adaptive Structures Using the Genetic Algorithm. Proc. SPIE Int. Soc. Opt. Eng. 4512 (2001) 212-218
2. Abdullah, M.M., Richardson, A., Hanif, J.: Placement of Sensors/Actuators on Civil Structures Using Genetic Algorithms. Earthquake Engineering and Structural Dynamics, 8 (2001) 1167-1184

3. Liu, F.Q., Zhang, L.M., Link, M.: Optimal Placement of Actuators and Sensors for Vibration Active Control. Proc. Int. Modal Analysis Conf., 2 (1999) 1820-1825
4. Onoda, J., Hanawa, Y.: Actuator Placement Optimization by Genetic and Improved Simulated Annealing Algorithms. AIAA Journal, 6 (1993) 1167-1169
5. Chen, G.S., Bruno, R. J., Salama, M.: Optimal Placement of Active/Passive Members in Truss Structures Using Simulated Annealing. AIAA J., 8 (1991) 1327-1334
6. Guo, H.Y., Zhang, L., Zhou, J.X., Jiang, J.: Optimal Placement of Actuators in Actively Controlled Structures by Genetic Algorithms. Process in Safety Science and Technology, 3 (2002) 62-67
7. Xu, B, Jiang, J.S.: Integrated Optimization of Structure and Control for Piezoelectric Intelligent Trusses with Uncertain Placement of Actuators and Sensors. Computational Mechanics, 5 (2004): 406-412
8. Gao, W., Chen, J.J., Ma, H.B., et al.: Optimal Placement of Active Bars in Active Vibration Control for Piezoelectric Intelligent Truss Structures with Random Parameters. Computers and Structures, 1 (2003) 53-60
9. Chen, G.S., Lurie, B.J., Wada, B.K.: Experimental Studies of Adaptive Structures for Precision Performance. AIAA Paper No. 89-1327-CP, 1989
10. Sun, C.T., Wang, R.T.: Enhancement of Frequency and Damping in Large Space Structures with Extendible Members. AIAA J., 12 (1991) 2269-2271
11. Xu, K., Warnitchai, P., Igusa, T.: Optimal Locations and Gains of Sensors and Actuators for Feedback Control. AIAA paper 93-1660 (1993) 3137-3145
12. Abdullah, M.M.: Optimal Location and Gains of Feedback Controllers at Discrete Locations. AIAA J., 11 (1998) 2109-2116
13. Levine, W.S., Athans M.: On the Determination of the Optimal Constant Output Feedback Gains for Linear Multivariable Systems. IEEE Trans. Automat. Contr., 1 (1970) 44-48
14. Sepulveda, A.E., Jin, I.M., Schmit, L.A.J.: Optimal Placement of Active Elements in Control Augmented Structural Synthesis. AIAA Journal, 10 (1993) 1906-1915
15. Tsahalis, D. T., Katsikas, S. K., Manolas, D. A.: A Genetic Algorithm for Optimal Positioning of Actuators in Active Noise Control: Results form the ASANCA Project. Aircraft Engineering and Aerospace Technology, 3 (2000) 252-257

Performance Comparison of SCTP and TCP over Linux Platform

Jong-Shik Ha, Sang-Tae Kim, and Seok J. Koh

Department of Computer Science,
Kyungpook National University, Korea
{mugal1, saintpaul1978, sjkoh}@cs.knu.ac.kr

Abstract. Stream Control Transmission Protocol (SCTP) is the third transport layer protocol next to TCP and UDP. The SCTP provides some distinctive features over the TCP. This paper is purposed to compare SCTP and TCP in the performance perspective. We compare the throughput of SCTP and TCP for the three different test scenarios: the performance comparison of SCTP and TCP for the different size of the user input data for the socket system call, the analysis of the fairness under competition of SCTP and TCP traffic, and the performance comparison of the SCTP multi-homing and single-homing cases. From the results, it is shown that the SCTP provides better throughput over TCP for a larger user input data. We also see that the SCTP traffic tends to compete fairly with TCP and that the multi-homing SCTP provides better performance than the single-homing case.

1 Introduction

Stream Control Transmission Protocol (SCTP) is a new transport protocol next to TCP and UDP, which was standardized in the IETF [1]. Similarly to the TCP, the SCTP is a connection-oriented reliable transport protocol. Differently from the TCP, the SCTP uses the four-way handshake procedure for association establishment and the three-way handshake scheme for association termination. In particular, the SCTP provides the ‘multi-streaming’ and ‘multi-homing’ features.

Some previous studies [2, 3] include the performance analysis of the SCTP itself. In this paper, we focus on the comparison of the performance of SCTP and TCP in the viewpoint of the throughput over the Linux platform [4, 5]. The SCTP performance is analyzed for the three kinds of test scenarios: 1) performance comparison of SCTP and TCP for the different size of the user input data in the socket system call, 2) analysis of the fairness under the competition of SCTP and TCP traffics, and 3) performance gain of the SCTP multi-homing.

This paper is organized as follows. Section 2 briefly summarizes the features of the SCTP. In Section 3, we describe the three test scenarios for the performance comparison over the Linux platform. Section 4 shows the experimental results for the performance testing. Section 5 concludes this paper.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 396 – 404, 2005.

2 SCTP Features

In this section we describe the distinctive features of SCTP, which include the SCTP association setup, association termination, multi-streaming and multi-homing features.

2.1 Four-Way Association Establishment

Differently from the 3-way handshake mechanism of TCP, the SCTP uses the 4-way handshake scheme for establishment of an SCTP association. Let us consider the two SCTP endpoints, A and B.

First, the endpoint A sends an SCTP INIT chunk to the endpoint B for initiation of an SCTP association. The endpoint B will respond with the INIT-ACK chunk to A, which contains the ‘cookie’ information for the security purpose. The endpoint A will then send the COOKIE-ECHO chunk to the B. The endpoint B completes the association establishment by sending the COOKIE-ACK chunk to the A.

It is noted in Figure 1 that this 4-way handshake scheme of SCTP is employed for preventing the so-called TCP SYN flooding. That is, the SCTP endpoint B will allocate the relevant kernel memory for the connection from the endpoint A, only after receiving the third COOKIE-ECHO chunk (after confirmation that the peer endpoint is a secure host).

2.2 Three-Way Association Terminations

The SCTP also uses the 3-way handshake mechanism for termination of an SCTP association for the purpose of the graceful close (shutdown). It is noted that the TCP provides the 4-way connection termination scheme, as shown in Figure 2.

To terminate an association, the endpoint A may send a SHUTDOWN chunk to the endpoint B. If there is no data to send, the endpoint B will respond with the SHUTDOWN-ACK chunk to the A. Finally, the endpoint A completes the association termination by sending the SHUTDOWN-COMPLETE chunk to the endpoint B.

Differently from the TCP, the SCTP does not support the so-called ‘half-open state’, wherein one side may continue sending data while the other end is closed.

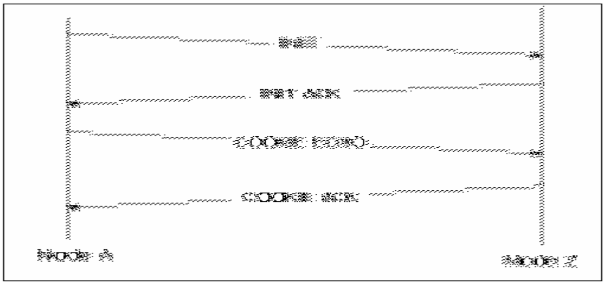


Fig. 1. SCTP four-way association establishment

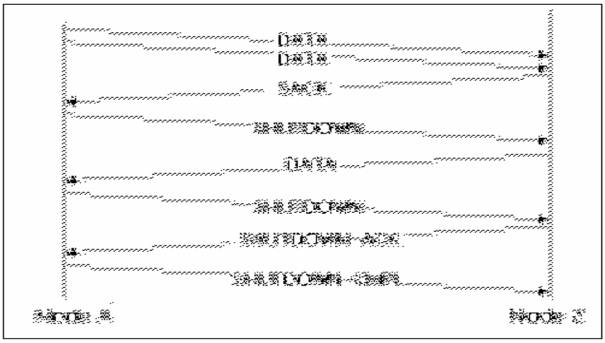


Fig. 2. SCTP three-way association terminations

2.3 SCTP Multi-streaming

The multi-streaming is a distinctive feature of SCTP. The SCTP user may assign each datagram to one of multiple streams within an association. In the association establishment phase, the two SCTP endpoints will exchange the number of available streams in the association each other. For each stream in the association, the SCTP increases the Stream Sequence Number (SSN) for the data chunk generated by the application user, as shown in Figure 3.

These SSN numbers are used by the receiver to determine the sequence of delivery. The SCTP performs in-sequence delivery per stream. This mechanism helps to avoid the head-of-line (HoL) blocking of TCP, since each stream data can be independently delivered to the peer endpoint within one association.

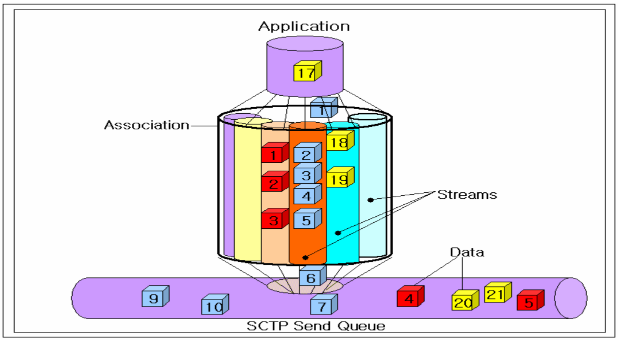


Fig. 3. SCTP multi-streaming

2.4 SCTP Multi-homing

From the multi-homing feature, the SCTP endpoint can use one or more IP addresses for data transport in the association, as shown in Figure 4.

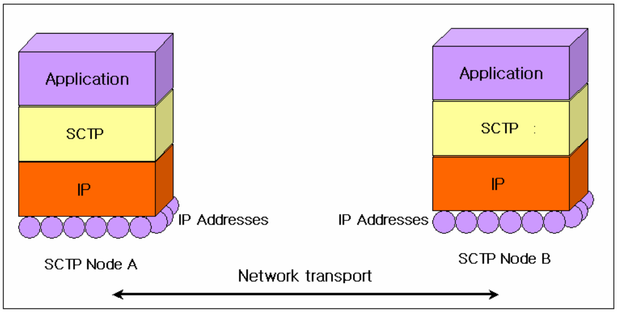


Fig. 4. SCTP multi-homing

The SCTP multi-homing feature can be used to protect an association from potential network failures by steering traffic to alternate IP addresses. During the initiation of an association, SCTP endpoints exchange the lists of IP addresses used at the remote endpoint. One of the listed IP addresses will be designed as the primary address. If the primary address repeatedly drops chunks, however, all chunks will be transmitted to an alternate address.

3 Test Scenarios

In this section, we describe the test scenarios employed in the experimentations for comparison of the SCTP and TCP performance.

3.1 Scenario 1: Different Size of User Input Data

The first test scenario is employed to compare the throughput of the SCTP and TCP for the different size of the use input data in the socket system call.

For the test purpose, a test network is configured as shown in Figure 5.

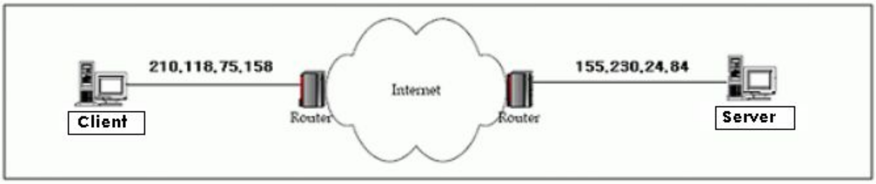


Fig. 5. Network configuration for Scenario 1

In the figure the client and server hosts are equipped with the Linux-Kernel 2.6.10 and LK-SCTP toolkit [4]. After establishing an SCTP association with the server, the client begins to download a file of 100 Mbytes from the server. As a performance metric, we measured the throughput of data transmission (i.e., the totally transmitted data bytes during the association period).

3.2 Scenario 2: Competition of SCTP and TCP Traffic

This scenario is tested to see how fairly the SCTP and TCP traffics compete in the network. Both the SCTP and TCP connections are established at the same time between the client and the server, as shown in Figure 6.

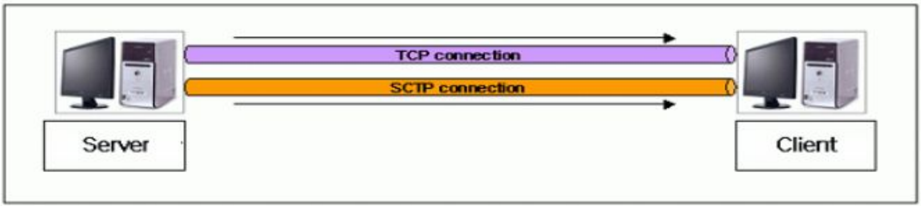


Fig. 6. Competition of SCTP and TCP traffic

For the two connections, we measured the traffic between the client and the server.

3.3 Scenario 3: Performance of SCTP Multi-homing

For the SCTP multi-homing, the test network is configured, as shown in Figure 7.

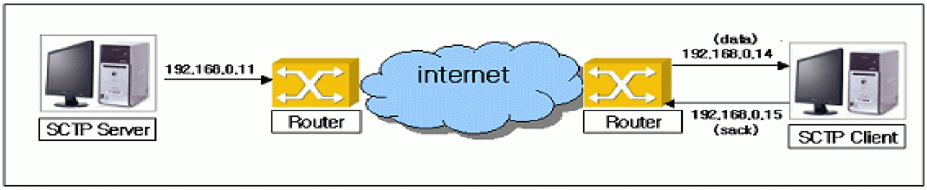


Fig. 7. Network configuration for SCTP multi-homing

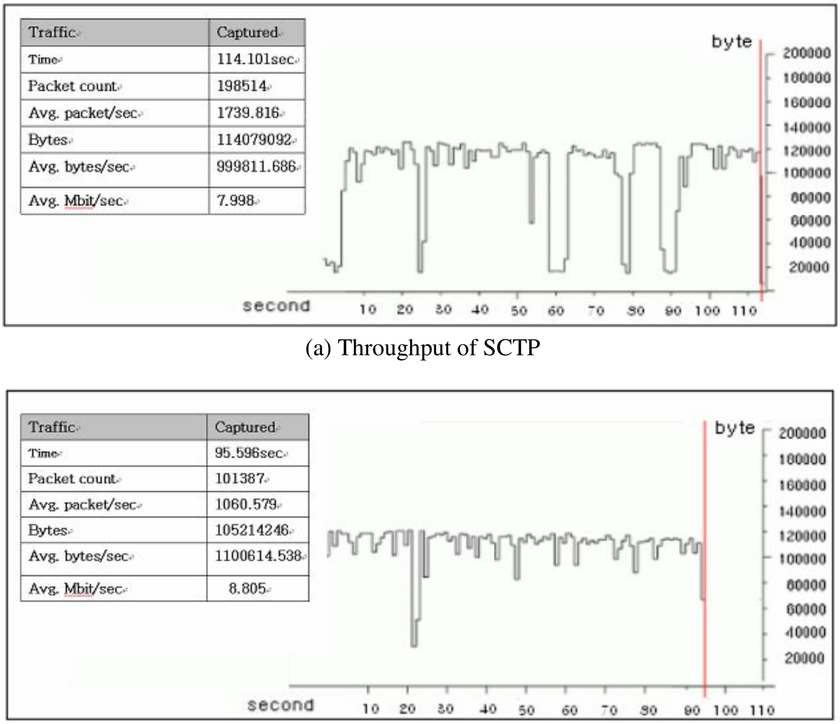
In the figure, the client is in the dual-homing state and uses the two different IP addresses for data packets and SACK packets, respectively.

4 Experimental Results

In this section, we discuss the results for the test experimentations of SCTP and TCP.

4.1 Results for Scenario 1

Figure 8 and 9 show the test results for the different sizes of user input data for each send() socket system call by using the 'ethereal' tool [6].



Traffic	Captured
Time	114.101sec
Packet count	198514
Avg. packet/sec	1739.816
Bytes	114079092
Avg. bytes/sec	999811.686
Avg. Mbit/sec	7.998

(a) Throughput of SCTP

Traffic	Captured
Time	95.596sec
Packet count	101387
Avg. packet/sec	1060.579
Bytes	105214246
Avg. bytes/sec	1100614.538
Avg. Mbit/sec	8.805

(b) Throughput of TCP

Fig. 8. Results for the user input data size of 2048 bytes

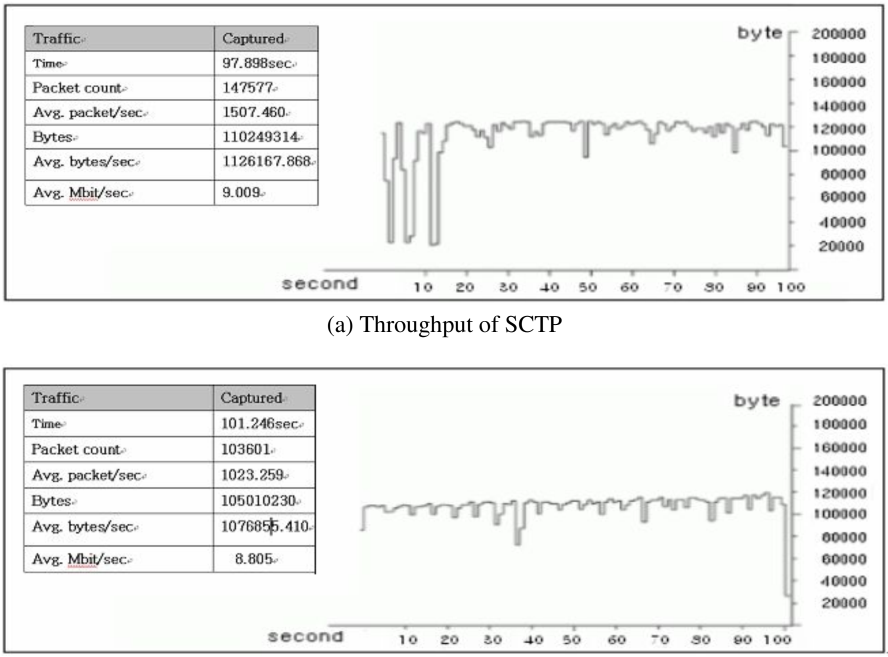
In Figure 8, we show the data packets (in byte) transmitted over the association period, for SCTP (Fig. 8(a)) and for TCP (Fig. 8(b)), in which the user input data of 2048 bytes are sent by the socket *send()* call.

In Fig. 8(a), we see that the SCTP transmits the total 198,514 packets and 114,079,092 bytes (including the data and control packets) over the association period of 114 seconds, which corresponds to the average throughput of 999,811 bytes per second.

On the other hand, we see in Fig. 8(b) that the TCP sends 101,387 packets over the connection period of 95 second, with the average throughput of 1,100,614 bytes per second. In summary, from the figure we see that the TCP provides better throughput than the SCTP for the user input data of 2,048 bytes.

In Figure 9, we show the results of the throughput for SCTP (Fig. 9(a)) and for TCP (Fig. 9(b)) with the user input data of 8,192 bytes.

It is noted that the results in Figure 9 are different from those in Figure 8. Fig. 9(a) shows that the SCTP gives the average throughput of 1,126,167 bytes per second, whereas Fig. 9(b) shows that the TCP provides the throughput of 1,076,685 bytes per second, for the user input data of 8,192 bytes.



Traffic	Captured
Time	97.898sec
Packet count	147577
Avg. packet/sec	1507.460
Bytes	110249314
Avg. bytes/sec	1126167.868
Avg. Mbit/sec	9.009

(a) Throughput of SCTP

Traffic	Captured
Time	101.246sec
Packet count	103601
Avg. packet/sec	1023.259
Bytes	105010230
Avg. bytes/sec	1076855.410
Avg. Mbit/sec	8.805

(b) Throughput of TCP

Fig. 9. Results for the user input data size of 8192 bytes

From the results of Figure 8 and 9, it is interesting to note that the SCTP tends to provide better throughput performance over the TCP, when the size of the user input data for each socket system call gets larger. That is, the SCTP performance will benefit from the transport of the large bulk data, compared to TCP.

On the other hand, this performance gain of SCTP over TCP seems to come from the congestion control schemes associated [7, 8]. That is, the TCP uses the initial *Congestion Window* (*CWND*) as 1*MTU, whereas the SCTP starts from the *CWND* of 2*MTU. Overall, the SCTP tends to provide better throughput than TCP for the large-scale bulk data transport.

4.2 Results for Scenario 2

Figure 10 shows the results of the traffic traces for SCTP and TCP, in which the two SCTP and TCP connections are activated at the same time in the single computer.

From the figure, we see that the SCTP competes with the TCP for the data transmission under the same condition, in which the traffic generated by SCTP and TCP is almost equally distributed. The TCP connection completes the data transmission earlier than the SCTP, since the SCTP generates more data and control chunks.

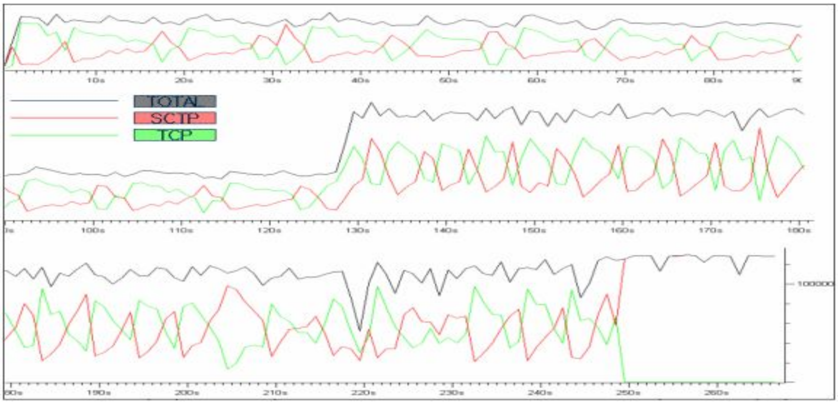
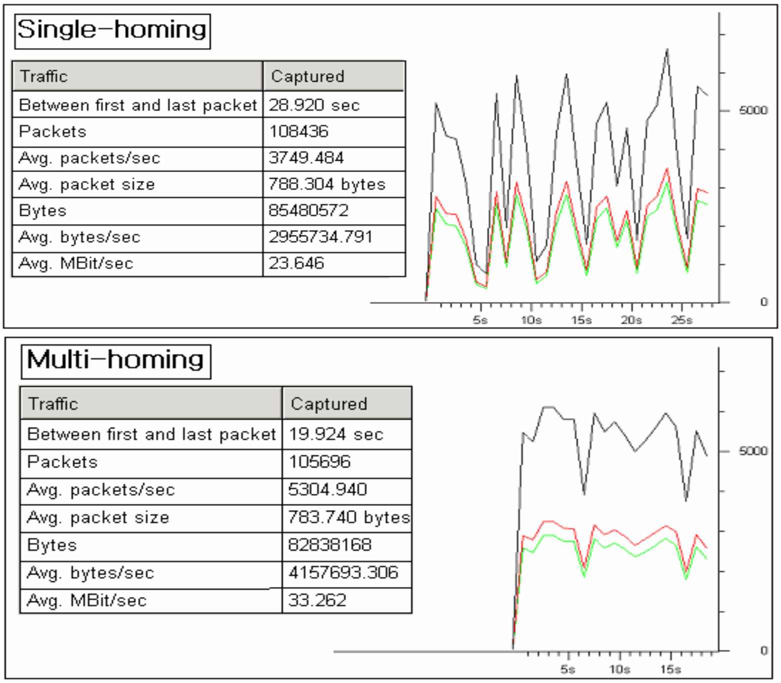


Fig. 10. Results for competition of TCP and SCTP traffic

4.3 Results for Scenario 3

Figure 11 shows the results of the SCTP single-homing and multi-homing association, as shown in Figure 7. It is noted in the multi-homing SCTP that the data and control SACK chunks are delivered over the different IP addresses [9].



Single–homing

Traffic	Captured
Between first and last packet	28.920 sec
Packets	108436
Avg. packets/sec	3749.484
Avg. packet size	788.304 bytes
Bytes	85480572
Avg. bytes/sec	2955734.791
Avg. MBit/sec	23.646

Multi–homing

Traffic	Captured
Between first and last packet	19.924 sec
Packets	105696
Avg. packets/sec	5304.940
Avg. packet size	783.740 bytes
Bytes	82838168
Avg. bytes/sec	4157693.306
Avg. MBit/sec	33.262

Fig. 11. Effects of SCTP single-homing and multi-homing

From the figure, we see that the multi-homing SCTP completes the data transmission earlier, with the better throughput of 4,157,693 bytes per second, than the single-homing SCTP (2,955,734 bytes per second).

It is clear from the results that the SCTP multi-homing feature can be used to improve the throughput of the data transmission. In this experiment, the SCTP control chunks are delivered using the different IP address from the SCTP data chunks.

5 Conclusion

In this paper, we have described the comparison of SCTP and TCP in the viewpoint of the throughput performance over the Linux platform. We compare the throughput of SCTP and TCP for the three different test scenarios: the performance comparison of SCTP and TCP for the different size of the user input data for the socket system call, the analysis of the fairness under competition of SCTP and TCP traffic, and the performance comparison of the SCTP multi-homing and single-homing cases.

From the results, it is shown that the SCTP provides better throughput over TCP for a larger user input data. We also see that the SCTP traffic tends to compete fairly with TCP, and that the multi-homing SCTP provides better performance than the single-homing case.

References

1. Stewart, R., et al.: Stream Control Transmission Protocol. RFC 2960, October 2000
2. Jungmajer, M schopp and M. Tuxen.: Performance Evaluation for the Stream Control Transmission Protocol. IEEE ATM Workshop 2000, June 2000
3. Ravier, T., et al.: Experimental studies of SCTP multi-homing. First Joint IEI/IEE Symposium on Telecommunications Systems Research, 2001
4. Linux Kernel SCTP Project. Available from http://lksctp.sourceforge.net/
5. Stewart, R., et al.: Sockets API Extensions for Stream Control Transmission Protocol. IETF Internet Draft, draft-ietf-tsvwg-sctpsocket-10.txt, Feb. 2005
6. Ethereal, available from http://www.ethereal.com
7. Allman, M., et al.: TCP Congestion Control. RFC 2581, April 1999
8. J. Hoe.: Improving the Startup Behavior of a Congestion Control Scheme for TCP. ACM SIGCOMM, August 1996
9. Koh, S., et al.: mSCTP for Soft Handover in Transport Layer. IEEE Communications Letters, Vol. 8, No.3, pp.189-191, March 2004

Multiresolution Fusion Estimation of Dynamic Multiscale System Subject to Nonlinear Measurement Equation

Peiling Cui[1,2], Quan Pan[2], Guizeng Wang[1], and Jianfeng Cui[3]

[1] Department of Automation, Tsinghua University, Beijing 100084, China
ljhcpl@263.net
[2] Department of Automatic Control, Northwestern Polytechnical University,
Xi'an, Shaanxi 710072, China
[3] ZhengZhou Institute of Aeronautical Industry Management, Zhengzhou,
Henan, 450015, China

Abstract. Fusion of the states of a nonlinear dynamic multiscale system (DMS) on the basis of available noisy measurements is one of the well-known key problems in modern control theory. To the best of our knowledge, all of the previous work focused attention on linear DMS. However, nonlinear DMS has never been investigated. In this paper, modeling and fusion estimation of dynamic multiscale system subject to nonlinear measurement equation is proposed. Haar wavelet is used to link the scales. Monte Carlo simulation results demonstrate that the proposed algorithm is effective and powerful in this kind of nonlinear dynamic multiscale system estimation problem.

1 Introduction

In multiscale system theory, a well-known achievement is the multiscale stochastic model that stems from the work of Willsky *et al* and that allows the modeling of multiresolution data at different levels in the bintree [1-2]. This model has made success in a number of applications [3-8]. Motivated by the success of this methodology in solving static estimation problems, a dynamic algorithm is proposed by propagating the static estimator over time with alternating update and prediction steps in a manner analogous to Kalman filtering [9-10]. In [11-13], measurements available at multiple resolution levels are integrated by the wavelet transform to deal with the target tracking by L. Hong. In [14], an optimal estimation of a class of dynamic multiscale system (DMS) is discussed. The sampling frequencies of the sensors are supposed to decrease by a factor of two.

Methods mentioned above are unrealistic for many real-world signals because they all assumed that the system is linear. In fact, this is a real limitation in many practical situations. For example, the observation model of Radar or IR sensor is linear in a spherical coordinate system. However, the state model of target is best described in a rectangular coordinate system. This results in either the state model or observation model being nonlinear in the same coordinate system. Up to now, the nonlinear estimation problem of dynamic multiscale system is still open. In this paper, the multiresolution fusion estimation method of a class of dynamic multiscale system subject to nonlinear measurement equation is given.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 405 – 413, 2005.

2 Haar-Wavelet-Based Modeling and Estimation Algorithm

For convenience, let the sampling rate decrease from sensor 1 to sensor J by a factor of two. Obviously, sensor 1 corresponds to the finest scale. The state at all scales in time interval ΔT is called a state block, and the measurement a data block. In every ΔT , the state estimation must be updated when a new data block is available. We hope the approximation of any node at all scales is accomplished in time interval ΔT , not using the state nodes outside of it. We choose the Haar wavelet [15]. This choice is motivated by the particularly simple realization of the Haar wavelet transform in our multiscale framework by using a bintree structure. Haar wavelet is the simplest and most widely used one with low-pass filter $\left[\sqrt{2}/2,\ \sqrt{2}/2\right]$.

For clarity, we unify the notations of state nodes firstly. Fig. 1 shows the system bintree structure at time interval $k\Delta T$. $x_J(k)$ is denoted as the state of scale J, $x_{J-1}(2k)$ and $x_{J-1}(2k+1)$ the states at scale $J-1$. Analogically, at scale j there are 2^{J-j} state nodes, wh0ich are denoted as $x_j(2^{J-j}k)$, $x_j(2^{J-j}k+1) \ldots x_j\left(2^{J-j}(k+1)-1\right)$. Assuming that the multiscale system state structure satisfies the dyadic structure of Haar wavelet, the node $x_j(2^{J-j}k+m_j)$ can be expressed with the finest scale nodes as follows[14]

$$x_j(2^{J-j}k+m_i)=\left(\sqrt{2}\big/2\right)^{j-1}\sum_{i=0}^{2^{j-1}-1}x_1(2^{J-1}k+2^{j-1}m_j+i)\ . \tag{1}$$

where $m_j=0,1,\ldots,2^{J-j}-1; j=1,2,\cdots,J$.

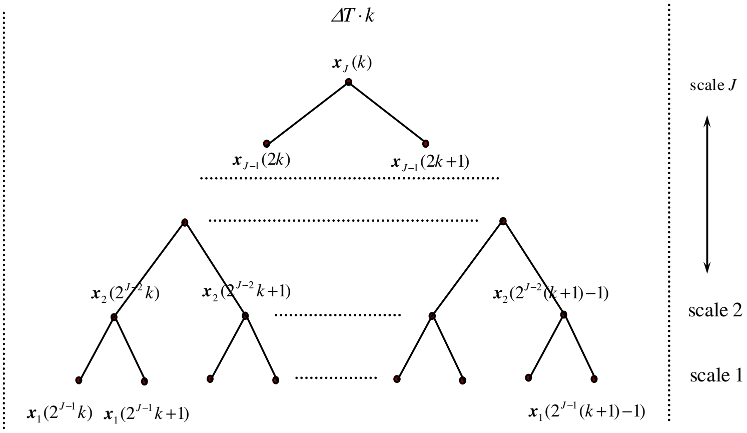


Fig. 1. Tree structure of the dynamic multiscale system state nodes at time interval $k\Delta T$

2.1 State Equation[14]

For simplicity, the system is assumed to be time-invariant. The discrete time state transition equation of the finest scale at time interval $(k+1)\Delta T$ is

$$x_1(2^{J-1}(k+1)) = Ax_1(2^{J-1}(k+1)-1) + Bw(2^{J-1}(k+1)-1) \ . \tag{2}$$

where $x_1(\cdot) \in R^{N_x}$, state transition matrix $A \in R^{N_x \times N_x}$, noise stimulus matrix $B \in R^{N_x \times u}$, $w(\cdot) \in R^u$ is white noise with variance q. Then

$$x_1(2^{J-1}(k+1)+1) = A^2 x_1(2^{J-1}(k+1)-1) + ABw(2^{J-1}(k+1)-1) + Bw(2^{J-1}(k+1)) \ . \tag{3}$$

$$\vdots$$

$$x_1(2^{J-1}(k+1)+m_1) = A^{m_1+1} \cdot x_1(2^{J-1}(k+1)-1) + \sum_{i=-2}^{m_1-2} A^{m_1-i-1} Bw(2^{J-1}(k+1)+i) \ . \tag{4}$$

where $m_1 = 0,1,...,2^{J-1}-1$. Letting

$$\overline{x}(k) = col(x_1(2^{J-1}k), x_1(2^{J-1}k+1), \cdots, x_1(2^{J-1}(k+1)-1)) \ . \tag{5}$$

$$\overline{w}(k) = col\left(w(2^{J-1}(k+1)-1), w(2^{J-1}(k+1)), ..., w(2^{J-1}(k+2)-2) \right) \ . \tag{6}$$

$$\overline{A}(m_1) = A^{m_1+1}, \ \overline{B}(m_1) = [A^{m_1}B, A^{m_1-1}B, \ldots, B, \mathrm{O}, \ldots, \mathrm{O}] \ . \tag{7}$$

where col denotes arranging the data in the bracket into column vector, $\overline{x}(k) \in R^{2^{J-1}N_x \times 1}$, $\overline{w}(k) \in R^{2^{J-1}u \times 1}$, $\overline{A}(m_1) \in R^{N_x \times N_x}$, and $\overline{B}(m_1) \in R^{N_x \times 2^{J-1}u}$ is with zero elements on the last $(2^{J-1} - m_1 - 1) \times u$ columns. Letting

$$\overline{A} = \begin{bmatrix} \mathrm{O} & \cdots & \mathrm{O} & \overline{A}(0) \\ \mathrm{O} & \cdots & \mathrm{O} & \overline{A}(1) \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{O} & \cdots & \mathrm{O} & \overline{A}(2^{J-1}-1) \end{bmatrix}, \ \overline{B} = \begin{bmatrix} \overline{B}(0) \\ \overline{B}(1) \\ \vdots \\ \overline{B}(2^{J-1}-1) \end{bmatrix} . \tag{8}$$

where $\overline{A} \in R^{2^{J-1}N_x \times 2^{J-1}N_x}$, $\overline{B} \in R^{2^{J-1}N_x \times 2^{J-1}u}$, then we have

$$\overline{x}(k+1) = \overline{A}\overline{x}(k) + \overline{B}\overline{w}(k) \ . \tag{9}$$

2.2 Measurement Equation

Suppose that discrete measurement equation at scale j is

$$z_j(2^{J-j}k + m_j) = g_j\left(x_j(2^{J-j}k + m_j)\right) + v_j(2^{J-j}k + m_j) \ . \tag{10}$$

where the dimension of $z_j(\cdot)$ is N_z, $g_j(\cdot)$ is a nonlinear function. $v_j(\cdot)$ is Gaussian white noise with zero mean and variance $R_j(\cdot)$, and is uncorrelated with $w(\cdot)$. Letting

$$M_j(m_j)=\begin{bmatrix}\underbrace{0\cdot\mathrm{I},...,0\cdot\mathrm{I}}_{m_j 2^{j-1}} & \underbrace{\left(\sqrt{2}/2\right)^{j-1}\cdot\mathrm{I},...,\left(\sqrt{2}/2\right)^{j-1}\cdot\mathrm{I}}_{2^{j-1}} & \underbrace{0\cdot\mathrm{I},...,0\cdot\mathrm{I}}_{\left(2^{J-j}-m_j-1\right)2^{j-1}}\end{bmatrix} . \tag{11}$$

$$\overline{x}(k)=col\left(x_1(2^{J-1}k),x_1(2^{J-1}k+1),...,x_1(2^{J-1}(k+1)-1)\right) . \tag{12}$$

then

$$x_j(2^{J-j}k+m_j)=M_j(m_j)\overline{x}(k) . \tag{13}$$

and then

$$z_j(2^{J-j}k+m_j)=g_j\left(M_j(m_j)\overline{x}(k)\right)+v_j(2^{J-j}k+m_j) . \tag{14}$$

defining

$$\overline{z}_j(k)=col\left(z_j(2^{J-j}k),z_j(2^{J-j}k+1),...,z_j(2^{J-j}(k+1)-1)\right) . \tag{15}$$

$$\overline{G}_j\left(\overline{x}(k)\right)=\begin{bmatrix} g_j\left(M_j(0)\overline{x}(k)\right) \\ g_j\left(M_j(1)\overline{x}(k)\right) \\ \vdots \\ g_j\left(M_j(2^{J-j}-1)\overline{x}(k)\right)\end{bmatrix}, \overline{v}_j(k)=\begin{bmatrix} v_j(2^{J-j}k) \\ v_j(2^{J-j}k+1) \\ \vdots \\ v_j(2^{J-j}(k+1)-1)\end{bmatrix} . \tag{16}$$

It can be seen that, $\overline{G}_j\left(\overline{x}(k)\right)$ is the nonlinear function of $\overline{x}(k)$, the covariance of $\overline{v}_j(k)$ is

$$\overline{R}_j(k)=diag\left[R_j(□),R_j(□),...,R_j(□)\right] . \tag{17}$$

then

$$\overline{z}_j(k)=\overline{G}_j\left(\overline{x}(k)\right)+\overline{v}_j(k) . \tag{18}$$

defining

$$\overline{z}(k)=col\left(\overline{z}_J(k),\overline{z}_{J-1}(k),...,\overline{z}_1(k)\right) . \tag{19}$$

$$\overline{G}\left(\overline{x}(k)\right)=col\left(\overline{G}_J\left(\overline{x}(k)\right),\overline{G}_{J-1}\left(\overline{x}(k)\right),....,\overline{G}_1\left(\overline{x}(k)\right)\right) . \tag{20}$$

$$\overline{v}(k)=col\left(\overline{v}_J(k),\overline{v}_{J-1}(k),...,\overline{v}_1(k)\right) . \tag{21}$$

then

$$\overline{z}(k)=\overline{G}\left(\overline{x}(k)\right)+\overline{v}(k) . \tag{22}$$

It can be seen that, $\bar{G}(\bar{x}(k))$ is the nonlinear function of $\bar{x}(k)$. The covariance of white noise $\bar{v}(k)$ is

$$\bar{R}(k) = diag\left[\bar{R}_J(k), \bar{R}_{J-1}(k), ..., \bar{R}_1(k)\right] . \tag{23}$$

then the model of new system after augmentation is

$$\begin{cases} \bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}\bar{w}(k) \\ \bar{z}(k) = \bar{G}\left(\bar{x}(k)\right) + \bar{v}(k) \end{cases} . \tag{24}$$

where $\bar{w}(k)$ and $\bar{v}(k)$ are white noise and are uncorrelated. The above equation is a pair of nonlinear equation, and the estimation value of $\bar{x}(\square)$ can be obtained by performing nonlinear filtering.

The estimation at coarser scales can be obtained from $\hat{\bar{x}}(k)$ directly, and the estimation of node $x_j(2^{J-j}k + m_j)$ is $M_j(m_j) \cdot \hat{\bar{x}}(k)$ [14].

3 Simulation Results

For verifying the validity of our algorithm, consider the following constant-velocity dynamic system with measurements at two scales. The state equation at the finest scale (scale 1) is

$$x_1(2k+2) = Ax_1(2k+1) + Bw(2k+1) . \tag{25}$$

where

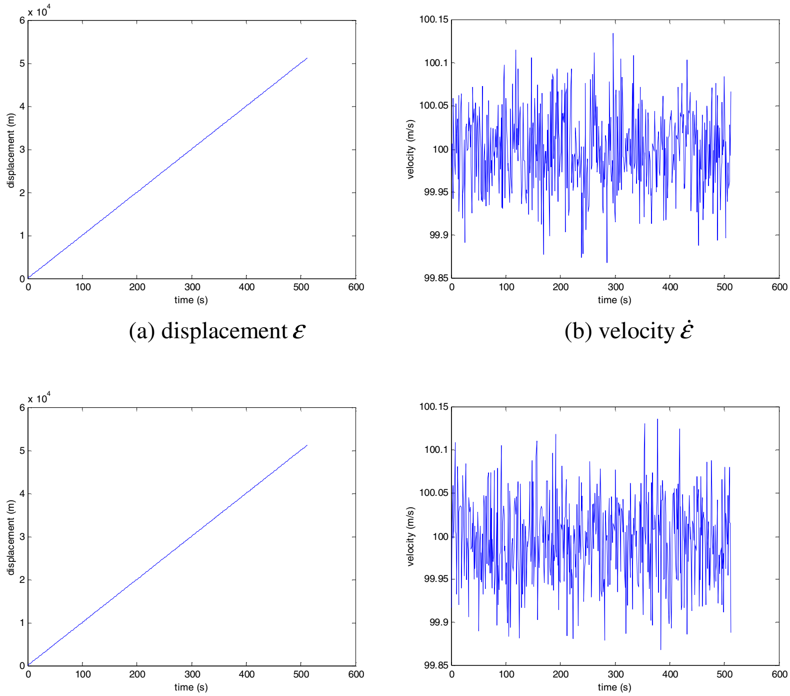
$$A = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0.5 \cdot T^2 & 0 \\ T & 0 \\ 0 & 0.5 \cdot T^2 \\ 0 & T \end{bmatrix} . \tag{26}$$

state vector $x_1 = \left[\varepsilon \quad \dot{\varepsilon} \quad \eta \quad \dot{\eta}\right]$, ε and η represent the displacements of two coordinates, $w(\square)$ is Gaussian white noise with zero mean and its variance is $Q(\square)$.

The measurements include radial distance and angle, and the measurement equation at scale j is

$$z_j(2^{2-j}k + m_j) = \begin{bmatrix} \sqrt{\left(\varepsilon_j(2^{2-j}k + m_j)\right)^2 + \left(\eta_j(2^{2-j}k + m_j)\right)^2} \\ \arctan\left(\eta_j(2^{2-j}k + m_j) / \varepsilon_j(2^{2-j}k + m_j)\right) \end{bmatrix} + v_j(2^{2-j}k + m_j) . \tag{27}$$

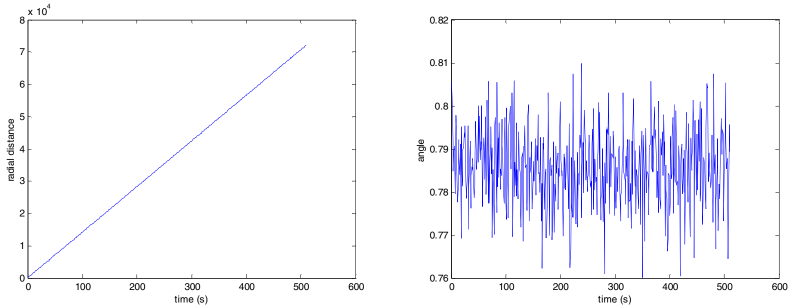
where $m_j = 0, 1, \ldots, 2^{2-j} - 1; j = 1, 2$. The measurement noise of radial distance and angle at scale j are zero-mean Gaussian white noise and with variances $R_j(\square)$ and $a_j(\square)$.



(a) displacement ε (b) velocity $\dot{\varepsilon}$

(c) displacement η (d) velocity $\dot{\eta}$

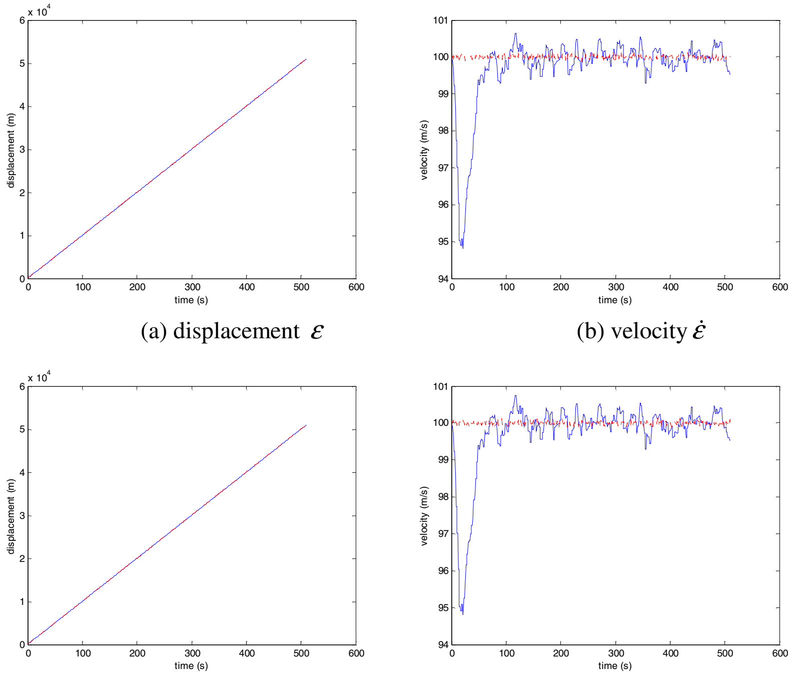
Fig. 2. True state at scale 1



(a) radial distance (b) angle

Fig. 3. Measurements at scale 1

Letting $T = 1\text{s}$, $Q = 100 \cdot I_2$, I_2 is 2×2 identity matrix, $R_1 = R_2 = 2500\, m^2 / s^2$, $a_1 = a_2 = 7.5 \times 10^{-5}$. The state is estimated with Extended Kalman Filter(EKF) by Monte Carlo simulation (200 runs).



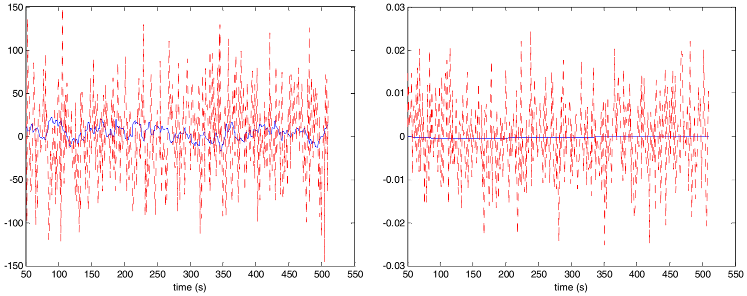
(a) displacement ε

(b) velocity $\dot{\varepsilon}$

(c) displacement η

(d) velocity $\dot{\eta}$

Fig. 4. True state (dotted) and the estimated state (solid 1) at scale



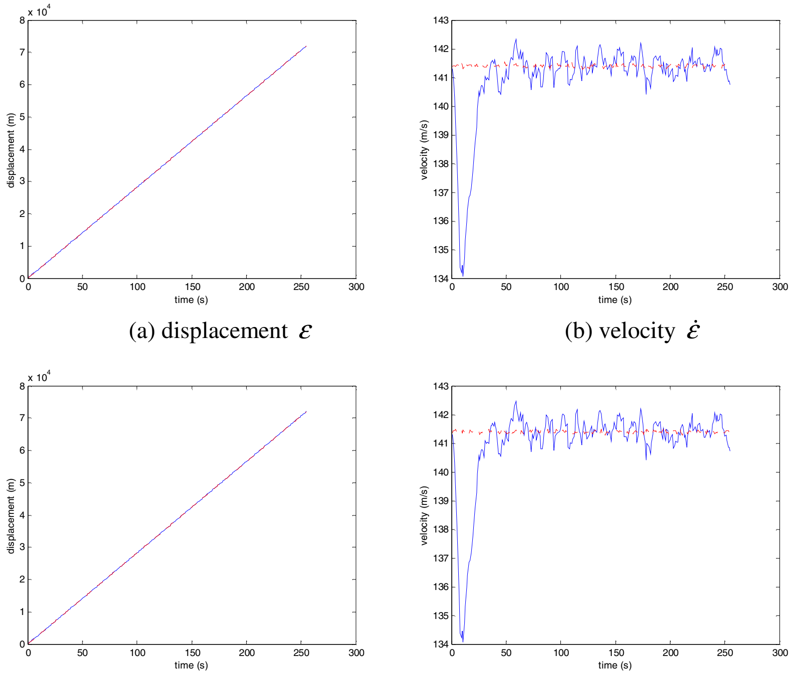
(a) radial distance

(b) angle

Fig. 5. Measurement noise (dotted) and estimation error (solid) at scale 1

Fig. 2 shows a sequence of true state at scale 1. Fig. 3 shows the measurements at scale 1. Fig. 4 shows a sequence of the true state and the estimated state at scale 1. Fig. 5 compares the measurement noise with the estimation error at scale 1. The noise compression ratio of radial distance is 15.8dB, and the noise compression ratio of angle is 30.6dB.

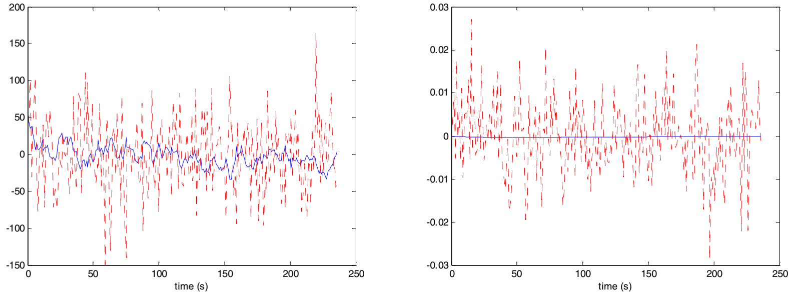
Fig. 6 shows a sequence of the true state and the estimated state at scale 2. Fig. 7 compares the measurement noise with the estimation error at scale 2. The noise compression ratio of radial distance is 11.7dB, the noise compression ratio of angle is 30.6dB.



(a) displacement ε (b) velocity $\dot{\varepsilon}$

(c) displacement η (d) velocity $\dot{\eta}$

Fig. 6. True state (dotted) and the estimated state (solid) at scale 2



(a) radial distance (b) angle

Fig. 7. Measurement noise (dotted) and estimation error (solid) at scale 2

It can be seen that, the estimation error is smaller than the measurement noise at each scale because of the fusion of measurement information at two scales.

4 Conclusion

In this paper, modeling and fusion estimation of a class of nonlinear dynamic multiscale system is proposed. The measurement equation is nonlinear, and the system is observed by multiresolution multisensors. Haar wavelet is used to link the scales, and a centralized model is built. Monte Carlo simulation results verify the validation of our algorithm.

Acknowledgement

This research has been partially supported by National Natural Science Foundation of China under grant 60172037.

References

1. Basseville, M., Benveniste, A., Chou, K., Golden, S., Nikoukhah, R., Willsky, A. S.: Modeling and Estimation of Multiresolution Stochastic Processes. IEEE Transactions on Information Theory (1992) 38(2): 766–784
2. Chou, K., Willsky, A. S., Benveniste, A.: Multiscale Recursive Estimation, Data Fusion, and Regularization. IEEE Transactions on Automatic Control (1994) 39(3): 464–478
3. Tsai, A., Zhang, J., Willsky, A. S.: Expectation-Maximization Algorithms for Image Processing using Multiscale Models and Mean-Field Theory, with Applications to Laser Radar Range Profiling and Segmentation. Optical Engineering (2001) 40(7): 1287–1301
4. Schneider, M. K., Fieguth, P. W., Karl, W. C., Willsky, A. S.: Multiscale Methods for the Segmentation and Reconstruction of Signals and Images. IEEE Transactions on Image Processing (2000) 9(3): 456–468
5. Daniel, M., Willsky, A. S.: The Modeling and Estimation of Statistically Self-Similar Processes in a Multiresolution Framework. IEEE Transactions on Information Theory (1999) 45(3): 955–970
6. Frakt A. B., Willsky, A. S.: Efficient Multiscale Stochastic Realization. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, Washington (1998) 4: 2249–2252
7. Fieguth, P., Karl, W., Willsky, A. S.: Efficient Multiresolution Counterparts to Variational Methods for Surface Reconstruction. Computer Vision and Image Understanding (1998) 70(2):157–176
8. Fosgate, C., Krim, H., Irving W. W., Willsky, A. S.: Multiscale Segmentation and Anomaly Enhancement of SAR Imagery. IEEE Transactions on Image Processing (1997) 6(1): 7–20
9. Ho, T.: Large-scale Multiscale Estimation of Dynamic Systems. Phd Thesis, Massachusetts Institute of Technology, September (1998)
10. Luettgen, M., Willsky, A. S.: Multiscale Smoothing Error Models. IEEE Transactions on Automatic Control (1995) 40(1): 173–175
11. Hong, L.: Multiresolution Distributed Filtering. IEEE Transactions on Automatic Control (1994) 39(4): 853–856
12. Hong, L., Cheng, G., Chui, C. K.: A Filter-Bank-Based Kalman Filtering Technique for Wavelet Estimation and Decomposition of Random Signals. IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing (1998) 45(2): 237–241
13. Hong, L., Scaggs, T.: Real-time Optimal Filtering for Stochastic Systems with Multiresolutional Measurements. Systems & Control Letters (1993) 20: 381–387
14. Zhang, L.: The Optimal Estimation of a class of Dynamic Multiscale Systems. PhD thesis, Northwestern Polytechnic University, Xi'an, PRC (2001)
15. Daubechies, I.: Ten Lectures on Wavelets. CBMS-NSF Series in Appl. Math., SIAM (1992)

On a Face Detection with an Adaptive Template Matching and an Efficient Cascaded Object Detection

Jin Ok Kim[1], Jun Yeong Jang[2], and Chin Hyun Chung[2]

[1] Faculty of Multimedia, Daegu Haany University,
290, Yugok-dong,Gyeongsan-si, Gyeongsangbuk-do, 712-715, Korea
bit@dhu.ac.kr

[2] Department of Information and Control Engineering, Kwangwoon University,
447-1, Wolgye-dong, Nowon-gu, Seoul, 139-701, Korea
chung@kw.ac.kr

Abstract. We present a method for a template matching and an efficient cascaded object detection. The proposed method belongs to wide criteria which can regard to the "feature-centric". Furthermore, the proposed cascade method has some merits to the face changes. The proposed method for an object detection uses to find the object to most approach better than to find the object to correspond completely. Therefore, this method can use to detect the many faces mixed with different objects. We expect that the result of this paper can be contributed to develop more detection methods and recognition system algorithm.

1 Introduction

Traditionally, computer vision systems have been used in specific tasks such as performing tedious and repetitive visual tasks of assembly line inspection[1]. Current trend in computer vision is moving toward generalized vision applications. For example, face recognition and video coding techniques etc.. Many of the current face recognition techniques assume the availability of frontal faces of similar sizes[1]. Most face recognitions are achieved practically in such condition. Consequently, we must consider such condition for efficient face recognition. As well, we must consider efficient recognition about various objects. However, in computer vision is not easy. The solution to the problem involves segmentation, extraction, and verification of faces and possibly facial features from an uncontrolled background[1]. To be concrete, the problem of such object recognition uses a object image of two dimension normally to the input and use the output to assort who that image is. For example, a face image of the three dimension is to be reflected to the two dimension. Accordingly, the information deeply, magnitude, rotation etc. has the loss of important many information in the recognition. The recognition comes to be difficult basically as the pattern due to the complication which the object has and illumination, background and the complication of the environment etc.. The object recognitions of comprehensive concept must

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 414–422, 2005.

accomplish the course to find the location of the object in a two dimension image of the random above all. Afterwards, preprocessing course of the back of a noise removal is performed. And also, normalization course of a object image comes to be continuously performed which set the size of the object or location with inside the image's size to want to the location. Fig.1 pictures are typical test images used in face classification research. The background in Fig.1 images is necessary for face classification techniques[1]. However, with the face of the Fig.2 could exist in a complex background and in many different positions[1][2].

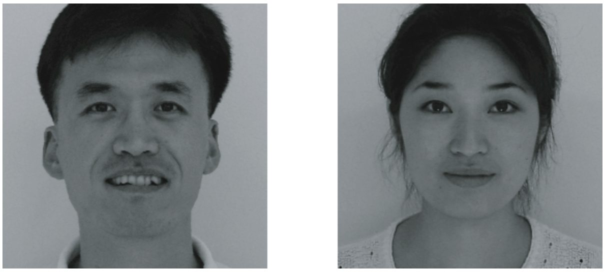


Fig. 1. Typical training images for face recognition

Fig. 2. A realistic face detection scenario

2 Detection Algorithm for Recognition

The active shape models can express which it appears in the Fig.3[3]. They can distinguish generally to three types. The first type uses a generic active contour called snakes, first introduced by Kass et al. in 1987[4]. Deformable templates were then introduced by Yuille et al.[5] to better the performance of snakes. Cootes et al.[6] later proposed the use of a new generic flexible model which they termed smart snakes and PDM to provide an efficient interpretation of the human face.

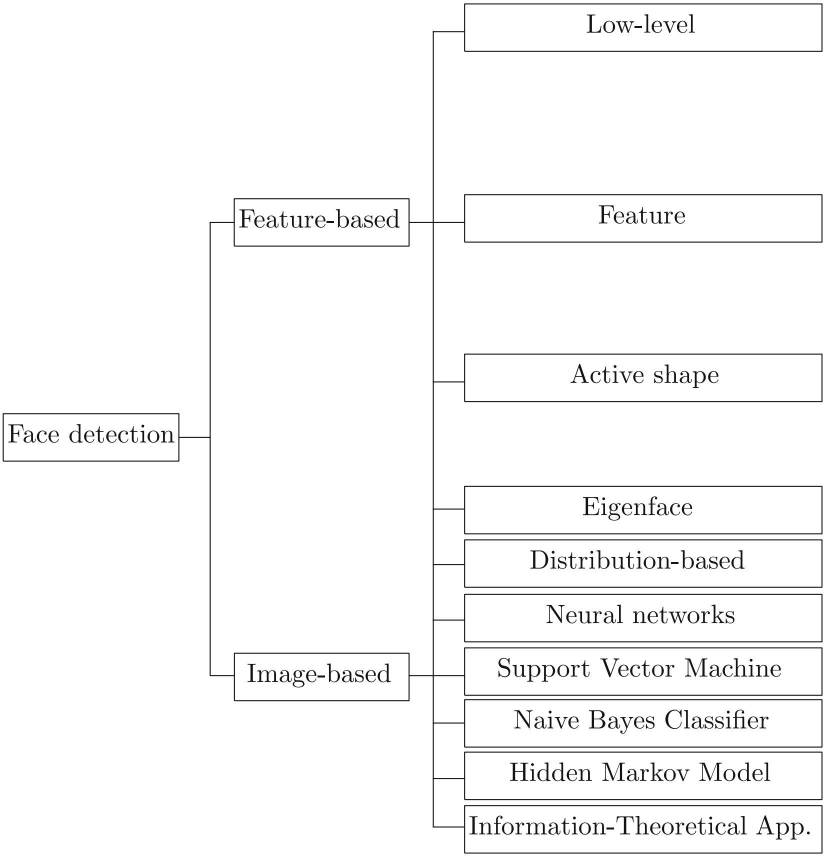


Fig. 3. Schematic depiction of a the detection cascade

3 Cascaded Method Detection for Object Recognition

The cascaded method for object detection approach uses a novel organization of the first cascade stage called "feature-centric" like the templates.

One of the point of detection is coping with variation in object size and location. There are general two approach methods for this. The first is "Invariant" methods. These attempt to use features or filters that are invariant to geometry[7][8][9][10] or photometric properties[11][12][13]. Another method is "exhaustive-search". This method finds the object by scanning classifier over an exhaustive range of possible locations and scales in an image[14]. But the defect of this method has very time consuming to find the object to want. The method to do the supplementation the defect of the "exhaustive search" is the method that "the cascade of sub-classifiers" of Fig.4. The algorithm for constructing a cascade of classifiers achieves increased detection performance while radically reducing time consuming. Each sub-classifier stage makes a decision to reject or accept the input window. The window to be accepted goes to a next stage(next Sub-Classifier) and the window to be rejected goes to the classify as non-object. The window to remain in the last goes via such course is classified as the object. This is designed to remove many non-object windows to the computation of the minimum.

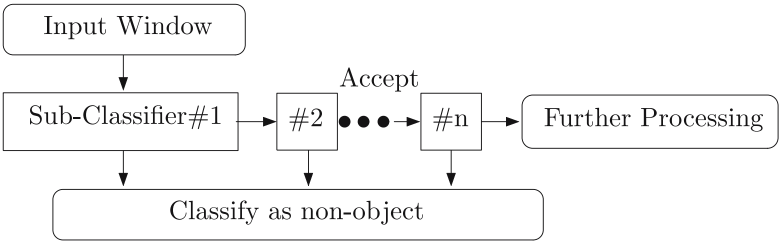


Fig. 4. Schematic depiction of a the detection cascade

The idea of using cascade-like methods has existed for several decades, and, in particular, was used widely in the automatic target recognition techniques of the 1970s and 80s[15].
Each sub-classifier is represented as a semi-naïve Bayes classifier[16].

$$\begin{aligned} H(X_1,\ldots,X_n) &= \log\frac{P\left(S_1|\omega_1\right)}{P\left(S_1|\omega_2\right)} + \log\frac{P\left(S_2|\omega_1\right)}{P\left(S_2|\omega_2\right)} + \ldots + \log\frac{P\left(S_m|\omega_1\right)}{P\left(S_m|\omega_2\right)} \\ &> \lambda \\ S_1,\ldots,S_m &\subset \{X_1,\ldots,X_n\} \end{aligned} \tag{1}$$

where $X_1,\ldots,X_n$ are the input variables within the classification window, $S_1,\ldots,S_r$ are subsets of these variables, and ω_1 and ω_2 indicate the two classes. If ω_1 is the face and ω_2 is the non-face, the classifier chooses class ω_1. Otherwise, it chooses class ω_2 (if $f(X_1,\ldots,X_n) < \lambda$).

Each stage in the cascade reduces the false detections rate and decreases the detection rate. In most classifiers including more features will achieve higher detection rates and lower false detections rates. At the classifiers with more features require more time to compute[17].

4 Template Matching Detection for Object Recognition

The template matching belongs to wide criteria which can regard to the "feature-centric". Minute explanation about this explains in a next section.

The template matching finds similar image pattern in the image inside to check to be given beforehand when the image was given. At this time, template

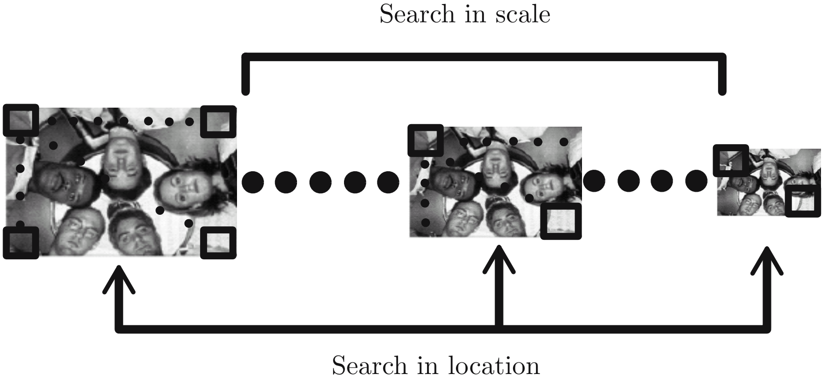


Fig. 5. Exhaustive search object detection

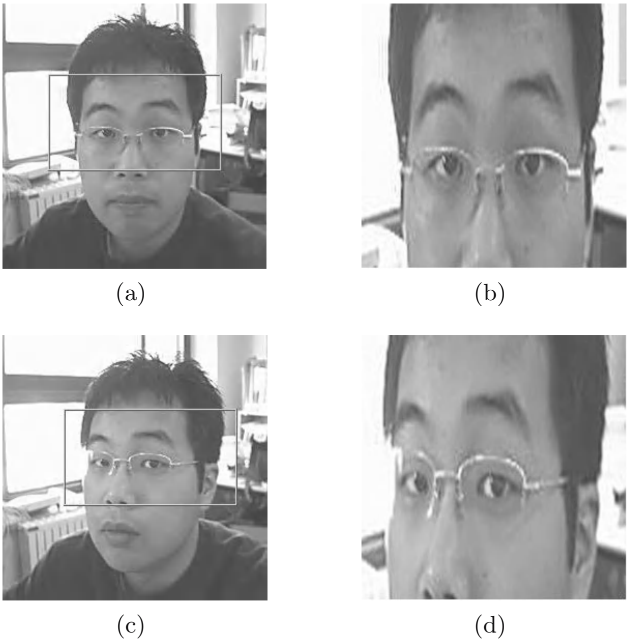
is the kind of a model image. We overlap small template of the image at the starting point on left corner which compare a template image with the part of the overlapping image to check. This comparison standard amount can choose so that it is suitable according to the purpose. After store comparison standard amount to be calculated, We shift again a one pixel to left which does the template. And we compare again a template image with the part of the overlapping image to check.

The template matching is important to well select the comparison standard amount. The determine the comparison standard amount has some kind of the subject to consider. It must be insensible at an image noise and at intensity variation. It must have also small computation quantity.
The current standard of template matching is based on computed in Fast Fourier Transform(FFT)[18]. This can be extended to shift of template by a suitable sampling of the template[18]. We can use generally two kinds method. The first method is MAD(Mean Absolute Difference) by

$$MAD = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N} \mid T\left(x_i, y_i\right) - I\left(x_i, y_i\right) \mid \tag{2}$$

where M and N are width and length of template image, $T(x_i, y_i)$ is template image, $I(x_i, y_i)$ is the overlapping image to check. The second method is MSE(Mean Square Error) by

$$MSE = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}[T\left(x_i, y_i\right) - I\left(x_i, y_i\right)]^2. \tag{3}$$



(a) (b)

(c) (d)

Fig. 6. (a) The image to set the template (The comparison standard amount) (b) Template image (c) The image to search (d) The image to compare a template image with the part of the overlapping image to check

If template and the overlapping image are similar each other, MAD or MSE will become computation near to zero. The other way, if they are different each other, the two value will grow bigger.

5 Experiments

The face image data and object data did not put the limit in the lighting for a face detection. Also, we did not consider to wear glasses and the size of face area. The data used to be holding at the laboratory and on MIT-CMU test set. We experimented with the gray scale image of 256×256 size on an pentium4, 1.7GHz processor.

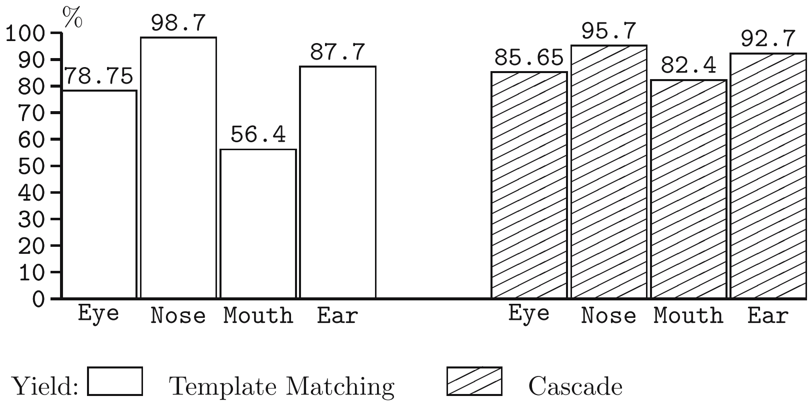
5.1 The Environment of the Cascaded Method

The complete face detection cascade has 38 stages with over 6000 features[17]. On a difficult data, containing 507 faces and 75 million sub-windows, faces are detected using of 10 feature evaluations per sub-window. This system is about 15 times faster than an implementation of the detection system constructed by Rowley et al.[17][19].

This experiments was processed also to two kinds as the template matching.

5.2 The Result of the Cascaded Method

The detection probability of the cascade method is superior generally but the template matching method fells off remarkably from the detection probability of the nose and eye. Because, this regards the open mouth of the appearance to smile wrong as the eye and is the experiment result of the case to wear glasses.



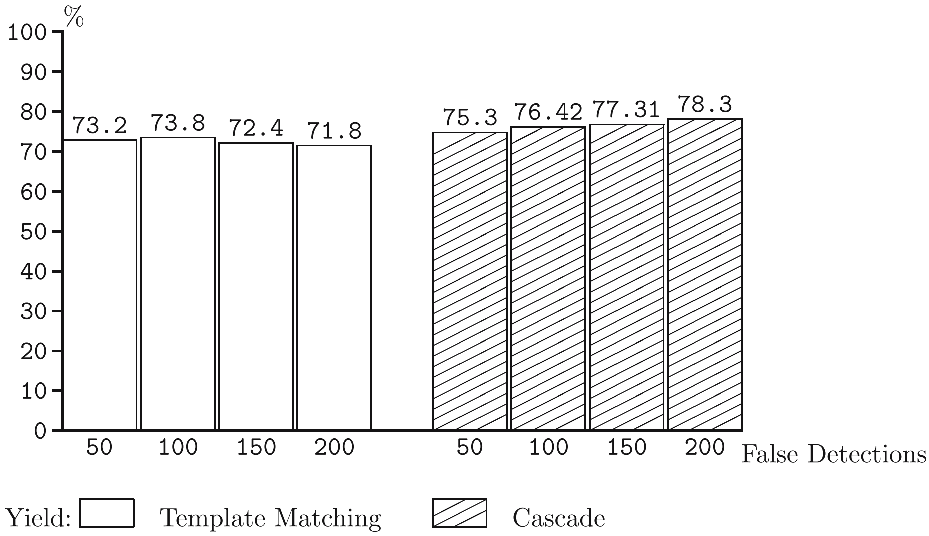
5.3 The Environment of the Template Matching

Like formula (1)(2), if the template of (M×N) size and the image to check of R×C size are given, the number to compare of the overlapping image happens (R-M)×(C-N) times. If size of template is 100×100 and size of image to check is 640×640, the number of the overlapping image happens 540×380 times. This is not little the number to overlapping. According to, it happens time complexity. If the time complexity is high, the computation time to similar pattern takes long.

We trained for finding the part of the face to want from image. The experiments was processed to two kinds. The first, the image used male of 56 persons and female of 51 persons including expression of four kinds. The template set in the sequence of eye, nose, mouth and ear. The second, it applies the template matching at the layer when the face of many people exists.

5.4 The Result of the Template Matching

This experiment's detection probability is lower than the first experiment. we can see high cascade method's detection probability as the false detections come to be high.



6 Conclusion

We propose a method for a template matching and cascaded object detection for an efficient face detection. The template matching method is superior to previous methods for face detections. Since the template matching and cascade method has an advantage to find the object better, it can find the object to correspond completely. Therefore, the method can detect many faces mixed with different objects better and can detect the various expressions of a face, provided that the cascade method can maximize the face detection probability. This paper has an advantage to detect many faces mixed with different objects better than to detect the various expression of face. The proposed method can contribute to more efficient detection methods and recognition systems.

References

1. Hjelmäs, E., Low, B.K.: Face Detection (A Survey). Computer Vision and Image Understanding (2001) 236–274
2. Yang, M.H.: Recent Advances In Face Detection. International Conference on Pattern Recognition (2004)
3. Ji, E.M., Yoon, H.S., Lee, S.H.: A New Face Detection Method Using Combined Features Of Color And Edge Under The Illumination Variance. Information Science (2002)
4. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. International Journal of Computer Vision **1** (1987) 321–331
5. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature Extraction From Faces Using Deformable Templates. IEEE Computer Vision and Pattern Recognition (1989) 104–109

6. Cootes, T.F., Taylor, C.J.: Active Shape Models-Smart Snakes. in Proc. of British Machine Vision Conference (1992) 266–275
7. Fergus, R., Perona, P., Zisserman, A.: Object Class Recognition By Unsupervised Scale-Invariant Learning. IEEE International Conference on Computer Vision and Pattern Recognition **2** (2003)
8. Forsyth, D.A.: Invariant Descriptors For 3D Recognition And Pose. IEEE Transactions on Pattern Analysis and Machine Intelligence (1991) 971–991
9. Wood, J.: Invariant Pattern Recognition: A Review. Electronics and Computer Science School of The University of Southampton on Pattern Recognition **29** (1996) 1–17
10. Zisserman, A., Forsyth, D., Mundy, J., Rothwell, C., Liu, J., Pillow, N.: 3D Object Recognition Using Invariance. Artificial Intelligence, Special Volume on Computer Vision **1-2** (1995) 239–288
11. Chen, H., Belhumeur, P., Jacobs, D.: In Search Of Illumination Invariants. IEEE International Conference on Computer Vision and Pattern Recognition (2000) 254–261
12. Nagao, K., Grimson, W.E.L.: Using Photometric Invariants For 3D Object Recognigion. Computer Vision and Image Understanding (1998) 74–93
13. Slater, D., Healey, G.: The Illumination-Invariant Recognition Of 3D Objects Using Local Color Invariants. IEEE Transactions on Pattern Analysis and Machine Intelligence (1996) 206–210
14. Schneiderman, H.: Feature-Centric Evaluation For Efficient Cascaded Object Detection. IEEE International Conference on Computer Vision and Pattern Recognition (2004) Robotics Institute, Carnegie Mellon University, Pittsburgh.
15. Bhanu, B.: Automatic Target Recognition: A State Of The Art Survey. IEEE Transactions Aerospace Electronic Systems (1986) 364–379 on Aerospace and Electronic Systems.
16. Kononenko, I.: Semi-Naïve Bayesian Classifier. Sixth European Working Session on Learning (1991) 206–219
17. Viola, P., Jones, M.: Rapid Object Detection Using A Boosted Cascade Of Simple Features. IEEE Computer Vision and Pattern Recognition **1** (2001) 511–518
18. Fredriksson, K., Ukkonen, E.: Faster Template Matching Without FFT. IEEE Image Processing **1** (2001)
19. Rowley, H., Baluja, S., Kanade, T.: Neural Network-Based Face Detection. IEEE Pattern Analysis and Machine Intelligence **20** (1998) 23–38

Joint Limit Analysis and Elbow Movement Minimization for Redundant Manipulators Using Closed Form Method

Hadi Moradi and Sukhan Lee

The School of Information and Communication,
Sungkyunkwan University
{moradi, lsh}@ece.skku.ac.kr

Abstract. Robot arms with redundant degrees of freedom (DOF) are crucial in service robots where smooth trajectories and obstacle avoidance in the working area are needed. This paper presents a closed form analysis on the joint limits, and elbow movement minimization using the redundancy of a seven DOF manipulator based on a closed form inverse kinematics. Using the redundancy circle of the redundant arm, the NULL space, the motion planning has been performed. The solution provided has the advantage of being exact with low computational cost. Consequently, this method eliminates the need for trial-and-error which takes times and may not result in a desirable solution. Experimental results have been provided showing the advantages of closed form inverse kinematics over the iterative methods.

1 Introduction

To be able to manipulate objects like humans, a robot arm needs to be kinematically redundant. Consequently, many robot manipulator manufacturers build arms with 7 DOF to imitate human's 7 DOF arm. This redundancy leads to infinite number of solutions, which is called the NULL space, for each given hand pose. The NULL space can be used to avoid joint limits [1], obstacles [2] and singular configurations [3], [8]. Moreover, a subset of the NULL space can optimize the robot arm dynamics [4].

To solve for the joint angles based on a given EEF (End EFector) pose, which is called the inverse kinematics, several techniques exist. Iterative approach using Jacobian matrix is one of these techniques. Basically for sufficiently small changes, the changes in the joints and in the position are linearly related by the Jacobian matrix (based on first order Taylor expansion):

$$\Delta x = J(\theta).\Delta\theta \text{ or } \Delta\theta = J^{-1}(\theta).\Delta x \ . \qquad (1)$$

Although this looks a simple equation but requires calculating the inverse of the Jacobian matrix which is not a simple task especially for redundant manipulators that need pseudo-inverse since the Jacobian matrix is not square. There are also other iterative methods which do not rely on the inverse of Jacobian but still need many iterations to determine the inverse kinematics solution and the given solution may not be optimal or even desirable.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 423 – 432, 2005.

Dahm and Joublin [5] proposed an elegant method to solve the inverse kinematics of a 7 DOF, which can be used for many DOF, in closed form. There are similar approaches that provide the closed form inverse kinematics for redundant arms, especially for the human's arm like manipulators [6], [9]. All these studies show that for any given EEF pose (position and orientation) the robot's elbow would be free to move along a circle which is called the Redundancy Circle. Thus a solution for the inverse kinematics would be a solution from the Redundancy Circle.

Closed form solution provide many advantages to other techniques such as: a) no approximation with Taylor expansion is needed, b) no Jacobian inversion needed so no singularity treatment is necessary as far as the pure mapping is concerned, c) the use of redundancy takes place in the Cartesian space rather than the joint space.

Using the closed form inverse kinematics provide us the Redundancy Circle, i.e. the NULL space, in which the robot is free for a given hand pose. Any solution is acceptable as long as there is no joint limit or other constraints. However the joints are limited and a solution selected from the Redundancy Circle may not be valid due to joint limits. Consequently finding a valid solution will be another trial-and-error to resolve the joints' limit problem. In this paper we address the joint limit and elbow movement minimization using the closed form inverse kinematics described in [5].

In the following section we review the closed form solution for a 7DOF manipulator. In section 2 the joints' limit analysis has been discussed. Section 3 addresses the elbow movement minimization. In the results we show some example of using the closed form method for a robot arm.

1.1 Closed form Inverse Kinematics Review

In order to better understand the notations and geometry involved in this analysis, we review the closed inverse kinematics for a 7 DOF manipulator [5], with some minor corrections. In this method, the location of the next joint is calculated w.r.t. the current joint and the joint angles for the current joint are calculated using the following formulas:

$$\begin{aligned} roll &= a\tan 2({}^{cj}nj^{y}, {}^{cj}nj^{x}) \\ pitch &= a\cos({}^{cj}nj^{z} / |nj|) \end{aligned} \quad . \tag{2}$$

In which ${}^{cj}nj^{x}, {}^{cj}nj^{y}, {}^{cj}nj^{z}$ are the position values of the next joint (nj) w.r.t. current joint (cj). In this paper we use: ${}^{d}a_{b}^{c}$ in which "*b*" shows the joint that "a" refers to, "*c*" shows the axis (x, y or z) and "*d*" show the frame in which "*a*" is represented. Figure 1 shows a 7 DOF manipulator that each joint consist of two rotary joints making a spherical joint. The 7th joint is along the EEF (End EFfector). For the sake of simplicity the world frame is considered at the shoulder joint. As it can be seen from this figure, if the EEF is fixed then the wrist's position, not its orientation, is fixed.

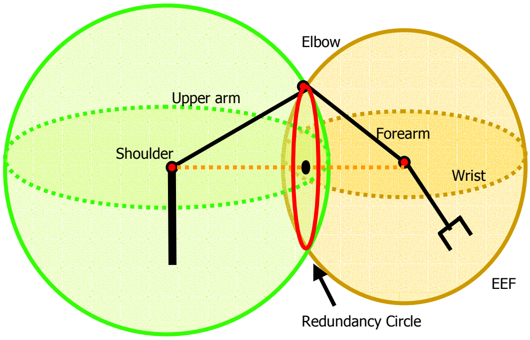


Fig. 1. The Redundancy Circle (RC) of a 7 DOF arm centered on the line connecting the shoulder and wrist

Consequently, the spherical motion of the upper arm and the forearm generate spheres that their intersection would be a circle on which the elbow is located. If there is no joint limit, then the elbow is free to move around this circle that is called the Redundancy Circle. If the upper arm and the form arm spheres do not intersect, then it mean that the desired EEF pose is out of the workspace.

It should be mentioned that pitch calculation using acos will return only one solution out of the 2 possible solutions. If the manipulator has the human arm restrictions, i.e. it does not fold back the elbow or any other joint, then one solution it is enough. However, if all solutions are needed, then the 2nd solution should be calculated based on the sign of joint angle.

Figure 2 shows how the kinematics of the arm can be used to determine the radius and the center of the Redundancy Circle. ${}^{Shoulder}\vec{r}_w$ has been calculated using the transformation matrices: ${}^{Shoulder}_{w}T = {}^{Shoulder}_{EEF}T \times {}^{EEF}_{w}T$. R, the radius of the RC, and the center of it can be calculated as follows:

$$R=\sqrt{r_U^2-\left(\frac{r_U^2-r_F^2+r_W^2}{2\left|r_W\right|}\right)^2} \quad . \tag{3}$$

$${}^{Shoulder}\vec{r}_M=\frac{r_U^2-r_F^2+r_W^2}{2r_W^2}\,{}^{Shoulder}\vec{r}_W \quad . \tag{4}$$

Equation (3) is valid as long as EEF stays inside the workspace. Otherwise, equation (3) becomes invalid since r_w would be bigger than $| r_U + r_F |$ or smaller than $| r_U - r_F |$. If equation (3) is valid, then the elbow position $\vec{r}_U$ is:

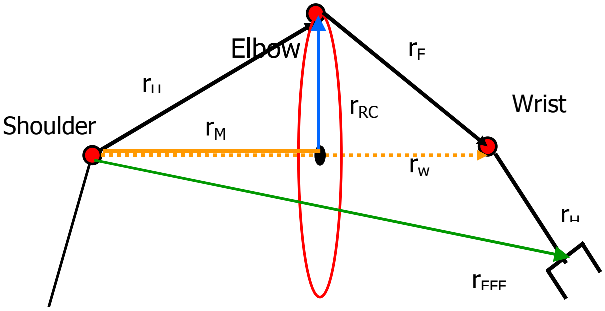


Fig. 2. The center and the radius of the Redundancy Circle and the elbow position can be calculated based on the EEF pose

$$\vec{r}_U = (R_Z^{\theta_W} . R_Y^{\varphi_W} . R_Z^{\alpha} . e^x).R + {}^{Shoulder}\vec{r}_M \quad . \tag{5}$$

In which α is the Redundancy Angle (RA), θ_W and φ_W are the spherical angles for ${}^{Shoulder}\vec{r}_w$. $R_Z^{\varphi_w}.R_Y^{\theta_w}.R_Z^{\alpha}$ determines the rotation to be applied to e^x, a unit vector along x axis, to align it with $\vec{r}_{RC}$. Having ${}^{Shoulder}\vec{r}_U$ then the first and second joint values can be calculated based on ${}^{Shoulder}\vec{r}_U$:

$$\theta_1 = a\tan 2({}^{Shoulder}r_U^x , {}^{Shoulder}r_U^y) \quad . \tag{6}$$

$$\theta_2 = a\cos({}^{Shoulder}r_U^z / |r_U|) \quad . \tag{7}$$

The same methodology is used to calculate the joint values for the elbow joints, joint 3 and 4, and the wrist joints, 5 and 6. The following equations show calculation of the wrist position w.r.t. elbow and the EEF position w.r.t. wrist

$${}^{Elbow}\vec{r}_F = (R_Y^{-\theta_2} . R_Z^{-\theta_1}) . {}^{Shoulder}\vec{r}_F \quad . \tag{8}$$

$$\theta_3 = a\tan 2({}^{Elbow}r_F^y , {}^{Elbow}r_F^x) \quad . \tag{9}$$

$$\theta_4 = a\cos({}^{Elbow}r_F^z / | r_F |) \quad . \tag{10}$$

$${}^{Wrist}\vec{r}_H = (R_Y^{-\theta_4} . R_Z^{-\theta_3} . R_Y^{-\theta_2} . R_Z^{-\theta_1}) . {}^{S}\vec{r}_H \quad . \tag{11}$$

$$\theta_5 = a\tan 2({}^{Wrist}r_H^y , {}^{Wrist}r_H^x) \quad . \tag{12}$$

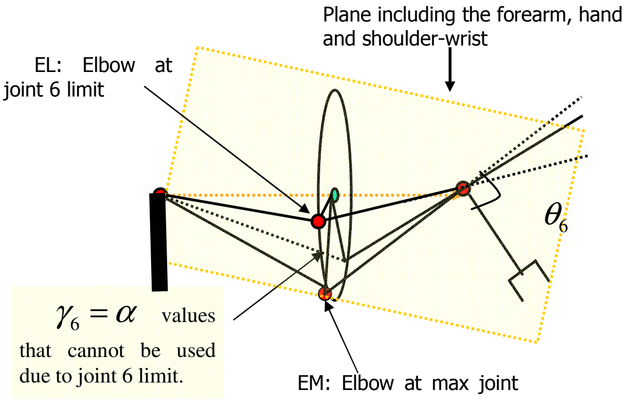

Plane including the forearm, hand and shoulder-wrist
EL: Elbow at joint 6 limit
θ_6
$\gamma_6 = \alpha$ values that cannot be used due to joint 6 limit.
EM: Elbow at max joint

Fig. 3. Joint 6 will reach the maximum when the hand, the elbow and the shoulder-wrist are on the same plane

$$\theta_6 = a\cos({}^{Wrist}r_H^z / |r_H|) . \tag{13}$$

θ_7 is calculated assuming that the Z axis of the EEF is along the hand. It is also possible to calculate it using the orientation of EEF's x axis (${}^SR_{EFF}^{11}, {}^SR_{EFF}^{21}, {}^SR_{EFF}^{31}$) w.r.t. the wrist.

$${}^{Wrist}R_{EFF} = (R_Y^{-\theta_6}.R_Z^{-\theta_5}..R_Y^{-\theta_4}.R_Z^{-\theta_3}.R_Y^{-\theta_2}.R_Z^{-\theta_1}).{}^SR_{EFF} . \tag{14}$$

$$\theta_7 = a\tan 2({}^{Wrist}R_{EFF}^{21}, {}^{Wrist}R_{EFF}^{11}) . \tag{15}$$

2 Joint Limits in the Redundancy Circle

The previous section showed how a joint angle can be calculated based on the position of the elbow on the Redundancy Circle, i.e. angle α . However, a desired α may not be feasible due to joint limits. In other words, the calculated joint angles are not valid because they pass the joint limits. To avoid joint limits two methods can be used: a) trial-and-error: try another α and calculated joint values until a valid joint value is calculated, b) map the joint limits to the redundancy circle and select α from the valid region of the Redundancy Circle. There is no question that the 2nd method is preferable over the first one since it eliminates many unsuccessful trials and provides a mean to calculate the joint angles directly and determine an optimal solution. Moreover, for real-time applications such as service robots, the trial-and-error method may not be feasible due to the computation cost.

To map joints' limits to RC, we start with joint 4 (pitch angle for the elbow joint) since it is constant in a given NULL space. In other words if joint 4 limit is violated in a given configuration, then there is no solution at all for the desired configuration. Equation (16) gives a direct way to calculated joint 4's value w/o the need to go through equations (3) to (10).

$$\theta_4 = \pi - a\cos(\frac{r_U^2 + r_F^2 - r_W^2}{2|r_W|}) \ . \tag{16}$$

If joint 4's limit is not violated, then the other joints' limits should be checked. We start with joint 6. In a given hand configuration, joint 6's maximum value would be reached when r_w, r_F and r_H are all on the same plane. Consequently by projecting r_H on the Redundancy Circle, we can determine the region in which α is not valid. First, the maximum value of joint 6 in the current configuration is calculated. In such a case r_F , elbow to wrist link, can be represented by a vector from EM to the wrist (Figure 3). Let's assume a coordinate frame, called the RC coordinate frame, at the center of Redundancy Circle with Z axis pointing toward the wrist.

$$\begin{aligned}
&{}^{RC}\vec{r}_F{}^{x} = R.{}^{RC}\vec{r}_H{}^{x} / |{}^{RC}r_H{}^{xy}| \\
&{}^{RC}\vec{r}_F{}^{y} = R.{}^{RC}\vec{r}_H{}^{y} / |{}^{RC}r_H{}^{xy}| \\
&{}^{RC}\vec{r}_F{}^{z} = (r_W - r_M).{}^{RC}\vec{r}_W \\
&|{}^{RC}r_H{}^{xy}| = sqrt({}^{RC}\vec{r}_H{}^{x}.{}^{RC}\vec{r}_H{}^{x} + {}^{RC}\vec{r}_H{}^{y}.{}^{RC}\vec{r}_H{}^{y})
\end{aligned} \ . \tag{17}$$

Then the max joint 6 angle is calculated from the dot product of ${}^{RC}\vec{r}_F$ and ${}^{RC}\vec{r}_H$:

$$\theta_6{}^{\max} = a\cos({}^{RC}\vec{r}_F.{}^{RC}\vec{r}_H) \ . \tag{18}$$

$$EM = a\tan 2({}^{RC}\vec{r}_H{}^{y}, -{}^{RC}\vec{r}_H{}^{x}) \ . \tag{19}$$

To calculate γ_6, joint 6's limit on the RC, the dot product of r_F projection on RC with r_H projection on RC is used:

$$\gamma_6 = a\cos(({}^{RC}\vec{r}_F{}^{x}.{}^{RC}\vec{r}_H{}^{x} + {}^{RC}\vec{r}_F{}^{y}.{}^{RC}\vec{r}_H{}^{y})/(R.|{}^{RC}\vec{r}_H{}^{xy}|)) \ . \tag{20}$$

in which ${}^{RC}\vec{r}_F{}^{x}.{}^{RC}\vec{r}_H{}^{x} + {}^{RC}\vec{r}_F{}^{y}.{}^{RC}\vec{r}_H{}^{y}$ is equal to:

$$\cos(\theta_{6\_\lim}).(|\vec{r}_F|.|\vec{r}_H|)) - (r_W - r_M).{}^{RC}\vec{r}_H{}^{z}) \ . \tag{21}$$

Because of the similarity between joints 2 and 6, the same method can be used to map joint 2 limits to the Redundancy Circle. In such a case, joint 2's maximum for the current configuration may be reached when the elbow, the wrist and the shoulder link are on the same plane.

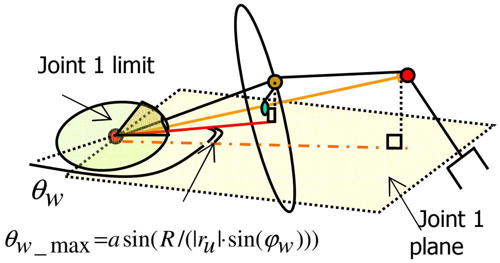


Fig. 4. Depending on θ_w, elbow may be in joint 1's limit region

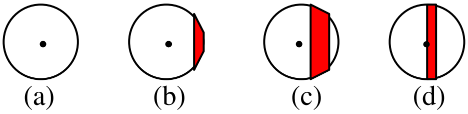


Fig. 5. Mapping joint 1 limit's on the Redundancy Circle. a) No limit, b, c and d show when joint 1 limit is violated. c and d show situation in which elbow can be on the right side of the Redundancy Circle but cannot go to the left side.

Joints 1, 3 and 5 are all roll angles and their limits on the RC can be calculated similarly. Thus we only present joint 1's limit on the RC. Figure 4 shows the span of elbow based on the RC. There would be 4 cases. Figure 5 shows all possible mapping of joint 1's limit violation on the RC. The maximum wrist angle would be calculated based on equation:

$$\theta_{w\_\max} = a\sin(R/(|r_u| \cdot \sin(\varphi_w))) \ . \tag{22}$$

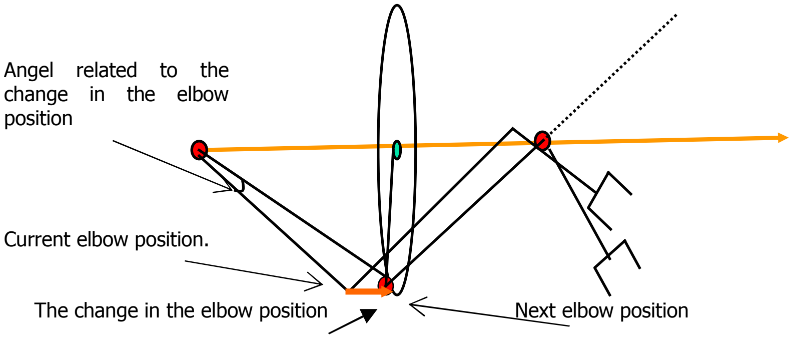

Angel related to the change in the elbow position
Current elbow position.
The change in the elbow position
Next elbow position

Fig. 6. Redundancy angle change based on the minimum elbow movement

For case (b), the limit on the RC can be calculated using the following equations:

$$
\begin{aligned}
&\gamma = a\cos(h / R)\\
&h{=}\tan(J1\_\mathrm{lim}{-}\theta_W).|r_M|.\sin(\varphi_W)\quad .\\
&\bar{\alpha} = \frac{PI}{2} \pm \gamma
\end{aligned}
\tag{23}
$$

in which $\overline{\alpha}$ represents the prohibited Redundancy Angle. As it was mentioned earlier the same approach is used to calculate the joint limit mapping for joints 3 and 5.

3 Minimum Elbow Movement

One of the heuristics in motion planning of a 7DOF arm, to select a good plan from a set of plans, is to minimize the elbow movement. This leads to less arm movement that may be desirable for service robots. To calculate the joints values based on this heuristic, the Redundancy Angle can be selected such that it includes this criterion. Figure 6 shows the elbow movement from one Redundancy Circle to another one.

$$
\begin{aligned}
&\cos(\gamma) = \vec{r}_{U_c}.\vec{r}_{U_n} = r_{U^{x}_{c}}.r_{U^{x}_{n}} + r_{U^{y}_{c}}.r_{U^{y}_{n}} + r_{U^{z}_{c}}.r_{U^{z}_{n}}\\
&\frac{d\cos(\gamma)}{d\alpha}{=}[\sin(\theta_W)\cos(\alpha){+}\sin(\alpha)\cos(\theta_W)\cos(\varphi_W)].R.r_{U^{x}_{c}}-\\
&\qquad[\sin(\theta_W)\sin(\alpha)\cos(\varphi_W){-}\cos(\alpha)\cos(\theta_W)].R.r_{U^{y}_{c}}-\\
&\qquad\sin(\alpha)\sin(\varphi_W).R.r_{U^{z}_{c}}{=}0\\
&\cos(\alpha)[\sin(\theta_W).r_{U^{x}_{c}}-\cos(\theta_W).r_{U^{y}_{c}}]{=}\qquad\qquad .\\
&\sin(\alpha)[-\cos(\theta_W)\cos(\varphi_W).r_{U^{x}_{c}}-\sin(\theta_W)\cos(\varphi_W).\cos(\theta_W).r_{U^{y}_{c}}+\sin(\varphi_W).r_{U^{z}_{c}}]\\
&a{=}\sin(\theta_W).r_{U^{x}_{c}}-\cos(\theta_W).r_{U^{y}_{c}}\\
&b{=}-\cos(\theta_W)\cos(\varphi_W).r_{U^{x}_{c}}-\sin(\theta_W)\cos(\varphi_W).\cos(\theta_W).r_{U^{y}_{c}}+\sin(\varphi_W).r_{U^{z}_{c}}\\
&\alpha{=}a\tan 2(a,b)
\end{aligned}
\tag{24}
$$

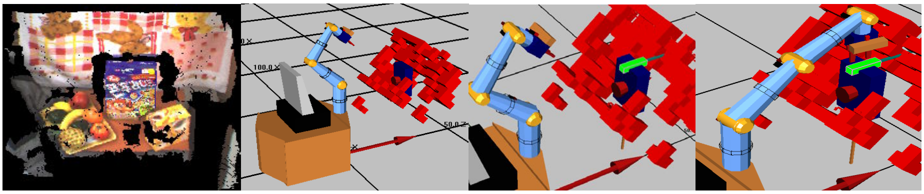

(a) (b) (c) (d)

Fig. 7. a) 3D point cloud, b) simulated scene showing object in dark blue and octree cells in red, c) accessibility result showing the point and direction to be grasped, d) the motion planning used the accessibility results to move the robot to the grasp point

To minimize the elbow movement, the angle ψ between the current elbow and the next elbow positions should be minimized. So the derivate of this angle with respect to α is calculated and set to zero. $\vec{r}_U$ in equation 24 is replace by its equivalent from equation 5.

4 Experimental Results

We have implemented and tested the proposed algorithm on Amtec's lightweight arm which is a 7 DOF robotic manipulator. We have also used RRG from University of Texas at Austin [7] and compared the two. Although RRG is a general inverse kinematics package and can handle general form of manipulators based on the Denavit-Hartenberg values, however, we found the closed form inverse kinematics more suitable for service robots in which real time performance is needed. RRG provides the inverse solution using iterative approach which may not result into a good solution. In such a case a desired solution should be determined using trial-and-error method.

We tested our approach on a service robot (Figure 7) in which the world model is provided through a 3D Modeling engine [11]. The 3D Modeling engine determines the location of the desired objects, that are already stored in a database, and registers the objects in the 3D model. Then the poses of the planes around the scene are determined and the planes are registered into the model. Finally any remaining object in the environment is represented using Octree representation.

Then an accessibility analysis [11] is used to determine the best direction for grasping which is used by the motion planning to determine the path using the above method.

5 Conclusion

This paper presents a closed form analysis on the redundancy of a seven DOF manipulator using closed form inverse kinematics. Using the Redundancy Circle of the redundant arm, the NULL space of the robot arm is determined and used for motion planning. The solution provided has the advantage of being exact with low computational cost. The future work will be focused on more direct mapping for joints 1, 3 and 5 of the manipulator. Moreover, the singularity may be mapped to the Redundancy Circle too.

Acknowledgment

This study was performed for the intelligent robotics development program, one of the 21st century frontier R&D programs funded by the Ministry of commerce, industry and energy of Korea. This work is partly supported by the science and technology program of Gyeonggi province.

References

1. Ligeois: Automatic Supervisory Control of the Configuration and Behavior of Multi-Body Mechanism. IEEE Transaction on Systems, Man and Cybernetics. (1977)
2. Maciejewski, A.A., Klein, C.A.: Obstacle Avoidance for Kinematically Redundant Manipulators in Dynamically Varying Environments. The International Journal of Robotics Research. 4(3) (1985) 109 - 117
3. Senft, V., Hirzinger, G.: Redundant Motions of Non Redundant Robots- A Nnew Approach to Singularity Treatment. IEEE International Conference on Robotics and Automation. (1995)
4. Pholsiri, Kapoor, C., Tesar, C., D.: Manipulator Task-Based Performance Optimization. Proceedings of ASME 2004 Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Salt Lake City. UT. (Sept. 28 - Oct. 2. 2004)
5. Dahm, P., Joublin, F.: Closed Form Solution for the Inverse Kinematics of a Redundant Robot Arm. Technical Reports. Institute For Neuroinformatic. Lehrstuhl Fur Theoretische Biologie. Ruhr-University. Bochum. Germany
6. Asfour, T., Dillmann, R.: Human-Like Motion of a Humanoid Robot Arm Based on Closed-Form Solution of the Inverse Kinematics Problem. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003). Las Vegas. USA. (Oct. 27-31. 2003)
7. Pryor, Taylor, M., Kapoor, R.C., Tesar, C., D.: Generalized Software Components for Reconfiguring Hyper-Redundant Manipulators. IEEE/ASME Transactions on Mechatronics. Vol. 7. No. 4. (Dec. 2002) 475 - 478
8. Yigit, S., Burghart, C., Woern, H.: Avoiding Singularities of Inverse Kinematics for a Redundant Robot Arm for Safe Human Robot Co-operation. CCCT. (2003)
9. Asfour, T., Berns, K., Dillmann, R.: The Humanoid Robot ARMAR: Design and Control. International Conference on Humanoid Robots (Humanoids 2000). Boston (MIT). USA. (Sept. 7-8. 2000)
10. Brock, O., Khatib, O., Viji, S.: Task-Consistent Obstacle Avoidance and Motion Behavior for Mobile Manipulation. Submitted to IEEE International Conference on Robotics and Automation. (2002)
11. Jang, H., Han, J., Lee, S., Jang, D., Kim, E., Moradi, H.: A Visibility-based Accessibility Analysis of the Grasp Points for Real-time Manipulation. The International Conference on Intelligent Robots and Systems. (2005)

Detecting Anomalous Network Traffic with Combined Fuzzy-Based Approaches

Hai-Tao He[1], Xiao-Nan Luo[1], and Bao-Lu Liu[2]

[1] Computer Application Institute,
Sun Yat-sen University, Guangzhou, Guangdong, 510275, China
{hthe, lnslxn}@sysu.edu.cn
[2] College of Textile & Garment,
Guangzhou University, Guangzhou, Guangdong, 510310, China
liubaolu@gzfx.edu.cn

Abstract. This paper introduces the combined fuzzy-based approaches to detect the anomalous network traffic such as DoS/DDoS or probing attacks, which include Adaptive Neuro-Fuzzy Inference System (*ANFIS*) and Fuzzy C-Means (*FCM*) clustering. The basic idea of the algorithm is: at first using *ANFIS* the original multi-dimensional (*M-D*) feature space of network connections is transformed to a compact one-dimensional (*1-D*) feature space, secondly *FCM* clustering is used to classify the *1-D* feature space into the anomalous and the normal. *PCA* is also used for dimensional reduction of the original feature space during feature extraction. This algorithm combines the advantages of high accuracy in supervised learning technique and high speed in unsupervised learning technique. A publicly available DRAPA/KDD99 dataset is used to demonstrate the approaches and the results show their accuracy in detecting anomalies of the network connections.

1 Introduction

With the explosive development in the last decade, Internet has been one of the most important communication infrastructures and millions of people depend heavily on it. Unfortunately, the global Internet infrastructure is under great risks caused by network intrusions now. For example, in January of 2003 the SQL Slammer worm infected more than 90% of vulnerable hosts within 10 minutes [1], and the probing traffic overwhelmed the whole Internet during it's target selection; in April of 2004, a new worm named Sasser circulated on the Internet and infected more than a million personal computers worldwide within a few days, which caused them to repeatedly shut down and reboot [2]. Hence, *intrusion detection* is of big importance in the field of Internet security research. Generally network intrusions fall into four categories of attacks: *DoS, Probing, R2L, U2R* [3], the first two are at network level while the last two are at host level. In this paper only the *DoS* and *Probing* attacks are concerned, a simple cause is that the harm done by them outweighed the host level ones greatly.

Pattern Recognition (PR) is the scientific discipline whose goal is the classification of objects into a number of categories or classes, which is an integral part

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 433–442, 2005.

in most machine intelligence systems built for decision making [4]. And *intrusion detection* is an important area of *PR*, which is a two-class or multi-class *PR* problem. In this paper, we simply regard it as a two-class one, for our focus is on how to discriminate the anomalous network traffic from the normal. However, one assumption should be made to this two-class *PR* problem that the feature vectors of the anomalies are distinct from the those of the normals. The theory and techniques of *PR* have been widely applied to *intrusion detection* and many literatures appeared in recent years, such as Support Vector Machine (*SVM*) [5], Fuzzy Logic [6], Self-Organizing Map (*SOM*) [7], Hidden Markov Model (*HMM*) [8], Bayes decision theory [10], clustering [9], etc.. Traditionally, *PR* is divided into two types: *supervised PR* and *unsupervised PR*. If given a trained data with known labels or priori known information, this type of problem is called as *supervised PR*; and if given the feature sets only (without the labels), the goal is to reveal the underlying clusters (groups) based on similarities, this type of problem is known as *unsupervised PR*. To our knowledge, either the *supervised* or the *unsupervised* (not both) was employed by most of the intrusion detection researchers. However, here our focus is on combining these two types. The basic idea of the algorithm is: at first using Adaptive Neuro-Fuzzy Inference System (*ANFIS*), which is the *supervised*, the original multi-dimensional (*M-D*) feature space of network traffic is mapped to a compact one-dimensional (*1-D*) feature space having two distinct centroids, secondly Fuzzy C-Means (*FCM*) clustering, which is the *unsupervised*, is used to classify the *1-D* feature space into the anomalous and the normal. Additionally, during the preprocessing Principal Component Analysis (*PCA*), which is a powerful data analysis tool and often ignored by some researchers, is used to reduce the dimensions of the original feature space.

The rest of the paper is organized as follows. In Section 2 the related works such as Intrusion Detection, Principal Component Analysis (*PCA*), Adaptive Neuro-Fuzzy Inference System (*ANFIS*) and Fuzzy C-Means (*FCM*) clustering are introduced. Section 3 is the overview of Network Intrusion Detection System (*NIDS*) using the combined fuzzy-based approaches. In section 4 an experiment is carried out to demonstrate the approaches with the publicly available DRAPA/KDD99 dataset. Section 5 summarizes our conclusions and introduces the future work.

2 Related Works

2.1 Intrusion Detection

Network anomalies typically refer to circumstances when network operations deviate from normal network behavior, which can be broadly classified into two categories: the first category is related to network failures and performance problems; and the second is related to security-related problems [11]. *Network intrusion*, which is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource [12], falls into the second major category. Intrusion detection system (*IDS*) detects automatically computer intrusions to protect computers and computer networks safely from malicious uses or attacks [13]. According to the aspect of the target environment for detection, *Host-based* IDS (*HIDS*) running on a individual computer detects possible attacks into the computer, *Network-based* IDS (*NIDS*) monitors network behavior by examining the content as well the format of network data packets [14]. Another categorization generally used is *misuse detection* and *anomaly detection*: *misuse detection* can detect the known attacks defined in the knowledge base accurately but is deficient in finding the novel intrusions outside the knowledge base; while *anomaly detection* can detect the novel attacks but with high error detection rates.

The techniques generally used in *anomaly detection* are statistical models, immunity system approach, protocol verification, file checking, training checking, neural nets and whitelisting, while the techniques generally used in *misuse detection* are expression matching, state transition analysis, dedicated languages, genetic algorithms and burglar alarms, the further information of these techniques can be found in [15]. For the time now the commercial *IDS*s mainly are *misuse-detection-based IDSs* and the *anomaly-detection-based IDSs* are left to the researchers in labs. In this paper, we discuss the *anomaly-detection-based NIDS* only.

2.2 Principal Component Analysis (PCA)

In the real world, we have to understand some phenomenon of the system by measuring it's various quantities (e.g. spectra, voltages, velocities.). The dataset containing the system's various quantities is called *feature space*, which is always high-dimensional, unclear and even with noise. *PCA* is such a way to reveal the hidden dynamics and best expresses the feature space. Suppose the feature space is m-dimensional and each feature vector is denoted by a *row vector* $\boldsymbol{x}$ (x_1 x_2 $\ldots x_m$), by *linear transformation* the new coordinate system formed by m orthogonal *Principal Components* (PCs, which are basis vectors) is setup, then the original row vector $\boldsymbol{x}$ can be projected to the new axes and the new feature vector $\boldsymbol{y}$ (y_1 y_2 $\ldots y_m$) is the result. With an assumption that greater variance means greater importance when working with zero-mean feature space in every dimension, the *PCs* (from 1 to m) are ordered by the amount of data variances that they capture, PC_i ($1 \leqq i \leqq m$) represents the direction that captures the maximum variance among the remaining $m - i + 1$ orthogonal directions. But, after transformation it is nearly impossible to interpret the physical meanings of the new feature vectors. *PCA* is a very important method of multivariate data analysis and has been widely applied to image compression, face recognition etc., in this paper *PCA* is used to reduce the dimensions of the feature space.

2.3 Adaptive Neuro-fuzzy Inference System (ANFIS)

Adaptive Neuro-Fuzzy Inference System (*ANFIS*, also known as Adaptive Network-based Fuzzy Inference System) combines the neural network with the fuzzy system, which implements the fuzzy inference system under the framework of neural network and makes *ANFIS* have the self-learning ability in identifying the system's parameters. During the learning process either *back propagation* or in a combination with a *least squares method* is used. The *ANFIS* used here is a typical zero order Sugeno (or Takagi-Sugeno-Kang) fuzzy model with properties of single output, no rule sharing and unit-weight of each fuzzy rule, which has good performance of nonlinear function approximation. Unlike the *objective* probability-based inference system, *ANFIS* seems *subjective* and could employ experts' domain knowledge much more. The further discussion about *ANFIS* can be found in [17].

2.4 Fuzzy C-Means (FCM) Clustering

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [16]. Generally clustering can be categorized into two classes: *hierarchical clustering* and *partitional clustering.* Compared with the partitional algorithms the hierarchical ones are more versatile for their working well on non-isotropic clusters such as chain-like and concentric ones besides the compact isotropic clusters, whereas the partitional algorithms (such as the typical k-means algorithm) can only deal well with the data set having compact isotropic clusters. However, the time and space complexities of the partitional algorithms are typically lower than those of the hierarchical algorithms [16], so they are more efficient.

FCM clustering is a type of soft clustering, which gives every feature vector a degree of membership in several distinct clusters, while a hard clustering algorithm (such as *k-means* clustering) assigns a feature vector to only one cluster. In this paper, we use *FCM* clustering and convert it to a hard clustering by choosing the cluster with the largest membership.

3 Overview of the Combined Fuzzy-Based Network Intrusion Detection System (CFNIDS)

The overall structure and components of the Combined Fuzzy-based Network Intrusion Detection System (*CFNIDS*) are depicted in Fig.1. Here we just describe it briefly. First, training and test dataset are collected, which are the customized subsets of KDD99 dataset for our specific purposes. Secondly, these sets are preprocessed. In this progress, our domain knowledge of computer security is used to select several comparative important features of a network connection, and further, *PCA* is used to reduce the dimensions of the selected feature space in order to enhance the efficiency of the classifier. Thirdly, using *ANFIS*, which is the *first* Fuzzy-based approach, a transformation model mapping the feature space from *M-D* to *1-D* is learned. Fourthly, using *FCM* clustering, which is the *second* Fuzzy-based approach, the transformed *1-D* feature space is divided into the anomalous and the normal efficiently.

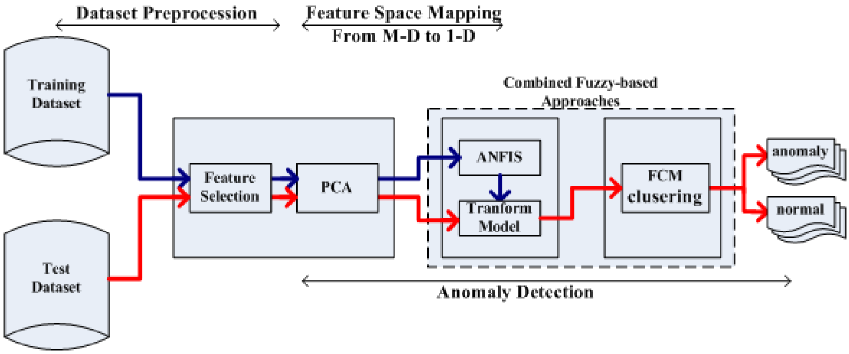


Fig. 1. Structure of combined fuzzy-based NIDS

What are the situations that the *CFNIDS* is suitable for? What problems can it solve and what limitations does it have? In order to answer such questions we should introduce the design philosophy of *CFNIDS*. As mentioned above, *CFNIDS* is a black box serving as a two-class pattern classifier, whose input is the feature vector extracted from the observed network connection in the real world, and the output is 0 or 1. Here 0 means *normal* and 1 means *anomaly*. Thus, *CFNIDS* is such an intrusion detection system using *anomaly detection* techniques only. *CFNIDS* could deal with network level intrusions with high efficiency, especially the flooding attacks (DDoS, Probing) which would deplete the limited resources of Internet. And three very important assumptions are accompanied with *CFNDIS*, which are as following:

1. The feature space of the anomalous is distinct from that of the normal, which was extracted directly from the observed network connections, and even after the transformation from *M-D* to *1-D* the distinction remains still.
2. The training dataset is large enough to model the normal network behavior appropriately.
3. For any input feature vector in the test dataset, it statistically belongs to the same distribution of that in the training dataset and has the same *mean* and *variance* values, which is very important for the normalization of the input of the learned *ANFIS* transformation model when testing.

The results will show that these three assumptions are useful and reasonable in practice.

4 Experiment and Results

We now describe the processes of the experiment as three phases. In the first phase, we introduce the publicly available KDD99 dataset, which is popular among the intrusion detection researchers, and how the dataset preprocessed according to our specific purposes and assumptions. In the second phase, *CFNIDS* is under learning and validating, several results are displayed. In the third phase, we evaluate the detection rates (*DRs*) of the trained *CFNIDS* using the test dataset especially the novel attacks in it.

4.1 Dataset and Preprocessing

Dataset Description The datasets used in this paper is from the 1999 KDD intrusion detection competition [18]. The task of the competition was to build a network intrusion detector, a predictive model capable of distinguishing between *bad* connections, called intrusions or attacks, and *good* normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, generally the attacks fall into four main categories [3]:

- DOS: denial-of-service, e.g. ping-of-death, syn flood;
- Probing: surveillance and other probing, e.g. port scanning;
- R2L: unauthorized access from a remote machine, e.g. guessing password;
- U2R: unauthorized access to local root privileges, e.g. various buffer overflow attacks.

There are several data files in the database, in which we use two files: kddcup.data_10_percent.gz, which is the training and validation dataset having exactly 494, 021 instances with 22 attack types; and corrected.gz, which is the test dataset having exactly 311, 029 instances with 37 attack types. Table 1 and Table 2 list the attack types, instances, categories in each dataset respectively. It *should* be noted that in prior literatures using KDD99 dataset such as [5] the number of novel attack types in the test dataset is 14, but, according on our experiment, we find it should be 17. The novel attacks in the test dataset are *apache2, httptunnel, mailbomb, mscan, named, processtable, ps, saint, sendmail, snmpgetattack, snmpguess, sqlattack, udpstorm, worm, xlock, xsnoop, xterm*, while *spy* and *warezclient* do *not* appear in the test dataset and just in the training dataset.

Table 1. Attack distributions in the training dataset (with 97, 278 normal instances)

Attack Type	Instances	Category	Attack Type	Instances	Category
back	2, 203	dos	buffer_overflow	30	u2r
ftp_write	8	r2l	guess_passwd	53	r2l
imap	12	r2l	ipsweep	1, 247	probing
land	21	dos	loadmodule	9	u2r
multihop	7	r2l	neptune	107, 201	dos
nmap	231	probing	perl	3	u2r
phf	4	r2l	pod	264	dos
portsweep	1, 040	probing	rootkit	10	u2r
satan	1, 589	probing	smurf	280, 790	dos
spy	2	r2l	teardrop	979	dos
warezclient	1, 020	r2l	warezmaster	20	r2l

Table 2. Attack distributions in the test dataset (with 60, 593 normal instances)

Attack Type	Instances	Category	Attack Type	Instances	Category
back	1, 098	dos	buffer_overflow	22	u2r
ftp_write	3	r2l	guess_passwd	4367	r2l
imap	1	r2l	ipsweep	306	probing
land	9	dos	loadmodule	2	u2r
multihop	18	r2l	neptune	58, 001	dos
nmap	84	probing	perl	2	u2r
phf	2	r2l	pod	87	dos
portsweep	354	probing	rootkit	13	u2r
satan	1, 633	probing	smurf	164, 091	dos
teardrop	12	dos	warezmaster	1, 602	r2l
apache2	794	dos	httptunnel	158	u2r
mailbomb	5000	dos	mscan	1053	probing
named	17	r2l	processtable	759	dos
ps	16	u2r	saint	736	probing
sendmail	17	r2l	snmpgetattack	7741	r2l
snmpguess	2406	r2l	sqlattack	2	u2r
udpstorm	2	dos	worm	2	r2l
xlock	9	r2l	xsnoop	4	r2l
xterm	13	u2r			

There are 41 extracted features in each network connection, in which 38 features are *continuous* and the others are *symbolic*. And Stolfo et al. defined higher-level features that help in distinguishing normal connections from attacks [19]. There are four categories of derived features, which are 9 *intrinsic* features, 13 *content* features, 9 *traffic* features and 10 *host* features.

Preprocessing. The focus of this paper is on the network level intrusions, according to it the preprocessing can be divided into three steps:

- First, pruning of the training dataset. We filter out the feature vectors of *U2R* and *R2L* categories from the training dataset, meanwhile what labelled as *satan (probing)* and *neptune (dos)* are removed also in order to inspect the *DRs* of them (as the novel attacks) in the test dataset.
- Secondly, feature selection using our domain knowledge. Among 41 features only 8 of them are selected here, which are *src_bytes, dst_bytes, count, srv_count, dst_host_count, dst_host_srv_count, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate.*
- Thirdly, feature extraction using *PCA*. After the above 2 steps, the feature matrix for training $\boldsymbol{X}$ is gotten. Then *PCA* is applied to $\boldsymbol{X}$ and 8 *PCs* is the result. Table 3 displays their variances and proportions. In order to reduce the size of the training dataset, only 5 *PCs* are selected, which are from $PC1$ to $PC5$ and account for more than 90% of the total variances.

Table 3. Variances and proportions of the 8 principal components

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$	$PC7$	$PC8$
Variance	3.7657	1.0285	0.9955	0.9685	0.8082	0.3706	0.0626	0.0004
Proportion	47.07%	12.86%	12.44%	12.11%	10.10%	4.63%	0.78%	0.01%

4.2 Training and Validating

The main goal of this process is to train a *ANFIS*-based transformation model, by which the *M-D* feature space can be transformed to the *1-D* feature space with keeping the distinction between the anomalies and normals, also the validation dataset is used to tune the model's parameters. It is noted that the validation dataset is the same as the training dataset. In this experimentation, all the algorithms are implemented by Matlab 7.0.1 running on a IBM Thinkpad T.40 notebook with Intel Pentium 1.5GMhz CPU and 512M memory.

Table 4. Results of training and validating

Features	$PC1-2$	$PC1-3$	$PC1-4$	$PC1-5$	$PC1-6$
Training Time(s)	2522	2776	4148	12092	61335
Detection Rate	98.04%	98.85%	99.05%	99.13%	99.25%
False Alarm Rate	0.37%	0.21%	0.15%	0.14%	0.20%

From Table 4, it is inferred that: *PCA* is a powerful tool used to reduce the dimensions of feature matrix, even using only two PCs ($PC1$ and $PC2$) the Detection Rate (*DR*) and False alarm Rate (*FR*) of the *CFNIDS* is good; PCs from $PC1$ to $PC5$ are chosen as the input of the transformation model for the balance between the *performance* (*DR, FR*) and the *cost* (Training Time). If we use 6 PCs from $PC1$ to $PC6$, the *DR* is the highest, however, the *FR* is the middle and the training time is especially intolerable when in practice.

4.3 Testing

The test dataset consists of three classes of network connection instances: the normals (60593), the anomalies appeared in the training dataset (166041) and the novel anomalies (84395), Table 5 summarizes the overall results of testing, while the *DR* of each novel anomaly is displayed in Table 6. *DRs* of the anomalies appeared in the training dataset and the normals are comparatively high, but those of the novel anomalies are low because their feature vectors aren't distinct from the those of the normals. With regard to the novel anomalies, the *DRs* of *DOS* and *Probing* are rather higher than them about *R2L* and *U2R*, it is because here our approaches mainly deal with the network level intrusions and *content* features about host level intrusions were not included during feature selection.

Table 5. Results of testing

	Anomalies Trained	Anomalies Novel	Normals
Detection Rates	99.81%	73.01%	96.94%

Table 6. Results of the novel anomalies in the test dataset

Attack Type	Instances	Detection Rate	Attack Type	Instances	Detection Rate
buffer_overflow	22	22.73%	perl	2	50.00%
ftp_write	3	66.67%	guess_passwd	4367	1.60%
imap	1	0.00%	loadmodule	2	50.00%
multihop	18	33.33%	neptune	58,001	99.70%
phf	2	0.00%	rootkit	13	30.77%
satan	1,633	79.42%	warezmaster	1,602	50.06%
apache2	794	60.45%	httptunnel	158	67.72%
mailbomb	5000	0.00%	mscan	1053	21.94%
named	17	23.53%	processtable	759	6.85%
ps	16	12.50%	saint	736	96.88%
sendmail	17	17.65%	snmpgetattack	7741	0.00%
snmpguess	2406	0.08%	sqlattack	2	0.00%
udpstorm	2	0.00%	worm	2	0.00%
xlock	9	44.44%	xsnoop	4	0.00%
xterm	13	15.38%	overall	84395	73.01%

This paper focuses on the network level intrusions and the problem here is a two-class *PR* problem, which is not like the multi-class problems of the some prior literatures, so, it is unappropriate to compare our results with theirs exactly. However, more correct verification, comparable method to our approach is needed, which is the future work.

5 Conclusions and Future Work

In this paper, we propose the combined fuzzy-based approaches to detect the anomalous network traffic, and demonstrate them through an experiment. The results show the approaches can be an effective choice of implementing *NIDS*. And *PCA*, as a mainstay of the modern multivariate data analysis, plays the important role in the implementation of *CFNIDS*. Besides these, we point out and redress the errors regarding the attack types and instances of the DARPA/KDD99 dataset in some prior literatures. However, the dataset used here seems too old to reflect the current conditions of Internet, our approaches should be expanded to the live Internet traffic collected using *Passive and Active Measurement* (*PAM*) methods, such as the campus networks or the ISP core links. Also, *kernel* methods such as *kernel PCA* can be employed as the replacement of *PCA* in the future work.

Acknowledgement

This work was partly supported by the National Natural Science Foundation of China under Grant No. 90304011. The authors would like to thank vice Prof. Jun Zhang and Prof. Shun-Zheng Yu for their helpful discussions and comments.

References

1. Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., Weaver, N.: Inside the Slammer Worm. IEEE Security and Privacy Magazine. **1(4)** (July 2003) 33–39
2. Chen, Thomas M., Jean-Marc Robert: Worm Epidemics in High-Speed Networks. IEEE Computer. (June 2004) 48–53
3. Lee, W., Stolfo, S., Mok, K.: A Data Mining Framework for Buiding Intrusion Detection Models. Proc. of the 1999 IEEE Symposium on Security and Privacy. Oakland. CA. (May 1999)
4. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Second Edtion. Elsevier Science. (2003)
5. Dong Seong Kim, Jong Sou Park: Network-Based Intrusion Detection with Support Vector Machines. ICOIN 2003. LNCS 2662. (2003) 747–756
6. Alexander Hofmann, Carsten Schmitz, Bernhard Sick: Intrusion Detection in Computer Networks with Neural and Fuzzy Classifiers. ICANN/ICONIP 2003. LNCS 2714. (2003) 316–324
7. Manikantan Ramadas, Shawn Ostermann, Brett Tjaden: Detecting Anomalous Network Traffic with Self–organizing Maps. RAID 2003. LNCS 2820. (2003) 36–54
8. Sung–Bae Cho, Sang–Jun Han: Two Sophisticated Techniques to Improve HMM–Based Intrusion Detection Systems. RAID 2003. LNCS 2820. (2003) 207–219
9. Sang Hyun Oh, Won Suk Lee: Optimized Clustering for Anomaly Intrusion Detection. PAKDD 2003. LNAI 2637. (2003) 576–581
10. Scott, Steven L.: A Bayesian Paradigm for Designing Intrusion Detection Systems. Computational Statistics & Data Analysis. **45** (2004) 69–83
11. Marina Thottan, Chuanyi Ji: Anomaly Detection in IP Networks. IEEE Tran. on Signal Processing. **51(8)** (Aug. 2003) 2191–2204
12. Anderson, J.P.: Computer Secuirty Threat Monitoring and Surveillance. Technical Report. Fort Washington. Pennsyslvania. (Apr. 1980)
13. Denning, D.E.: An Intrusion Detection Model. IEEE Trans. on Software Engineering. **13(2)** (1987) 222–232
14. Dit-Yan Yeung, Yuxin Ding: Host-Based Intrusion Detection Using Dynamic and Static Behavioral Models. Pattern Recognition. **36** (2003) 229–243
15. Theuns Verwoerd, Ray Hunt: Intrusion Detection Techniques and Approaches. Computer Commnications. **25** (2002) 1356–1365
16. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys. **31(3)** (Sep.1999) 264–323
17. Jang, J.S.R.: ANFIS: Adaptive–Network-Based Fuzzy Inferrence System. IEEE Trans. on Systems, Man And Cybernetics. **23(3)** (1993) 665–685
18. KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
19. Salvatore J.Stolfo, Wei Fan, Lee, W., Andreas Prodromidis, Philip K. Chan: Cost–Based Modeling and Evaluation for Data Mining with Application to Fraud and Intrusion Detection: Results from the JAM Project. Technical Report. (2000)

Support Vector Machines (SVM) for Color Image Segmentation with Applications to Mobile Robot Localization Problems

An-Min Zou, Zeng-Guang Hou, and Min Tan

Key Laboratory of Complex Systems and Intelligence Science,
Institute of Automation, The Chinese Academy of Sciences,
P.O. Box 2728, Beijing 100080, China
{anmin.zou, zengguang.hou, min.tan}@mail.ia.ac.cn

Abstract. In autonomous mobile robot industry, the landmark-based localization method is widely used in which the landmark recognition plays an important role. The landmark recognition using visual sensors relies heavily on the quality of the image segmentation. In this paper, we use seat numbers as the landmarks, and it is of great importance to the seat number recognition that correctly segment the number regions from images. To perform this assignment, the support vector machine method is adopted to solve the color image segmentation problems because of its good generalization ability. The proposed method has been used for the mobile robot localization problems, and experimental results show that the proposed method can bring robust performance in practice.

1 Introduction

To accomplish some designated tasks in indoor environment, such as send mails and documents in the office, an autonomous mobile robot has to know where it is. Localization, the ability that a robot can find its position, is fundamental to autonomous mobile robots. Landmark-based localization method is most commonly used in which landmark recognition plays an important role. In recent years, monocular vision-based localization method has been widely used since we can obtain abundant information from visual sensors [1,2,3,4]. In 1988, Sugihara presented one of the pioneering studies in the landmark-based localization using monocular vision [1]. In his paper, the problem of the mobile robot localization was described as a geometric constraint satisfaction problem. Since then, Krotkov [2], and Munoz and Gonzalez [3] extended this work. In [2], the effects of observation uncertainty were also analyzed. In [3], the authors have analyzed the effects of *false positive* (observed landmark that do not correspond to a known landmark). In our previous work [4], we presented a neural network based camera calibration method for the global localization of mobile robots using monocular vision. We used the camera to measure the relative location between the floor landmark and the robot, whereas in [1,2,3], the camera was used to measure the bearing of one landmark relative another. We made use

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 443–452, 2005.

of the different distance between every two neighboring landmarks to solve the landmark match problem and the Kalman filter technique to deal with the uncertainties. In these papers, the authors have solved the localization problems by denoting the position of the robot as (x, y, θ) and assumed that the task of the landmark recognition had been accomplished *a prior*. However, in some cases, we need not know the position of the robot in detail. In this paper, we use seat numbers as the landmarks. Since the seat numbers can be used to distinguish the landmarks, the problem of establishing matches between observed landmark and a known landmark can be solved easily. Then the seat number recognition plays an important role in the mobile robot localization applications, and it is of great importance to the seat number recognition that correctly segment the number regions from images. For a mobile robot while moving in the office, it recognizes the seat number of the desk near it using visual sensor, and then localizes itself in indoor environment.

Image segmentation is the process of division of an image into regions with similar attributes [5]. The color image segmentation plays an important role in many applications such as robot vision, object recognition, and medical imaging. The methods for the color image segmentation can be roughly classified into [6]: (1) Threshold-based; (2) Clustering; (3) Region growing; (4) Edge-based; (5) Physics-based; (6) Fuzzy approaches; (7) Neural networks methods [7,8]. Neural networks offer two important properties in pattern recognition tasks: high degree of parallelism, which allows for very fast computational time and makes them suitable for real time applications, and good robustness to disturbances, which allows for reliable estimates [9]. In [7], a Hopfield neural network is used to solve the color image segmentation problems and two segmentation algorithms are presented based on the idea that describes the segmentation problems as minimizing a suitable energy function for a Hopfield network. The first algorithm consists of three different networks (each for one color feature considered), and then combines the results. The second builds a single network according to the number of clusters obtained by histogram analysis. In [8], a three-layer neural network is used to segment the medical stained images, where the possible classes are three, nuclear cell, interstitium, and background represented by three different colors. R, G, B, R^2, G^2, and B^2 of each pixel are used as the input layer and the three desired classes as the output layer.

Support vector machine (SVM) is one of the most recent algorithms in artificial neural networks, and this new learning algorithm was proposed by Vapnik and is based on the statistical learning theory [10]. An important characteristic of SVM is that while "most classical neural network algorithms require an *ad hoc* choice of system's generalization ability, the SVM approach proposes a learning algorithm to control the generalization ability of the system automatically" [11]. SVM has been used for pattern recognition, regression estimation, density estimation, and ANOVA decomposition [12]. For pattern recognition problems, SVM has been used for handwritten digit recognition, speak identification, charmed quark detection, face detection in images, and text categorization [12]. In this paper, to correctly segment the number regions from color images, SVM is adopted to solve the color image segmentation problems because of its good generalization ability.

This paper is organized as follows. SVM is slightly described in Section 2. The color image is segmented using SVM in Section 3. Experimental studies are presented in Section 4 and conclusions are given in Section 5.

2 Support Vector Machines

The SVM implements the following idea [10]: It maps the input vector x into a high-dimensional feature space F through some nonlinear mapping. In this space, an optimal separating hyperplane is constructed. In the simplest two classes classification problem, SVM performs the classification by finding a decision surface determined by certain points of the training set, termed Support Vectors [12].

Let the linearly non-separable training set D be $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R^m$ is the input vector, and $y_i \in \{-1, +1\}$ is the corresponding class of pattern. The SVM first maps x to $z = \Phi(x) \in F$, and then constructs an optimal separating hyperplane $\omega \cdot z + b$ in the high-dimensional feature space by solving the following quadric programming (QP) problem:

$$\text{minimize } \frac{1}{2}(\omega \cdot \omega) + C\sum_{i=1}^{n} \xi_i, \tag{1}$$

$$\text{subject to } y_i(\omega \cdot z_i + b) \geq 1 - \xi_i,\ i = 1, \cdots, n \tag{2}$$

where ω is the weight vector; C is the penalty term; ξ_i is the non-negative slack variable; b is the threshold, and $\omega \cdot z_i$ is the inner product of vectors ω and z_i. This QP problem is called the primal problem, which is difficult to be solved directly. By the duality theorem, the dual optimization problem has the same optimal value as the primal problem with the Lagrange multipliers providing the optimal solutions. Thus, we convert the primal optimization problem to a dual optimization problem, which is easier to be solved.

A Lagrange function for the primal problem is defined as follows:

$$L_P = \frac{1}{2}(\omega \cdot \omega) + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i\{y_i(\omega \cdot z_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \mu_i \xi_i \tag{3}$$

where α_i is the Lagrange multiplier, and μ_i is the Lagrange multiplier introduced to enforce positivity of the ξ_i.

By the optimality conditions and the duality theorem, we can obtain the dual form of the primal problem as follows:

$$\text{maximize } \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{4}$$

$$\text{subject to } 0 \leq \alpha_i \leq C,\ i = 1, \cdots, n \tag{5}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{6}$$

where $K(x_i, x_j) = z_i \cdot z_j$ is a kernel; α_i is the Lagrange multiplier and C is the penalty term. Therefore, we may use the kernel $K(x_i, x_j)$ to construct the optimal separating hyperplane in the high-dimensional feature space without computing $z_i = \Phi(x_i)$ explicitly.

The kernels for SVM mainly include three classes: polynomial kernel, radial basis function (RBF) kernel, and two-layer neural network kernel, whose expressions are given as follows, respectively:

$$K(x_i, x) = (x_i \cdot x + 1)^p, \tag{7}$$

$$K(x_i, x) = \exp\left(-q \left\|x_i - x\right\|^2\right), \tag{8}$$

$$K(x_i, x) = \tanh(vx_i \cdot x + c). \tag{9}$$

In the solution of the dual QP problem, the x_i for which $\alpha_i \neq 0$ are defined as the support vectors, which are the critical elements of the training set [12], and the SVM decision function is obtained by the support vectors:

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right) \tag{10}$$

where x_i is the support vector; x is the input vector to classify; l is the number of the support vectors and b is the threshold.

3 Color Image Segmentation Using SVM

Many color spaces may be used in image processing, such as RGB, YUV, YIQ, YCrCb, HIS, etc. The most common is RGB space where colors are represented by their red, green, and blue components in an orthogonal Cartesian space. YUV, YIQ, and YCrCb spaces can be respectively obtained by linear transformation of R, G, and B, and HIS space can be obtained by nonlinear transformation. The RGB space is a natural choice for image processing since it avoids costly conversions of images into a different color space. In this paper, we use seat numbers as the landmarks, and it is of great importance to the seat number recognition that correctly segment the number regions from images. Since the object of the segmentation is that we have to correctly segment the number regions from images, we can take the images as two classes. One is the number region, i.e. the object, and the other is the remaining region, i.e. the background. The image segmentation can be taken as classification problems, which can be solved using anyone of well-known classification techniques. Since SVM is one of the classification techniques and good results of the SVM technique in pattern recognition have been obtained, the SVM is chosen for solving color image segmentation problems.

The SVM for color image segmentation is shown in Fig. 1. The inputs x^1, x^2, and x^3 are obtained from the red, green, and blue components of each pixel, respectively:

$$x^1 = \frac{R}{255}, \tag{11}$$

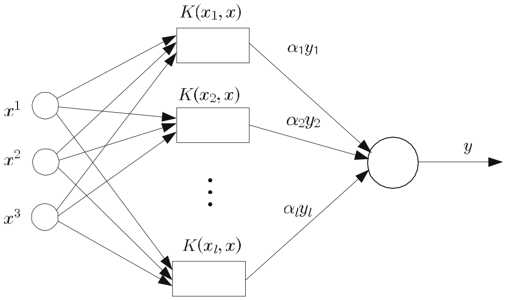


Fig. 1. Structure of the SVM for color image segmentation: x_i is the support vector, $i = 1, \cdots, l$; l is the number of the support vectors; the input vector is $x = (x^1, x^2, x^3)^T$; the output of the SVM is $y = \text{sign}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\right)$, if it is 1, then the pixel is the object; otherwise the pixel is the background

$$x^2 = \frac{G}{255}, \tag{12}$$

$$x^3 = \frac{B}{255}. \tag{13}$$

Training of SVM requires the solution of a QP optimization problem, which is a large-scale system optimization problem. Various algorithms have been proposed to solve the optimization problem of large-scale system. In [13,14], recurrent neural networks are used to solve the large-scale system optimization problems that have non-quadratic objective functions. For large QP optimization problem, Vapnik presented "chunking" algorithms to solve the SVM QP problems [15]. The chunking algorithm breaks the large QP problem into a series of smaller QP sub-problems, whose ultimate goal is identify all of the non-zero Lagrange multipliers and discard all of the zero Lagrange multipliers. Thus, the chunking algorithm requires the memory to store a matrix that has a number of elements approximately equal to the square of the number of non-zero Lagrange multipliers; however, it is difficult to be satisfied. The decomposition method is similar to chunking algorithm and this method breaks the large QP problem into fixed-size QP sub-problems [16]. However, the decomposition method requires the use of a numerical QP library, which can be costly or slow [17]. Similar to the decomposition method, sequential minimal optimization (SMO) decomposes the overall QP problem into fixed-size QP sub-problems (each involves only two Lagrange multipliers) and these sub-problems are solved analytically [17]. SMO algorithm is one of the efficient algorithms for solving the large QP problem, which is used to train the SVM.

4 Experimental Studies

In this Section, we will provide some experimental results using our proposed algorithm. The mobile robot CASIA-I used in our experiment is shown in Fig. 2,

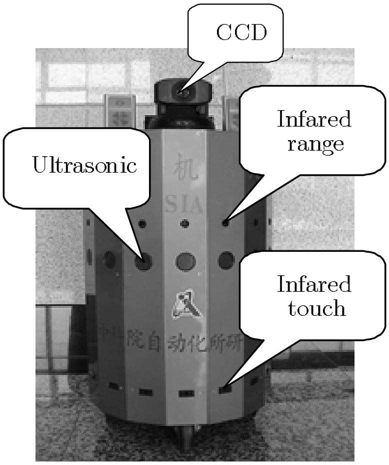


Fig. 2. Mobile robot CASIA-I

which has a CCD camera, 16 ultrasonic sensors, 16 infrared range sensors, and 16 infrared touch sensors. The size of the captured image is 240×360 pixels.

The experimental environment is shown in Fig. 3. We use the seat numbers of 10 desks in our laboratory as the landmarks, which are from 1 to 10. When the robot moves to the location near a desk, it uses the CCD camera to scan the environment around itself for searching the seat number, and then uses the proposed SVM-based method to segment the number region from the image captured by the camera, and then recognizes the seat number. In this way, the robot can localize itself in indoor environment.

In order to segment color images using SVM, the SVM has to be trained first. The size of the original color image used to train the SVM is 240×360 pixels. In order to speed up the training of the SVM, the size of the training

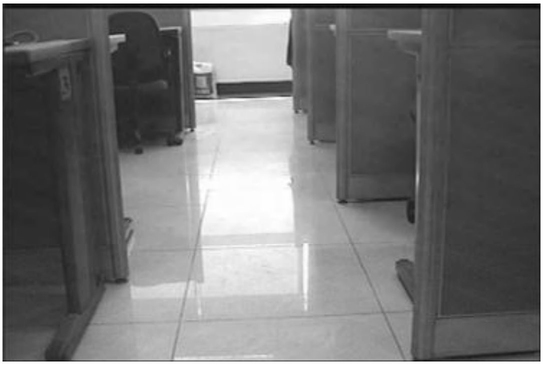
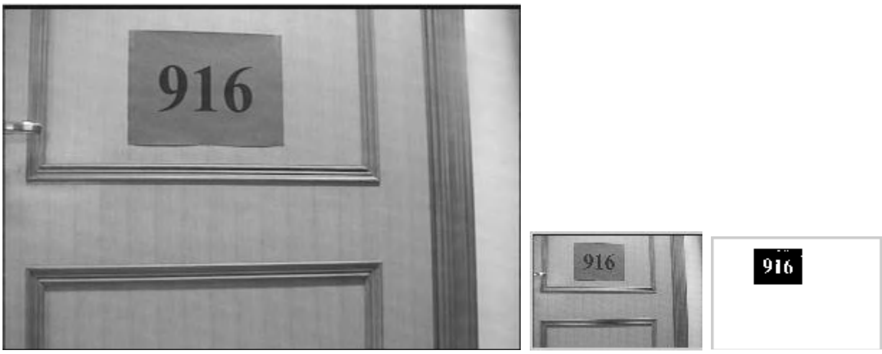


Fig. 3. Experimental environment



(a) (b) (c)

Fig. 4. The training images: (a) Original true color image; (b) Reduced image of (a), whose size is 1/9 of that of (a); (c) Binary image of (b), which is obtained by threshold-based method

Table 1. Number of support vectors and classification accuracy rate

$C = 100$			Number of support vectors	Classification accuracy rate
Polynomial kernel	p	2	91	99.73%
		3	80	99.73%
		4	77	99.81%
RBF kernel	q	0.5	113	99.73%
		5	71	99.86%
		10	59	99.91%

set is reduced to 1/9 of the size of original color image, then the number of the training set is 9600. Fig. 4 shows the images used to train the SVM. Fig. 4 (b) is the reduced image of (a), whose size is 80×120 pixels. Fig. 4 (c) is the binary image of (b), which is obtained by the following threshold-based method:

$$(R - B) > threshold_1 \textbf{ and } (R - G) > threshold_2 \textbf{ and } |G - B| < threshold_3. \quad (14)$$

If it is true, then the pixel is the object; otherwise the pixel is the background. The values of $threshold_1$, $threshold_2$ and $threshold_3$ are determined by the experiments, respectively.

Since an SVM is largely characterized by the choice of its kernel [12], we compare the number of the support vectors and classification accuracy rates by the algorithm using different kernels and different values of the parameters of kernels as shown in Table 1. The value of the penalty term is 100. We can see in the table that the SVM using RBF kernel with $q = 10$ has fewer support vectors and higher classification accuracy rate, which is chosen as the kernel of the SVM for color image segmentation.

In Table 2, we compare the number of the support vectors and classification accuracy rates by the algorithm using different values of C. We use the RBF kernel with $q = 10$ as the kernel of the SVM. C is the penalty term; a larger C corresponding to a higher penalty to errors during learning [12]. From the table we can see that, the bigger the value of C becomes, the fewer support vectors and the higher classification accuracy rate are obtained.

Table 2. Number of support vectors and classification accuracy rate

RBF kernel ($q = 10$)	C				
	1	10	100	1000	10000
Number of support vectors	203	102	59	41	37
Classification accuracy rate	99.61%	99.80%	99.91%	99.92%	99.92%

The results show that the number of the support vectors can be reduced obviously by the algorithm using different kernels, or different values of the parameters of kernels, or different values of penalty term, while the classification accuracy rate is high and is improved slowly. The SVM using RBF kernel with $q = 10$ and $C = 10000$ is chosen for the color image segmentation because by which we can obtain fewer support vectors and higher classification accuracy rate.

The proposed SVM-based method has been used for the mobile robot localization applications. Fig. 5 shows segmentation results using the proposed SVM-based method. Though there is a quite apparent diversity between the testing images and the training image, we are able to obtain good results using the proposed SVM-based method. The results show that the proposed SVM-based method can bring robust performance for the mobile robot localization.

5 Conclusions

In this paper, an SVM-based color image segmentation method is presented for the mobile robot localization applications. We use seat numbers as the landmarks, and it is of great importance to the seat number recognition that correctly segment the number regions from images. To perform this assignment, SVM is adopted to solve the color image segmentation problems because of its good generalization ability. We compare the number of support vectors and classification accuracy rates by the algorithm using different kernels, different values of the parameters of kernels, and different values of penalty term in order to choose an appropriate SVM. The proposed SVM-based method has been used for the mobile robot localization applications. The experimental results show that the proposed SVM-based method can bring robust performance in practice. The results also show that we are able to use seat number to localize the robot in indoor environment.

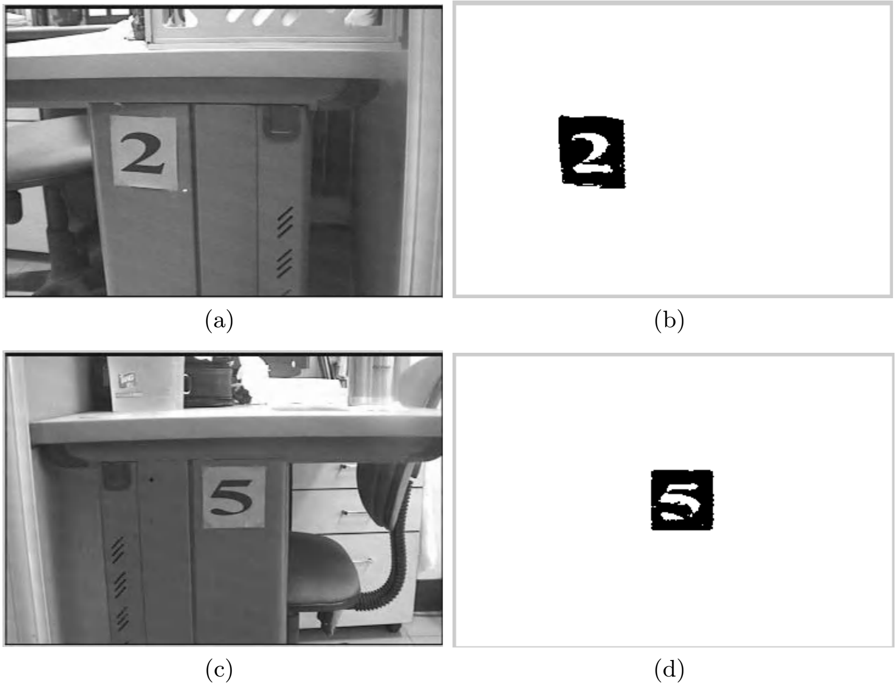

(a) (b) (c) (d)

Fig. 5. (a) and (c): True color images, which contain seat numbers in real indoor environment; (b) and (d): Binary images of (a) and (c), which are obtained respectively by the proposed method

Acknowledgements

This research has been supported in part by the National Natural Science Foundation of China (Grant Nos. 60205004, 50475179 and 60334020), the National Basic Research Program (973) of China (Grant No. 2002CB312200), the Hi-Tech R&D Program (863) of China (Grant Nos. 2002AA423160 and 2005AA420040), and the Science and Technology New Star Program of Beijing (Grant No. H02082 0780130).

References

1. Sugihara, K.: Some location problems for robot navigation using a single camera. Computer Vision, Graphics, and Image Processing, **42** (1988) 112-129
2. Krotkov, E.: Mobile robot localization using a single image. Proceedings of the IEEE International Conference on Robotics and Automation, **2** (1989) 978-983
3. Munoz, A. J., Gonzalez, J.: Two-dimensional landmark-based position estimation from a single image. Proceedings of the IEEE International Conference on Robotics and Automation, **4** (1998) 3709-3714

4. Zou, A. , Hou, Z.G., Zhang, L., Tan, M.: A neural network-based camera calibration method for mobile robot localization problems. Proceeding of the Second International Symposium on Neural Networks, Chongqing, China, Springer Lecture Notes in Computer Sciences (LNCS), **3498** (2005) 277-284
5. Pratt, W. K.: Digital Image Processing. John Wiley & Sons, Inc. 2nd ed (1991)
6. Cheng, H. D., Jiang, X. H., Sun, Y., Wang, J.: Color image segmentation: advances and prospects. Pattern Recognition, **34** (12) (2001) 2259-2281
7. Campadelli, P., Medici, D., Schettini, R.: Color image segmentation using Hopfield networks. Image and Vision Computing, **15** (3) (1997) 161-166
8. Okii, H., Kaneki, N., Hara, H., Ono, K.: Automatic color segmentation method using a neural network model for stained images. IEICE Transactions on Information and Systems (Japan), **E77-D** (3) (1994) 343-350
9. Lucchese, L., Mitra, S. K.: Color image segmentation: a state-of-the-art survey. Image Processing, Vision, and Pattern Recognition. Proceedings of the Indian National Science Academy (INSA-A), New Delhi, India, **67 A** (2) (2001) 207-221
10. Vapnik, V.: Statistical Learning Theory. John Willey & Sons (1998)
11. Pontil, M., Verri, A.: Properties of support vector machines. Artificial Intelligence Laboratory, C.B.C.L., MIT Press (1997)
12. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, Kluwer Academic Publishers, **2** (2) (1998) 121-167
13. Hou, Z.G.: A hierarchical optimization neural network for large-scale dynamic systems. Automatica, **37** (12) (2001) 1931-1940
14. Hou, Z.G.,Wu, C.P., Bao.: A neural network for hierarchical optimization of optimization of nonlinear large-scale systems. International Journal of Systems Science, **29** (2) (1998) 159-166
15. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer-Verlag, (1982)
16. Osuna, E., Freund, R., Giros, F.: Improved training algorithm for support vector machines. Proceedings of the IEEE workshop on Neural Networks for Signal Processing, (1997) 276-285
17. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In Advances in Kernel Methods - Support Vector Learning, MIT Press (1999) 42-65

Fuzzy Logic Based Feedback Scheduler for Embedded Control Systems

Feng Xia, Xingfa Shen, Liping Liu, Zhi Wang, and Youxian Sun

National Laboratory of Industrial Control Technology,
Institute of Modern Control Engineering,
Zhejiang University, Hangzhou 310027, China
{xia, xfshen, lpliu, wangzhi, yxsun}@iipc.zju.edu.cn

Abstract. The case where multiple control tasks share one embedded CPU is considered. For various reasons, both execution times of these tasks and CPU workload are uncertain and imprecise. To attack this issue, a fuzzy logic based feedback scheduling approach is suggested. The sampling periods of control tasks are periodically adjusted with respect to uncertain resource availability. A simple period rescaling algorithm is employed, and the available CPU resource is dynamically allocated in an intelligent fashion. Thanks to the inherent capacity of fuzzy logic to formalize control algorithms that can tolerate imprecision and uncertainty, the proposed approach provides runtime flexibility to quality of control (QoC) management. Preliminary simulations highlight the benefits of the fuzzy logic based feedback scheduler.

1 Introduction

Nowadays, embedded computing is playing an increasingly important role in the engineering community. More and more real-time control applications are built upon embedded systems. Due to economical and/or technical reasons, these systems representatively have constrained resources like limited computing speed [1,2]. Such examples can be seen in many fields such as automotive electronics, aircraft, robotics, etc. At the same time, practical control applications are becoming more and more complex. It is not uncommon for several independent control and/or non-control tasks to reside in one embedded CPU. In these cases, the limited CPU time becomes the shared resource for which several tasks compete. The performance of embedded control systems will then depend not only on the controller design, but also on the efficient scheduling of the shared computing resources [3].

Besides resource constraints, workload uncertainty is also unavoidable in today's embedded control systems. For a shared CPU, the workload may exhibit uncertain characteristics [4,5] due to: 1) task activation or termination, and 2) non-deterministic behavior of the underlying platforms. In a resource-constrained environment, this workload uncertainty will cause the allowable CPU utilization for control purpose to be limited and unexpectedly variable. Moreover, the computation time of a control algorithm may vary significantly, e.g. due to the changes in the data to be processed. This makes the requested CPU time of control task(s) also uncertain. Since noise inevitably exists in the measurement of tasks' timing parameters, the measured CPU

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 453 – 462, 2005.

utilization is imprecise, if not unavailable. As a consequence of separate concerns of control and scheduling [1], the quality of control (QoC) may be significantly impacted with traditional control systems design methodologies.

To address these problems, we follow the newly emerging methodology of feedback scheduling and suggest an intelligent approach based on fuzzy logic. Feedback control and real-time scheduling are synthetically integrated in embedded control systems. The motivation for using fuzzy logic for feedback scheduling of control tasks is mainly raised from its inherent capacity to formalize control algorithms that can tolerate imprecision and uncertainty [6]. Our goal is to provide runtime adaptation and flexible QoC management in the presence of CPU resource constraint as well as workload uncertainty. In our previous work [5], we have considered the case where only one single control task competes for CPU resources with other non-control tasks. Here we will instead focus on another typical case where multiple control tasks execute concurrently on the CPU. A fuzzy logic based feedback scheduler is developed to dynamically adjust the sampling periods of all control tasks. Preliminary simulations are conducted to evaluate its performance.

The rest of this paper is structured as follows. Section 2 observes related works associated with our study. The problem we consider here is described in Section 3. The fuzzy feedback scheduler is presented in Section 4. Section 5 evaluates the performance of the proposed approach. Section 6 concludes this paper.

2 Related Work

In recent years, the emerging field of codesign of control and scheduling has received considerable amount of attention, both from the control community and the computing community. The control and scheduling codesign problem has been stated in [1] as follows: *given a set of processes to be controlled and a computer with limited computational resources, design a set of controllers and schedule them as real-time tasks such that the overall control performance is optimized.* Examples of integrated offline design of control algorithms and scheduling algorithms include [7] and [8], etc. In these works, the codesign problem is treated as a performance optimization problem, and their strategies are static. On the contrary, what we concern is online integration of control and scheduling theory.

In the literature, there is a special interest in the use of feedback in scheduling algorithms for applications where the workload dynamics cannot be characterized accurately. Cervin and Eker [9] present a feedback scheduling mechanism for hybrid controllers where the execution time may change abruptly between different modes. The proposed solution attempts to keep the CPU utilization at a high level, avoid overload, and distribute the computing resources evenly among the tasks. Eker et al [10] design an optimal feedback scheduler to distribute computing resources over a set of real-time control loops in order to optimize the total control performance. Its approximation versions are exploited in [11], where feedback scheduling is performed by simple rescaling of the task periods. In [3], we utilize neural network technique in feedback scheduling of control tasks, and provide an almost optimal solution. Also in our previous work [5], we suggest an intelligent control theoretic approach to feedback scheduling based on fuzzy logic control technology. The case with one single control task that competes for CPU resources with other non-control tasks is considered, which is different from the scenario we consider in this paper. Feedback scheduling of anytime controllers is the topic of [12,13,14]. Preliminary results on dynamic scheduling of model predictive controllers in which a quadratic optimization problem is solved iteratively in every sample are presented in [12]. A feedback scheduling approach is employed in [13] to attack the impact of resource constraints on a class of iterative control algorithms. In [14], a fuzzy feedback scheduler is proposed to improve the performance of iterative optimal control applications. As an anytime algorithm, the execution time of the controller is dynamically adjusted while the sampling period is assumed to be constant. In this work, not execution times but sampling periods of control tasks will be adjusted by the feedback scheduler.

Lu et al [15] use a control theory based methodology to systematically design FCS (Feedback Control real-time Scheduling) algorithms to satisfy the transient and steady state performance specifications of real-time systems. Abeni et al [16] apply control theory to a reservation-based feedback scheduler and provide a precise mathematical model of the scheduler. An elastic methodology for automatically adapting the rates of a periodic task set is proposed in [17]. Actual executions are monitored by a runtime mechanism and used as feedback signals for predicting the actual load and achieving rate adaptation. The applications of feedback scheduling in networked control systems are presented in [18,19]. Although without explicit declaration, the methodology of feedback scheduling is also employed in other works, e.g. [20].Many further references can be found in [21,22,23].

3 The Feedback Scheduling Problem

As shown in Fig.1, we consider an embedded control system where a set of control tasks share one processor. Each of the tasks is responsible for controlling an individual plant. For convenience, all control tasks are assumed to be periodic, and they are independent of each other.

Let each control task execute with a sampling period h_i and an execution time C_i. According to basic real-time scheduling theory, the CPU utilization of the task set will then be $U = \sum_{i=1}^{n}(C_i / h_i)$, where n is the number of control tasks. In practical applications, the tasks' execution times may be unexpectedly variable, and their available values are imprecise. This causes the actual (measured) CPU utilization to be uncertain and imprecise.

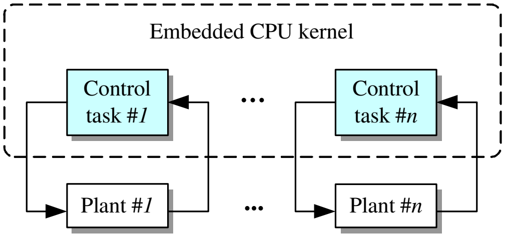


Fig. 1. An embedded control system with multiple concurrent control tasks

In the control community, control algorithms are often designed without taking into account computing resource availability. In the system considered, a new instance of a control task is not allowed to start execution until the previous instance has completed. It is common that the tasks are assigned fixed priorities. In this environment, when the computing resources become scarce, the time delay of low-priority control tasks will be cumulated due to continuous preemption from high-priority tasks. The control performance of the low priority loops will consequently be degraded, and even destabilized sometimes. Therefore, in order to achieve good QoC in each control loop, task schedulability condition should not be violated. That is

$$U = \sum_{i=1}^{n} (C_i / h_i) \le U_{sp} \tag{1}$$

where U_{sp} is the allowable (desired) CPU utilization, and it may change over time.

For a well-designed control algorithm, as we know, the QoC will highly depend on the sampling period [24]. Generally speaking, smaller sampling period leads to better QoC. Under the schedulability constraint in (1), the sets of smallest feasible sampling periods, which result in best control performance, must imply maximum use of the CPU. In this sense, the role of feedback scheduling will be to manipulate the sampling periods with respect to computing resource availability variations so that the CPU utilization U keeps as close as possible to the utilization setpoint U_{sp}.

Ideally, the sampling periods can be determined by solving an optimization problem as in [10]. However, the solutions may not be so optimal as expected due to the complex uncertain characteristics of task execution time and CPU workload. In extreme cases, the schedulability constraint may be violated. Furthermore, such an optimization routine is too computationally expensive to be used online and will cause a huge scheduling overhead. Instead, we refer to an intelligent control theoretic approach based on fuzzy logic.

4 Fuzzy Logic Based Feedback Scheduler

In this section, we propose a fuzzy logic based feedback scheduler for the system given in Fig.1. As a formal methodology for representing, manipulating, and implementing a human's heuristic knowledge, fuzzy logic provides a simple and flexible way to arrive at a definite conclusion based upon imprecise, noisy, or incomplete input information [6]. The inherent simplicity of fuzzy logic further makes it suitable for real-time applications like feedback scheduling.

4.1 Architecture

As mentioned above, the feedback scheduler is employed to dynamically adjust the sampling periods of all control tasks. It is implemented as a periodic task that runs concurrently with other tasks in the CPU. Instead of directly determining the sampling periods as done in [3,5], here we utilize a simple rescaling approach. Let $h_{i,N}$ denote the period of control task i at the Nth sampling instance of the feedback scheduler. Note that the period of the feedback scheduler is different from that of any control task. At each time instance N, the sampling periods will be recalculated by

$$h_{i,N+1} = \eta_N h_{i,N} \quad i = 1,\ldots,n \tag{2}$$

where η is the *rescaling factor*. The basic principle behind the simple rescaling algorithm is similar to elastic scheduling [17].

In ideal cases where both execution times and CPU utilization are accurately known, the rescaling factor can be obtained by $\eta_N = U_N / U_{sp}$, where U_N denotes current CPU utilization of the task set. One can notice that η varies with U_N from instance to instance. After all sampling periods are rescaled according to (2), the requested CPU utilization will then be

$$U_{N+1} = \sum (C_i / h_{i,N+1}) = \frac{1}{\eta_N} \sum (C_i / h_{i,N}) = \frac{U_{sp}}{U_N} U_N = U_{sp} \tag{3}$$

As we can see from (3), using the simple rescaling algorithm in (2), the requested CPU utilization of all control tasks will be exactly the desired utilization. However, this is only true for ideal cases. In this paper, the rescaling factor η_N will be determined using fuzzy technique. The architecture of the fuzzy feedback scheduling methodology is given in Fig.2.

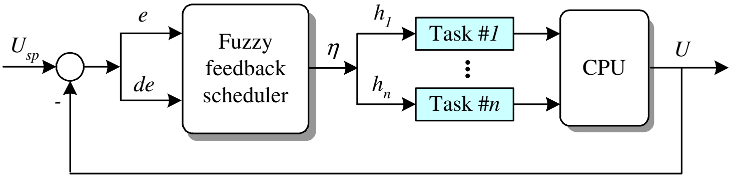


Fig. 2. Fuzzy feedback scheduling of a multitasking system

From the control perspective, the scheduler can be viewed as a feedback controller. The controlled variable is the CPU utilization. The sampling periods of all control tasks act as the manipulated variables. The fuzzy feedback scheduler attempts to control the CPU utilization to the desired level. It gathers the actual CPU utilization U and current allowable utilization U_{sp}, compares them, and then decides in an intelligent fashion what the rescaling factor should be. Finally, all sampling periods will be rescaled with this proportional factor.

4.2 Design Methodology

The internal structure of the fuzzy feedback scheduler is given in Fig.3. Similar to almost all fuzzy controllers [6], the fuzzy feedback scheduler consists of four main components: fuzzifier, rule base, inference mechanism, and defuzzifier. The vast amount of knowledge and experience of fuzzy applications in control community can be borrowed in the construction of feedback schedulers. Following the methodology of fuzzy control technology, we design the fuzzy feedback scheduler in the following.

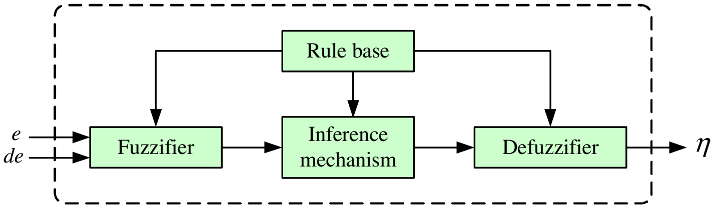


Fig. 3. Internal structure of fuzzy feedback scheduler

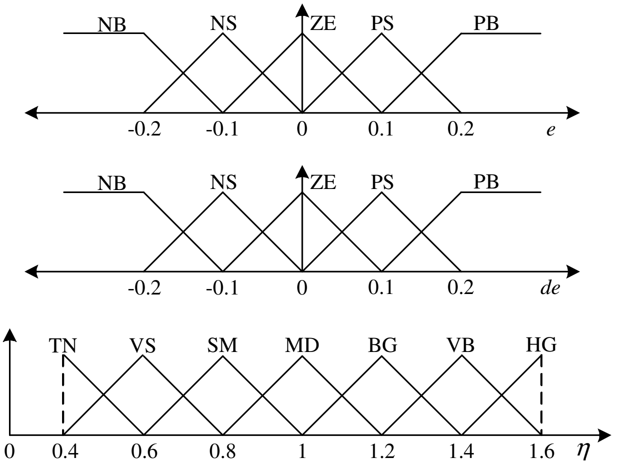


Fig. 4. Membership functions for input and output linguistic variables

The first step is to identify the input and output variables, i.e. choice and scaling of the linguistic variables. There are two input variables, the error between desired CPU utilization and actual CPU utilization, $e = U_{sp} - U$, and the change in error, $de = e_N - e_{N-1}$. The input linguistic variables are denoted by *E* and *DE*, respectively. The output of the scheduler is the period rescaling factor η, and the corresponding linguistic variable is denoted by *RF*.

Next, we specify meaningful linguistic values and the membership functions for each linguistic variable. In this work, the linguistic values are the same for input linguistic variables, which are NB (negative big), NS (negative small), ZE (zero), PS (positive small), and PB (positive big). The linguistic values of *RF* are TN (tiny), VS (very small), SM (small), MD (medium), BG (big), VB (very big), and HG (huge). Fig.4 shows the membership functions of *E*, *DE*, and *RF*, respectively.

The rule base will then be set up. Since feedback scheduling is a relatively new research area, we attempt to construct the rule base based on our experience in simulation studies and the well-established knowledge both from the control community and the real-time scheduling community. Table 1 describes all possible rules used in the fuzzy feedback scheduler. Each of them can be expressed as: if *E* is E_i and *DE* is DE_j, then *RF* is RF_{ij}, $\forall$ i= 1,…5, j=1,…5.

Table 1. Rule base within fuzzy feedback scheduler

RF		*DE*				
		NB	NS	ZE	PS	PB
E	NB	HG	HG	HG	VB	BG
	NS	HG	VB	BG	BG	MD
	ZE	VB	BG	MD	MD	SM
	PS	BG	MD	SM	SM	VS
	PB	MD	SM	SM	VS	TN

The fuzzy sets representing the conclusions will be obtained using the *max-min* inference mechanism. In the defuzzification procedure, we employ the *center of gravity* method to produce a crisp output.

Remark 1. In order to keep the scheduling overhead relatively small, one can employ a look-up table scheme when designing the fuzzy feedback scheduler.

5 Example

In this section, we consider an embedded control system that consists of three independent control loops. The transfer functions of these plants are given in (4). Each plant is controlled using a well-designed PID [24] algorithm. The period of each control task is equal to the corresponding sampling period.

$$G_1(s)=\frac{120}{s^2+35s+216},\quad G_2(s)=\frac{1000}{s^2+22s+72},\quad G_3(s)=\frac{980}{s^2+20s+60} \tag{4}$$

In our experiments, the nominal sampling periods of three loops are set to be 10 ms, 8 ms, and 6 ms, respectively. We assume that all sampling periods are always accurately known. The execution time of each control task varies according to the normal distribution with the same mean of 3 ms. A fixed-priority real-time kernel is utilized in the embedded CPU, and three control tasks are assigned rate monotonic (RM) priorities. Accordingly, control task 1 has the lowest priority and control task 3 has the highest priority among them. Using the Matlab/TrueTime toolbox [1], co-simulations of the controllers, the feedback scheduler and the plants are conducted as follows. At the start of the simulation, i.e. t = 0, control task 1 and 2 are switched on. Control task 3 remains off until time t = 1s. At t = 2s, an interfering task with a priority higher than that of control task 3 is released in the CPU. The requested utilization of the interfering task is 0.3.

In the first experiment (Case I), three control tasks work with their nominal periods all along. Before time t = 1s, there are only two tasks, and their average requested CPU utilization approximates (3/10+3/8) = 0.675. Because it is below the schedulable utilization of the RM algorithm for two tasks, which is 0.828, both plants perform well, see Fig.5 and 6. From time t = 1s, the average requested CPU utilization of the task set becomes (3/10+3/8+3/6) = 1.175. This then leads to an overload condition. As a result, control loop 1 becomes unstable, see Fig.5. Since control task 1 holds the

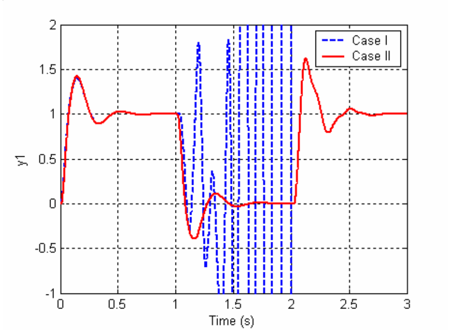


Fig. 5. Control performance of plant 1

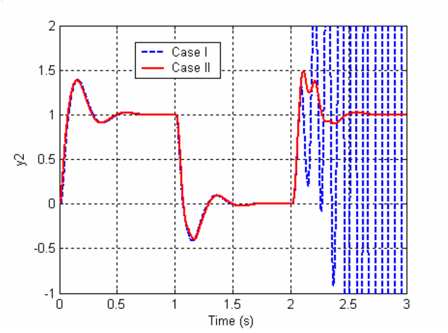


Fig. 6. Control performance of plant 2

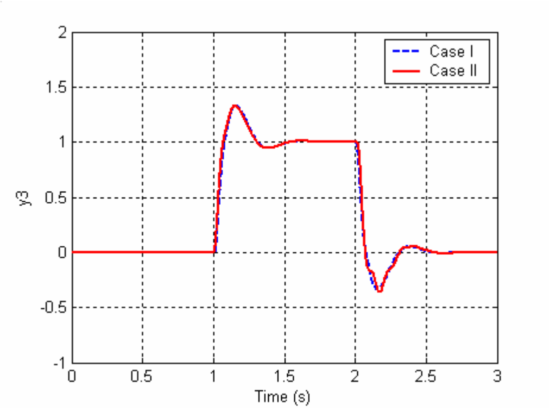


Fig. 7. Control performance of plant 3

lowest priority, it suffers the most during the overload. As shown in Fig.6 and 7, both control loop 2 and 3 still exhibit good performance. After the interfering task is released at t = 2s, both control loop 1 and 2 turn to be unstable, only the third plant performs well. Note that in Fig.5, the output of the first plant in this experiment (dashed blue line) goes out of the figure's scope after t = 2s.

In the second experiment (Case II), we implement the fuzzy logic based feedback scheduler presented in Section 4. The sampling periods of control tasks are dynamically adjusted in an intelligent fashion. The period of the feedback scheduler is set to be 25 ms and the scheduling overhead is neglected. The PID parameters are online updated with respect to control period adjustment. At time t = 1s, U_{sp} changes from 0.9 to 0.8, and then it changes to 0.5 at t = 2s. The resulting control performance is also given in Fig.5, 6 and 7. As we can see, the QoC is improved with the help of the fuzzy feedback scheduler, especially when the computing resources become scarce after t = 1s. All plants remain stable and exhibit satisfactory performance throughout the experiment. The results argue that the proposed approach can effectively handle the uncertain characteristics of CPU resource availability and provides flexible QoC management in multitasking embedded control systems.

6 Conclusion

In this paper, we suggest an intelligent control theoretic approach to feedback scheduling of multitasking embedded control systems. The well-known fuzzy logic is employed in the construction of the feedback scheduler. Combined with a simple rescaling algorithm, the fuzzy feedback scheduler provides runtime flexibility to QoC management of the whole system. The uncertain and imprecise behavior of the CPU resource availability is also effectively mastered. The potential benefits of the fuzzy feedback scheduler include: 1) graceful QoC degradation under overload conditions, 2) effectiveness in dealing with timing uncertainty inside CPU scheduling, 3) systematic design methodology, and 4) simple implementation.

References

1. Årzén, K.-E., Cervin, A.: Control and Embedded Computing: Survey of Research Directions. In: Proc. 16th IFAC World Congress, Prague, Czech Republic (2005)
2. Xia, F., Wang, Z., Sun, Y.: Integrated Computation, Communication and control: Towards Next Revolution in Information Technology. LNCS 3356, Springer-Verlag (2004) 117-125
3. Xia, F., Sun, Y.: Neural Network Based Feedback Scheduling of Multitasking Control Systems. In: Proc. KES2005, LNCS, Springer-Verlag (2005) (to appear)
4. Buttazzo, G., Velasco, M., Martí, P., Fohler, G.: Managing Quality-of-Control Performance Under Overload Conditions. In: Proc. ECRTS, Catania, Italy (2004) 1-8
5. Xia, F., Liu, L., Sun, Y.: Flexible Quality-of-Control Management in Embedded Systems Using Fuzzy Feedback Scheduling. In: Proc. RSFDGrC, LNCS, Springer-Verlag (2005) (to appear)
6. Passino, K.M., Yurkovich, S.: Fuzzy Control. Addison Wesley Longman, Menlo Park, CA (1998)
7. Seto, D., Lehoczky, J.P., Shin, K.G.: Trade-off Analysis of Real-time Control Peroformance and Schedulability. Real-Time Systems 21 (2001) 199-217
8. Ryu, M., Hong, S., Saksena, M.: Streamlining Real-time Controller Design: From Performance Specifications to End-to-end Timing Constraints. In: Proc. 3rd IEEE RTAS (1997) 91-99
9. Cervin, A., Eker, J.: Feedback Scheduling of Control Tasks. In: Proc. 39th IEEE CDC (2000) 4871-4876
10. Eker, J., Hagander, P., Årzén, K.-E.: A Feedback Scheduler for Real-time Controller Tasks. Control Engineering Practice 8:12 (2000) 1369-1378
11. Cervin, A., Eker, J., Bernhardsson, B., Årzén, K.-E.: Feedback-Feedforward Scheduling of Control Tasks. Real-Time Systems 23:1 (2002) 25-53
12. Henriksson, D., Cervin, A., Åkesson, J., Årzén, K.-E.: Feedback Scheduling of Model Predictive Controllers. In: Proc. 8th IEEE RTAS, San Jose, CA (2002) 207-216
13. Xia, F., Sun, Y.: NN-based Iterative Learning Control under Resource Constraints: A Feedback Scheduling Approach. LNCS 3498, Springer-Verlag (2005) 1-6
14. Xia, F., Sun, Y.: Anytime Iterative Optimal Control Using Fuzzy Feedback Scheduler. In: Proc. KES2005, LNCS, Springer-Verlag (2005) (to appear)
15. Lu, C., Stankovic, J.A., Tao, G., Son, S.H.: Feedback Control Real-time Scheduling: Framework, Modeling, and Algorithms. Real-time Systems 23:1/2 (2002) 85-126

16. Abeni, L., Palopoli, L., Lipari, G., Walpole J.: Analysis of a Reservation-Based Feedback Scheduler. In: Proc. 23rd IEEE RTSS, Austin, Texas (2002) 71-80
17. Buttazzo, G., Abeni, L.: Adaptive Rate Control through Elastic Scheduling. In: Proc. 39th IEEE CDC, Sydney, Australia (2000) 4883-4888
18. Xia, F., Li, S., Sun, Y.: Neural Network Based Feedback Scheduler for Networked Control System with Flexible Workload. In: Proc. ICNC'05, LNCS, Springer-Verlag (2005) (to appear)
19. Xia, F., Dai, X., Wang, Z., Sun, Y.: Feedback Based Network Scheduling of Networked Control Systems. In: Proc. 5th ICCA, Budapest, Hungary (2005)
20. Abdelzaher, T., Atkins, E.M., Shin, K.G.: QoS Negotiation in Real-Time Systems and Its Application to Automated Flight Control. IEEE Trans. on Computers 49:11 (2000) 1170-1183
21. Årzén, K.-E., Bernhardsson, B., Eker, J., Cervin, A., Persson, P., Nilsson, K., Sha, L.: Integrated Control and Scheduling. Research report: ISSN 0820-5316, Lund Institute of Technology, Sweden (1999)
22. Sha, L., Abdelzaher, T., Årzén, K.-E., Cervin, A., Baker, T., Burns, A., Buttazzo, G., Caccamo, M., Lehoczky, J., Mok, A.K.: Real Time Scheduling Theory: A historical perspective. Real-Time Systems 28 (2004) 101-155
23. Xia, F., Yin, H., Wang, Z., Sun, Y.: Theory and Practice of Real-time Scheduling in Networked Control Systems. In: Proc. 17th Chinese Control and Decision Conference, Harbin, China (2005) (in Chinese)
24. Åström, K.J., Wittenmark, B.: Computer Controlled Systems. Prentice Hall (1997)

The Application of FCMAC in Cable Gravity Compensation

Xu-Mei Lin[1,2], Tao Mei[2], Hui-Jing Wang[2], and Yan-Sheng Yao[1,2]

[1] Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China
[2] Department of Precision Machinery and Precision Instrumentation, School of Engineering Science, University of Science and Technology of China, Hefei, 230026, China
xmlin@iim.ac.cn

Abstract. The cable compensation system is an experiment system that performs simulations of partial or microgravity environments on earth. It is a highly nonlinear and complex system. In this paper, a network based on the theory of the Fuzzy Cerebellum Model Articulation Controller (FCMAC) is proposed to control this cable compensation system. In FCMAC, without appropriate learning rate, the control system based on FCMAC will become unstable or its convergence speed will become slow. In order to guarantee the convergence of tracking error, we present a new kind of optimization based on adaptive GA for selecting learning rate. Furthermore, this approach is evaluated and its performance is discussed. The simulation results shows that performance of the FCMAC based the proposed method is stable and more effective.

1 Introduction

Making microgravity on the ground would provide a very effective laboratory for space robotic systems. In this microgravity laboratory, we can achieve the robot grabbing and maintaining in outer space environment. The conducting earthbound research on robotic applications for outer space environments includes the neutral buoyancy, Air-bearing floors, Drop Towers and Tubes and passive counterweight, active motor control [1]. The neutral buoyancy, Air-bearing floors introduce effects of fluid or air dynamics and complexity of underwater or air operation but can be used with any number of objects for complex simulations and can provide long time training[2][3]. Drop Towers and Tubes can provide high precision but its experimental time is very short. The passive counterweight is simple and cheap, but brings additional inertia and friction. The active motor control does not add inertia, but the controller is very complex.

In this study, we propose to actively control the three-dimensional direction of the gravity compensation system. This system will be an active motor control schemes. The active control system is illustrated in Fig.1. This system includes a carriage, a boom, a cable, and a payload. Our motivation for this work is the payload will be always in microgravity when it contacts or collides with other objects.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 463 – 471, 2005.

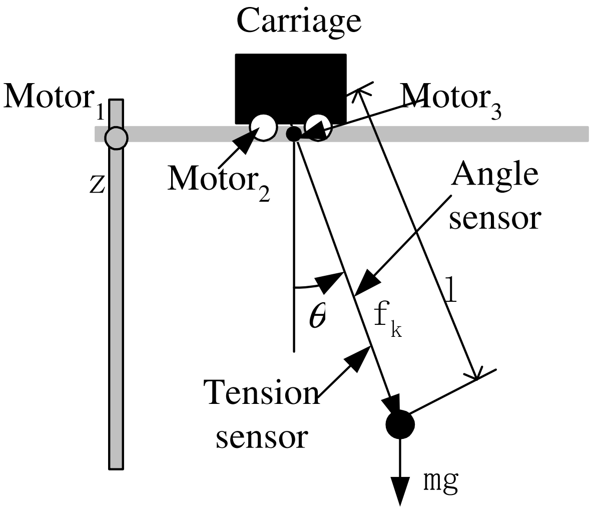


Fig. 1. The proposed active control system

This simulation system is a vertical microgravity control and mechanical motion control. Gravitational force acting on the payload is canceled by suspension cable following the motion of the carriage.

2 System Descriptions

The mechanical system of the microgravity simulation is shown as Fig.1. The system consists of a driven system and the suspension system. The driven system consists of a carriage with one degree of freedom and a boom with one degree of freedom. The carriage driven by motor2 can move along the boom. The carriage weighs 50kg. The boom can rotate around z axis. It is about 200cm long and has little kinetic friction. The driven system include two direct drive motors as actuators named as motor1 and motor2. The suspension system consists of a four-degree-of-freedom(rotational angles αβγ and z axis) cable which suspends the payload. When the payload collides with other objects, the suspension cable follows the trajectories of the payload motion, the motor1 and motor2 work as to keep the payload perpendicular to cancel gravitational force.

2.1 Sensor System

A tension sensor will be used to sense cable tension, and an angle sensor will sense the angle of the cable. There are three DC servo motors. The motor3 can be used to drive the cable, thereby adjusting the tension of cable. The DC servo motor2 can be used to drive the carriage and the DC servo motor1 can be used to drive the boom. The boom can rotate. The DC servo motor3 actively control the vertical direction of this cable with payload. The DC servo motor2 actively control the horizontal direction of this gravity compensation system. This motor2 drive the carriage so as the mobile can keep up with the bounced payload. The boom is rotational so the motor1 can drive it.

2.2 Modeling of System Dynamics

The system can be simplified to obtain approximations of the mathematical model[4]. In our paper, we only discuss the horizontal direction of this gravity compensation system. This simpler horizontal direction system consists of a single rigid cable which supports the masses of the payload as is shown in Fig.2.

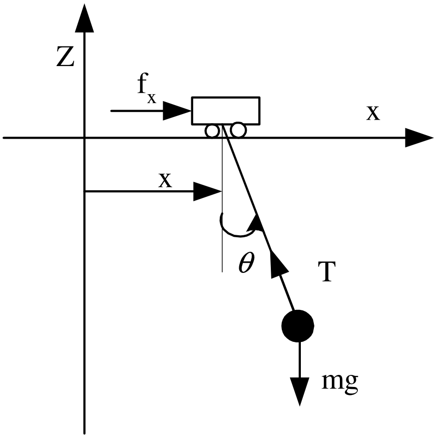


Fig. 2. Simplified system

This simpler system consists of a single pulley over which a rigid cable supports the payload's masses, the motor2 and a controller which can give some signal to the motor2. The motor2 can run based on the output of the controller. The carriage can freely run along the boom.

The following assumptions were made to reduce the model order to a more manageable degree: the mass of the cable is negligible compared to the mass of the payload, the friction of the carriage contributes insignificantly to the system, and the electrical time constant of the drive motor is considerably smaller than its physical time constant. Not considering these higher order effects will reduce the model order and make the controller design easier. Also, that the lengths of the cables change constantly with the payload moving vertically should be considered and the rotation of the boom will be taken in account when the payload is bounced. Controlling these effects are very difficult and very interesting and will be discussed in other paper.

Without the force colliding on the payload, the payload is under microgravity, the pulling force T is equaled to gravity G. When a force is collided on the payload, the payload will make some movement and make an angle with vertical direction. Its principle is shown as: the angle sensor dynamically senses the angle θ and gives this angleθto the controller. The output of this controller is based on some algorithm and it should be magnified so as to drive the motor2.

Let us see the collision of this system. At the first, the system is at rest and the three motors are in work which can shorten starts time of the motor. When the collision begins, the angle θ is given to the controller. The f_x in Fig.2 is the drive power of motor. The carriage can move based on f_x. If the f_x is large, the carriage will move faster and if the f_x is small then the carriage will slower.

The simplified system model is given as:

For carriage:

$$M\ \ddot{S}(t) = f_x \ - H(t) \tag{1}$$

For payload:

$$H(t) = m\frac{d^2}{dt^2}\left(S + l\sin\theta\right) = m\ddot{S} + ml\cos\theta\ddot{\theta} - ml\sin\vartheta\left(\dot{\theta}\right)^2 \tag{2}$$

$$v(t) - mg = m\frac{d^2}{dt^2}\left(l\cos\theta\right) = -ml\sin\theta\,\ddot{\theta} - ml\cos\theta\left(\dot{\theta}\right)^2 \tag{3}$$

$$\frac{H(t)}{V(t)} = \frac{\sin\theta}{\cos\theta} \tag{4}$$

Introducing the equation (2) to the equation (1):

$$\left(M + m\right)\ddot{S} + ml\cos\theta\,\ddot{\theta} - ml\sin\theta\left(\dot{\theta}\right)^2 = f_x \tag{5}$$

Introducing the equation (2) and (3) to (4):

$$\frac{\ddot{S} + l\cos\theta\,\ddot{\theta} - l\sin\theta\left(\dot{\theta}\right)^2}{g - l\sin\theta\,\ddot{\theta} - l\cos\theta\left(\dot{\theta}\right)^2} = \frac{\sin\theta}{\cos\theta} \tag{6}$$

Where:

$H(t)$ = the cable horizontal tension

$V(t)$ = the cable vertical tension

m = payload mass

M = carriage mass

$\ddot{S}(t)$ = acceleration of carriage

g =acceleration of gravity

l = the length of cable

θ = the angular position

Because the θ is too small, we can make an assumption that $\sin\theta(t)\approx\theta(t),\cos\theta(t)\approx 1$, then we get the linearization equation come from equation (5):

$$(M+m)\ddot{S}+ml\ddot{\theta}=f_x \tag{7}$$

$$\ddot{S}+l\ddot{\theta}-g\theta=0 \tag{8}$$

We assume that $x_1=S, x_2=\dot{S}, x_3=\theta, x_4=\dot{\theta}$

$$\dot{x}_2=\ddot{S}=-\frac{mg}{M}\theta+\frac{1}{M}f_x=-\frac{mg}{M}x_3+\frac{1}{M}f_x \tag{9}$$

$$v(t)-mg=m\frac{d^2}{dt^2}(l\cos\theta)=-ml\sin\theta\ddot{\theta}-ml\cos\theta\left(\dot{\theta}\right)^2 \tag{10}$$

Its state space equations are described as follows:

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\\ \dot{x}_3(t)\\ \dot{x}_4(t)\end{bmatrix}=\begin{bmatrix}0&1&0&0\\ 0&0&-\frac{m}{M}g&0\\ 0&0&0&1\\ 0&0&\frac{M+m}{Ml}g&0\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\\ x_3(t)\\ x_4(t)\end{bmatrix}+\begin{bmatrix}0\\ \frac{1}{M}\\ 0\\ -\frac{1}{Ml}\end{bmatrix}f_x \tag{11}$$

$$y=(1\quad 0\quad 0\quad 0)x(t) \tag{12}$$

Where $M=50Kg, m=20Kg, l=1.5m$

$$A=\begin{bmatrix}0&1&0&0\\ 0&0&-\frac{m}{M}g&0\\ 0&0&0&1\\ 0&0&\frac{M+m}{Ml}g&0\end{bmatrix} \tag{13}$$

$$B=\begin{bmatrix}0\\ \frac{1}{M}\\ 0\\ -\frac{1}{Ml}\end{bmatrix} \tag{14}$$

$$C=[1\quad 0\quad 0\quad 0] \tag{15}$$

Because $\text{rank}\left(B\quad AB\quad A^2B\quad A^3B\right)=4$, this system is controllable.

3 Fuzzy Cerebella Model Articulation Controller

The Cerebella Model Articulation Controller (CMAC) [5, 6] is a neural network that models the structure and function of the part of the brain known as the cerebellum. CMAC fundamentally learn to estimate the output by system states as an index to refer to a look-up table where the corresponding output derived from the sum of a set of synaptic weights are stored. Therefore, the relationship between input and output of an unknown system can be approximated by means of a CMAC if activated weights are being updated evolutionarily.

In order to improve the generalization property of CMAC and solve the memory questions for high dimension, FCMAC is proposed[7].It consists of five layers including input layer, fuzzy layer, fuzzy association layer, fuzzy post association layer and output layer [7].Fig.3 shows the structure of FCMAC.

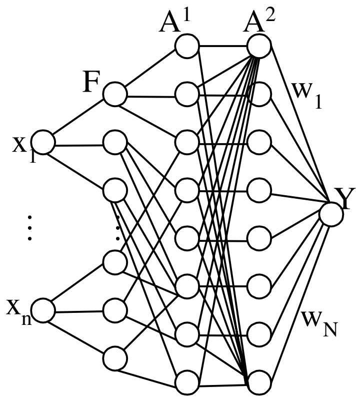


Fig. 3. The structure of FCMAC

The input vector $x=(x_1\cdots x_n)$ is transferred to the fuzzy layer. Each node at Fuzzy Layer corresponds to a linguistics variable, such as Positive Big, Positive Middle, Positive Small, Zero, Negative Middle, Negative Big. Fuzzy Association Layer links fuzzy layer and accomplishes the matching of fuzzy logic rule. Fuzzy Post association Layer will calculate the normalization of firing strength and prepare for fuzzy inference.

The output layer is:

$$y=w\cdot\Phi(x) \tag{16}$$

Where the parameter of the fuzzy CMAC weight is w , x is the input vector. $\Phi(x)$ is the Gaussian function.

In our paper, the output of FCMAC is one dimension variable that can drive the motor. So the learning algorithm is shown as follows:

$$E=\frac{1}{2}\left(y(t)-y_d(t)\right)^2 \tag{17}$$

$$\Delta w=\frac{\partial E}{\partial w}=\frac{\partial E}{\partial y}.\frac{\partial y}{\partial w}=\left(y(t)-y_d(t)\right).\Phi(x) \tag{18}$$

$$w(i+1)=w(i)+\eta.\Delta w(i) \tag{19}$$

Where $y(t)$ is the real output of FCMAC, $y_d(t)$ is the desired output of FCMAC, η is the learning rate. η is the most important parameter. Choosing appropriate η can improve the convergence of FCMAC. η is determined by adaptive GA .

4 Algorithms

The cable gravity compensation is very complex and nonlinear system, choosing PID controller is feasible but determining the parameters of PID is difficult. FCMAC can approximate nonlinear system. In our paper, the controller based FCMAC with PID is proposed. Miller et al.[5,6]successfully implemented CMAC for the control of an industrial manipulator. Fig.4 depicts the control scheme proposed by Miller et al. In our work, we use the FCMAC controller to replace the CMAC controller. It combined a conventional constant gain PID controller with a FCMAC controller. These two controllers work in parallel simultaneously. In the beginning, the PID controller is responsible for stabilization of the system. Later, The FCMAC learns and compensates for the nonlinearity. The procedure of learning cable gravity compensation based on FCMAC is as follows:

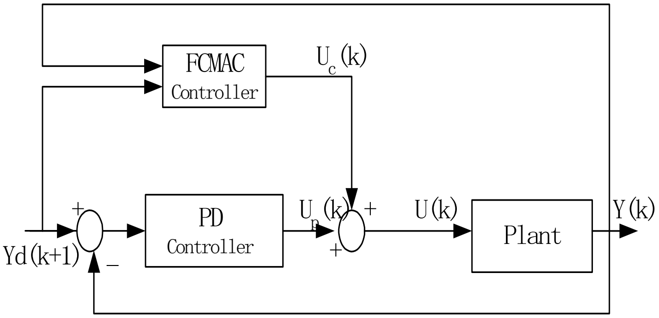


Fig. 4. Architecture of a FCMAC controller

Step 1. Initialize FCMAC neural network.

Step 2. Gather measured and desired value of process as training data.

Step 3. Run FCMAC neural network learning algorithm to achieve desired error.

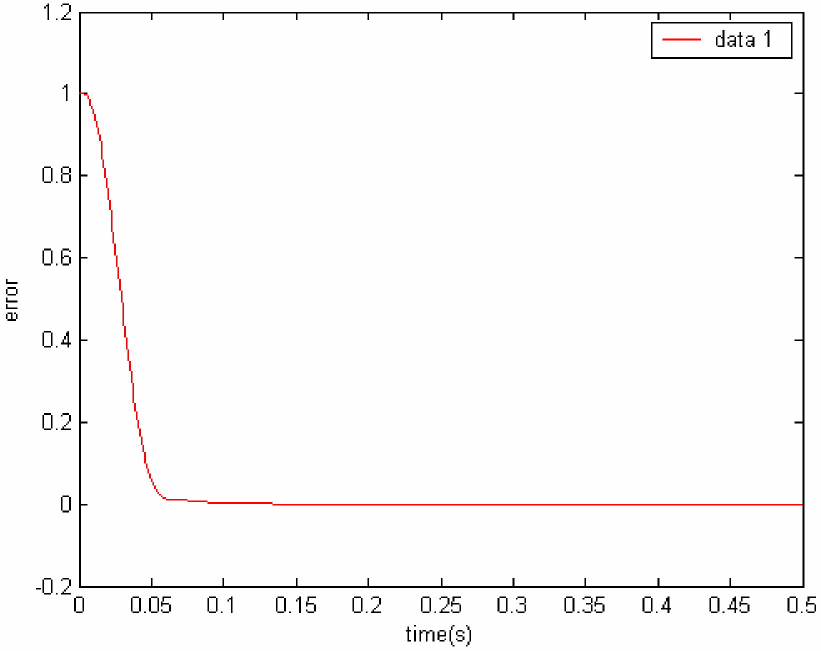


Fig. 5. The error based on FCMAC

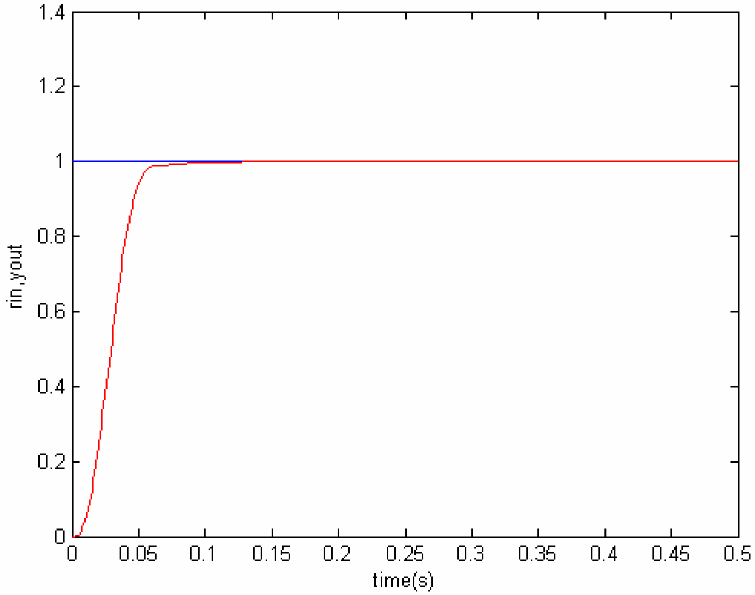


Fig. 6. The step signal response of simulation system

5 Simulation

In this section, we describe the simulation in matlab language. We make the model according to the analysis mentioned above. In this simulation, we generate all sensor information via theoretical mathematic models (this model is inaccurate). In the simulation, we select the step signal as reference command. Fig.5 shows the system error based on FCMAC and Fig.6 shows the track of step signal.

From Fig.5, this system based on FCMAC possesses great capabilities for fast learning.

6 Conclusions

According to these analyses, we can solve the problem of nonlinear system if we take FCMAC into the experiment. FCMAC can raise the control precision when the system model is nonlinear and complex, we can control the system according to angle position, speed, tension of the object. Taking the training data as control signal of servo system, the control system will get more stability.

This paper describes a nonlinear system offline, we should establish a nonlinear model online in order to control real time. We will be do much research on it.

Acknowledgment

This work was partly supported by Project 50275141of the National Natural Science Foundation of China.

References

1. White, G. C., Yangsheng, Xu.: An Active Vertical-direction Gravity Compensation System. IEEE Transactions on Instrumentation and Measurement, Vol.43, Issue 6, Dec.(1994)786 - 792
2. Huang, B. C., Ma, Y. L.: Space Environment Test Technology of Spacecraft. Defense Industry Press. 2002, 9 (in Chinese)
3. Liang, B., Chen, J. X., Liu, L. D., Li, C., Huang, X. L., Li, G.T.: The System of Extravehicular Mobile Robot. The Tenth Academic Annual Convention Thesis of a Space and The Sport Body Control Techniques in Country (2002)
4. Tong, M. D.: The Theories and Design of Linear System(M). The University of Science and Technology of China Press. (1998) 45-47
5. Miller, W. T., et al.: Application of a General Learning Algorithm to the Control of Robotics Manipulators. The International Journal of Robotics Reach, (1987)48-98
6. Miller, W. T., et al.: Real-time Dynamic Control of An Industrial Manipulator Using a Neural Network Based Learning Controllers. IEEE Trans. Robotics and Automation, (1990)1-9
7. Deng, Zhidong., Sun, Zengqi., Zhang, Zaixing.: A Fuzzy CMAC Neural Network. ACTA AUTOMATIC SINICA, Vol.21, (1995) 288-294

Oscillation and Strong Oscillation for Impulsive Neutral Parabolic Differential Systems with Delays

Yu-Tian Zhang and Qi Luo

College of Science, Wuhan University of Science and Technology,
430081, Wuhan, China
ytzhang81@163.com

Abstract. In respect that, in practical systems, we usually merely consider oscillation while strong oscillation is sometimes ignored which is also of wide applied background, this paper presents some results of the oscillation and strong oscillation of impulsive neutral parabolic differential systems with delays. Some criteria on the oscillation and strong oscillation are established by using analytical techniques. It is shown that, for impulsive parabolic differential systems with delays, although strong oscillation has more restriction than oscillation, the result of strong oscillation can be parallel to that of oscillation under certain conditions.

1 Introduction

As we all know, the term of impulsive control primitively came of evolution processes, which are characterized by the fact that at certain moments of time they experience a change of state abruptly because of short-term perturbations whose duration is negligible in comparison with the duration of the processes considered. Subsequently, researchers also found that, in fact, such impulsive control arises naturally in a wide variety of application, such as orbital transfer of satellite, ecosystems management, and control of money supply in a financial marker, pharmacokinetics and frequency modulated systems. Moreover, there are many cases where impulsive control can give better performance than continuous control. Sometimes even only impulsive control can be used for control purpose. For example, a central bank can not change its interest rate everyday in order to regulate the money supply in a financial market. Due to this important discovery, a great attention was paid to impulsive control problems, which are well described by impulsive differential systems. It is the reason for the development of the theory of impulsive ordinary differential systems and impulsive partial differential systems. In the past few years, the theory of impulsive ordinary differential systems [see 1, 2] and impulsive partial differential systems[see 3-11] has been elaborated to a considerable extent. Recently, the oscillation for impulsive delay parabolic differential systems [see 8] and the oscillation for impulsive hyperbolic differential systems with several delays [see 4] were studied, respectively. However, up to now, nobody studied the oscillation and strong oscillation of systems of impulsive neutral parabolic differential systems with delays, as far as we know.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 472–481, 2005.

In this paper, we study the oscillation and strong oscillation of impulsive neutral parabolic differential systems with delays. Some criteria on the oscillation and strong oscillation are established by using analytical techniques. It is shown that, for impulsive neutral parabolic differential systems with delays, the oscillation will be parallel to the strong oscillation under certain conditions.

2 Problem Statements and Preliminaries

In this paper, we consider the following impulsive neutral parabolic differential systems with delays

$$\begin{aligned} &\tfrac{\partial}{\partial t}[u_i(x,t)+z_i(t)u_i(x,t-\theta)] \\ &=\textstyle\sum_{k=1}^{m} a_{ik}(t)\varDelta u_k(x,t)+\sum_{k=1}^{m} b_{ik}\varDelta u_k(x,t-\tau)- \\ &q_i(x,t)u_i(x,t-\lambda)-c_i(x,t,(u_k(x,t))_{k=1}^{m},(u_k(x,t-\sigma)_{k=1}^{m}))\quad t\neq t_j \qquad (1) \\ &u_i(x,t_j^{+})-u_i(x,t_j^{-})=p_i(x,t_j,u_i(x,t_j)) \\ &i\in I_m,\quad j\in I_\infty,\quad (x,t)\in\varOmega\times R_+\equiv G \end{aligned}$$

where $I_m=(1,2,\cdots,m)$, $I_\infty=(1,2,\cdots)$, $R_+=[0,\infty)$, $\varOmega$ is a bounded domain in R^n with a smooth boundary $\partial\varOmega$, $\varDelta u_i(x,t)=\sum_{r=1}^{n}\frac{\partial^2 u_i(x,t)}{\partial x_r^2}$, $i\in I_m$, $u_i(x,t_j^{+})-u_i(x,t_j)=p_i(x,t_j,u_i(x,t_j))$ are the impulses at moments t_j and $0<t_1<t_j<\cdots$ is a strictly increasing sequence such that $\lim\limits_{j\to\infty}t_j=\infty$.

Consider the following boundary condition:

$$\frac{\partial u_i(x,t)}{\partial N}=0,\quad (x,t)\in\partial\varOmega\times R_+,\quad t\neq t_j,\quad i\in I_m,\quad j\in I_\infty \qquad (2)$$

and the initial condition

$$u_i(x,t)=\phi_i(x,t),\quad (x,t)\in\varOmega\times[-\delta,0], \qquad (3)$$

where N is the unit exterior normal vector to $\partial\varOmega$, and

$$\delta=\max(\theta,\tau,\sigma,\lambda),\quad \phi_i\in C^2(\varOmega\times[-\delta,0],R),\quad i\in I_m.$$

Throughout this paper, we assume that the following conditions hold:

(C_1) $a_{ik},\ b_{ik},\ z_i\in PC[R_+,R_+]$, $i,\ k\in I_m$, where PC denotes the class of functions, which are piecewise continuous in t with discontinuous of first kind only at $t=t_j$ and left continuous at $t\neq t_j, j\in I_\infty$,

(C_2) $\theta\geq 0,\ \tau\geq 0,\ \sigma\geq 0$, and $\lambda\geq 0$ are constants,

(C_3) $c_i\in PC[\bar{G}\times R^{2m},R]$ and

$$c_i(x,t,\xi_1,\cdots,\xi_i,\cdots,\xi_m,\eta_1,\cdots,\eta_i,\cdots,\eta_m)\begin{cases}\geq 0 & \text{if } \xi_i \text{ and } \eta_i\in(0,\infty)\\ \leq 0 & \text{if } \xi_i \text{ and } \eta_i\in(-\infty,0)\end{cases}$$

$$c_i(x,t,\xi_1,\cdots,-\xi_i,\cdots,\xi_m,\eta_1,\cdots,-\eta_i,\cdots,\eta_m)$$

$$=-c_i(x,t,\xi_1,\cdots,\xi_i,\cdots,\xi_m,\eta_1,\cdots,\eta_i,\cdots,\eta_m) \qquad i\in I_m$$

(C_4) $q_i\in PC[\bar{G},R_+],\ \ q_i(t)=\min\limits_{x\in\Omega}q_i(x,t),\ i\in I_m$

(C_5) $p_i\in\bar{G}\times R\to R,\ i\in I_m,\ j\in I_\infty$

(C_6) for any function $u_i\in PC[\bar{G},R_+]$, the following conditions are satisfied:

$$p_i(x,t_j,-u_i(x,t_j))=-p_i(x,t_j,u_i(x,t_j))$$

and

$$\int_\Omega p_i(x,t_j,u_i(x,t_j))dx=\alpha_{ij}\int_\Omega u_i(x,t_j)dx$$

where $\alpha_{ij}>0$ is a constant, $i\in I_m,\ \ j\in I_\infty$.

For convenience, we shall introduce the following notation:

$$U_i(t)=\int_\Omega u_i(x,t)dx,\quad t\in R_+,\ \ i\in I_m.$$

3 Some Definitions

In this section, Let us state some definitions for later use.

Definition 3.1. The vector function $u(x,t)=(u_1(x,t),\cdots,u_m(x,t))^T$ is said to be a solution of problem (1) and (2) if the following conditions are satisfied:

(1) $u_i(x,t)$ is a first differentiable function for $t,\ t\neq t_j,\ i\in I_m,\ j\in I_\infty$,

(2) $u_i(x,t)$ is a piecewise continuous function with points of discontinuity of the first kind at $t=t_j,\ j\in I_\infty$, and at the moments of impulse the following relations are satisfied:

$$u_i(x,t_j^-)=u_i(x,t_j),\ \ u_i(x,t_j^+)=u_i(x,t_j)+p_i(x,t_j,u_i(x,t_j)),\ \ i\in I_m,\ \ j\in I_\infty,$$

(3) $u_i(x,t)$ is a second-order differentiable function for $x,\ i\in I_m$,

(4) $u_i(x,t)$ satisfies (1) in the domain G and boundary condition (2), $i\in I_m$.

Definition 3.2. A nontrivial component $u_i(x,t)$ of the vector function $u(x,t)=(u_1(x,t),\cdots,u_m(x,t))^T$ is said to oscillate in $\Omega\times[\mu_0,\infty)$ if for each $\mu>\mu_0$ there is a point $(x_0,t_0)\in\Omega\times[\mu,\infty)$ such that $u_i(x_0,t_0)=0$.

Definition 3.3. The vector solution $u(x,t)=(u_1(x,t),\cdots,u_m(x,t))^T$ of problem (1) and (2) is said to oscillate in the domain $G=\Omega\times R$, if at least one of its nontrivial components oscillates in G. Otherwise, the vector solution $u(x,t)$ is said to be nonoscillatory in G.

Definition 3.4. The vector solution $u(x,t)=(u_1(x,t),\cdots,u_m(x,t))^T$ of problem (1) and (2) is said to strongly oscillate in the domain $G=\Omega\times R_+$, if each of its nontrivial components oscillates in G.

4 Some Lemmas

We shall introduce, in this section, some useful lemmas which play an important role in the proof of our conclusions.

Consider the following linear impulsive neutral differential systems with delays

$$(y(t)+p(t)y(t-\tau))'+q(t)y(t-\sigma)=0 \quad t\geq 0,\ t\neq t_k,\ k\in N$$

$$y(t_k^+-y(t_k^-))=b_k y(t_k) \quad k\in N,$$

where $0<t_1<t_k<\cdots$ is a strictly increasing sequence such that $\lim_{j\to\infty} t_k=\infty$, $p\in([0,\infty),R), q\in([0,\infty),R)$ are locally integratable functions, b_k (constant) satisfies $b_k>-1$, $k\in N$. As usual in the theory of impulsive differential systems, at the points of discontinuity t_k we assume that $y(t_(k))=y(t_k^-)$.

Lemma 4.1[12] Assume that

(1) $q(t)\geq 0$,

(2) $-1\leq \prod\limits_{t-\tau\leq t_k<t}(1+b_k)^{-1}p(t)\leq 0$,

(3) $\lim\limits_{t\to\infty}\inf\int_{t-\sigma}^{t}\prod\limits_{s-\sigma\leq t_k<s}(1+b_k)^{-1}q(s)ds>\frac{1}{e}$.

Then every solution of the following inequality

$$(y(t)+p(t)y(t-\tau))'+q(t)y(t-\sigma)\leq 0 \quad t\geq 0,\ t\neq t_k,\ k\in N$$

$$y(t_k^+-y(t_k))=b_k y(t_k) \quad k\in N$$

is oscillatory.

Lemma 4.2[12] Assume that

(1) $q(t)\geq 0$,

(2) $\prod\limits_{t-\tau\leq t_k<t}(1+b_k)^{-1}p(t)\leq -1$,

(3) $\lim\limits_{t\to\infty}\inf\int_{t}^{t+\tau-\sigma}\prod\limits_{s\leq t_k<s+\tau-\sigma}(1+b_k)\frac{q(s)}{-p(s+\tau-\sigma)}ds>\frac{1}{e}$.

Then every solution of the following inequality

$$(y(t)+p(t)y(t-\tau))'+q(t)y(t-\sigma)\leq 0 \quad t\geq 0,\ t\neq t_k,\ k\in N$$

$$y(t_k^+-y(t_k))=b_k y(t_k) \quad k\in N$$

is oscillatory.

Lemma 4.3[12] Assume that

(1) $\prod\limits_{t-\tau\leq t_k<t}(1+b_k)^{-1}p(t)\equiv\bar{p}$ is a positive constant,

(2) $\prod\limits_{t-\sigma\leq t_k<t}(1+b_k)^{-1}q(t)$ is a positive periodic function with respect to τ,

(3) $\lim\limits_{t\to\infty}\inf\int_{t-(\sigma-\tau)}^{t}\prod\limits_{s-\sigma\leq t_k<s}(1+b_k)^{-1}q(s)ds>\frac{1+\bar{p}}{e}$.

Then all solutions of the following inequality

$$(y(t)+p(t)y(t-\tau))'+q(t)y(t-\sigma)\leq 0 \;\; t\geq 0,\; t\neq t_k,\; k\in N$$

$$y(t_k^+-y(t_k))=b_k y(t_k) \;\; k\in N$$

is oscillatory.

5 Main Results

Based on the preparations made in section 3 and section 4, we shall obtain our main results in this section.

5.1 Oscillation

Firstly , we shall establish three theorems which provide sufficient conditions for oscillation of the solutions of problem (1) and (2).

Theorem 5.1.1. Assume that
(1)$-1\leq \prod\limits_{t-\theta\leq t_k<t}(1+\alpha_{ij})^{-1}k_i(t)\leq 0$,
(2) $\lim\limits_{t\to\infty}\inf\int_{t-\lambda}^{t}\prod\limits_{s-\lambda\leq t_k<s}(1+\alpha_{ij})^{-1}q_i(s)ds>\frac{1}{e}$.
Then every solution of problem (1) and (2) is oscillatory in G.

Proof. Assume to the contrary that $u(x,t)$ is a solution of problem (1) and (2) and $u(x,t)$ is nonoscillatory in G. Then, each of its nontrivial components is nonoscillatory in G. Hence, there must exists some $i_0\in I_m$ such that $u_{i_0}(x,t)$ is nonoscillatory in G. Alternatively, we suppose $u_{i_0}(x,t)$ is eventually positive, which implies there is a number $T>0$ such that $u_{i_0}(x,t)>0,\; u_{i_0}(x,t-\theta)>0,\; u_{i_0}(x,t-\sigma)>0,\; u_{i_0}(x,t-\lambda)>0$ in $\Omega\times[T,\infty)$.

Clearly, $u_{i_0}(x,t)$ satisfies the following equation:

$$\begin{aligned}
&\tfrac{\partial}{\partial t}[u_{i_0}(x,t)+z_{i_0}(t)u_{i_0}(x,t-\theta)]\\
=&\sum_{k=1}^{m}a_{i_0k}(t)\Delta u_k(x,t)+\sum_{k=1}^{m}b_{i_0k}\Delta u_k(x,t-\tau)-\\
&q_{i_0}(x,t)u_{i_0}(x,t-\lambda)-c_{i_0}(x,t,(u_k(x,t))_{k=1}^{m},(u_k(x,t-\sigma)_{k=1}^{m}))\;\; t\neq t_j\\
&u_{i_0}(x,t_j^+)-u_{i_0}(x,t_j)=p_{i_0}(x,t_j,u_{i_0}(x,t_j))\\
&j\in I_\infty,\;\; (x,t)\in\Omega\times R_+\equiv G
\end{aligned}\tag{4}$$

Case1: $t\neq t_j$. Integrating the first equation in (4) with respect to x over the domain Ω, we have

$$\begin{aligned}&\tfrac{d}{dt}[\textstyle\int_\Omega u_{i_0}(x,t)dx + z_{i_0}(t)\int_\Omega u_{i_0}(x,t-\theta)dx]\\ &= \sum_{k=1}^{m} a_{i_0k}(t)\textstyle\int_\Omega \Delta u_k(x,t)dx + \sum_{k=1}^{m} b_{i_0k}\int_\Omega \Delta u_k(x,t-\tau)dx-\\ &\textstyle\int_\Omega q_{i_0}(x,t)u_{i_0}(x,t-\lambda)dx - \int_\Omega c_{i_0}(x,t,(u_k(x,t))_{k=1}^m,(u_k(x,t-\sigma)_{k=1}^m))dx\\ &t \geq T,\ t \neq t_j,\ j \in I_\infty.\end{aligned} \tag{5}$$

Applying Green's formula and boundary condition (2), we have

$$\int_\Omega \Delta u_k(x,t)dx = \int_{\partial\Omega} \frac{\partial u_k(x,t)}{\partial N}dS = 0 \tag{6}$$

and

$$\int_\Omega \Delta u_k(x,t-\tau)dx = \int_{\partial\Omega} \frac{\partial u_k(x,t-\tau)}{\partial N}dS = 0 \quad t > T,\ t \neq t_j,\ j \in I_\infty. \tag{7}$$

Noting that $u_{i_0}(x,t) > 0,\ u_{i_0}(x,t-\sigma) > 0$, from assumption (C_3) , it is not difficult to show that

$$c_{i_0}(x,t,(u_k(x,t))_{k=1}^m,(u_k(x,t-\sigma)_{k=1}^m)) \geq 0, \quad t \geq T,\ t \neq t_j,\ j \in I_\infty. \tag{8}$$

Therefore, combining (5)-(8) and using assumption (C_4), we have

$$[U_{i_0}(t) + z_{i_0}(t)U_{i_0}(t-\theta)]' + q_{i_0}(t)U_{i_0}(t-\lambda) \leq 0$$

$$t \geq T,\ t \neq t_j,\ j \in I_\infty$$

Case2: $t = t_j$. It follows from the second equation in (4) and assumption (C_6) that

$$\begin{aligned}U_{i_0}(t_j^+) &= \textstyle\int_\Omega u_{i_0}(x,t_j^+)dx\\ &= \textstyle\int_\Omega u_{i_0}(x,t_j)dx + \int_\Omega p_{i_0}(x,t_j,u_{i_0}(x,t_j))dx\\ &= \textstyle\int_\Omega u_{i_0}(x,t_j)dx + \alpha_{i_0j}\int_\Omega u_{i_0}(x,t_j)dx\\ &= (1+\alpha_{i_0j})U_{i_0}(t_j) \qquad t = t_j,\ j \in I_\infty.\end{aligned} \tag{9}$$

Thus,

$$\begin{aligned}&(U_{i_0}(t) + z_{i_0}(t)U_{i_0}(t-\theta))' + q_{i_0}(t)U_{i_0}(t-\lambda) \leq 0 \quad t \geq T,\ t \neq t_j,\ j \in I_\infty\\ &U_{i_0}(t_j^+) = (1+\alpha_{i_0j})U_{i_0}(t_j) \quad j \in I_\infty,\end{aligned} \tag{10}$$

where $U_{i_0}(t) = \int_\Omega u_{i_0}(x,t)dx$ is eventually positive, which contradicts with Lemma 4.1.

If $u_{i_0}(x,t)$ is eventually negative, then $-u_{i_0}(x,t)$ is eventually positive. Furthermore, it is obvious that $u_{i_0}(x,t)$ satisfies

$$\begin{aligned}
&\tfrac{\partial}{\partial t}[u_{i_0}(x,t)+z_{i_0}(t)u_{i_0}(x,t-\theta)]\\
&=\sum_{k=1}^{m} a_{i_0k}(t)\mathit{\Delta} u_k(x,t)+\sum_{k=1}^{m} b_{i_0k}\mathit{\Delta} u_k(x,t-\tau)-\\
&q_{i_0}(x,t)u_{i_0}(x,t-\lambda)-c_{i_0}(x,t,(u_k(x,t))_{k=1}^m,(u_k(x,t-\sigma)_{k=1}^m))\quad t\neq t_j\\
&u_{i_0}(x,t_j^+)-u_{i_0}(x,t_j)=p_{i_0}(x,t_j,u_{i_0}(x,t_j))\\
&j\in I_\infty,\quad (x,t)\in \mathit{\Omega}\times R_+\equiv G
\end{aligned}$$

that is

$$\begin{aligned}
&\tfrac{\partial}{\partial t}[-u_{i_0}(x,t)+z_{i_0}(t)(-u_{i_0}(x,t-\theta))]\\
&=-\sum_{k=1}^{m} a_{i_0k}(t)\mathit{\Delta} u_k(x,t)-\sum_{k=1}^{m} b_{i_0k}\mathit{\Delta} u_k(x,t-\tau)-\\
&q_{i_0}(x,t)(-u_{i_0}(x,t-\lambda))+c_{i_0}(x,t,(u_k(x,t))_{k=1}^m,(u_k(x,t-\sigma)_{k=1}^m))\quad t\neq t_j\\
&-u_{i_0}(x,t_j^+)-(-u_{i_0}(x,t_j))=p_{i_0}(x,t_j,-u_{i_0}(x,t_j))\\
&j\in I_\infty,\quad (x,t)\in \mathit{\Omega}\times R_+\equiv G
\end{aligned}$$

The remainder is similar to the case $u_{i_0}(x,t)$ is eventually positive. we can still arrive at the desirable contradiction. This completes the proof.

Analogously, According to Lemma 4.2, it is easy to establish the following oscillation result of problem (1) and (2).

Theorem 5.1.2. Assume that
(1) $\prod\limits_{t-\theta\le t_k<t}(1+\alpha_{ij})^{-1}k_i(t)\le -1$,
(2) $\lim\limits_{t\to\infty}\inf\int_t^{t+\theta-\lambda}\prod\limits_{s\le t_k<s+\tau-\lambda}(1+\alpha_{ij})\frac{q_i(s)}{-k_i(s+\theta-\lambda)}ds>\frac{1}{e}$.
Then all solutions of problem (1) and (2) are oscillatory in the domain G.

Proof. We still adopt reduction to absurdity. Making use of the method which is same as in Theorem 5.1.1, it is easy to derive the following conclusion:

The system of the following form

$$\begin{aligned}
&(F_i(t)+z_i(t)F_i(t-\theta))'+q_i(t)F_i(t-\lambda)\le 0\quad t\ge T,\ t\neq t_j,\ j\in I_\infty\\
&F_i(t_j^+)=(1+\alpha_{ij})F_i(t_j)\quad j\in I_\infty
\end{aligned}\tag{11}$$

will have a oscillatory solution, which suggests the solution is either eventually positive or eventually negative. Obviously, This contradicts with Lemma 4.2. The proof is complete.

Similarly, according to Lemma 4.3, we can also obtain the following oscillation result of problem (1) and (2).

Theorem 5.1.3. Assume that

(1) $\prod\limits_{t-\theta\le t_k<t}(1+\alpha_{ij})^{-1}k_i(t)\equiv\bar{p_i}$ is a positive constant,

(2) $\prod\limits_{t-\lambda\le t_k<t}(1+\alpha_{ij})^{-1}q_i(t)$ is a positive periodic function with respect to θ,

(3) $\lim\limits_{t\to\infty}\inf\int_{t-(\lambda-\theta)}^{t}\prod\limits_{s-\lambda\le t_k<s}(1+\alpha_{ij})^{-1}q_i(s)ds>\frac{1+\bar{p_i}}{e}$.

Then every solution of problem (1) and (2) is oscillatory in the domain G.

Proof. The reduction to absurdity is yet adopted here. By the method in the proof of Theorem 5.1.1, it follows that the system of the following form

$$\begin{aligned}&(V_i(t)+z_i(t)V_i(t-\theta))'+q_i(t)V_i(t-\lambda)\le 0\ \ t\ge T,\ t\ne t_j,\ j\in I_\infty\\&V_i(t_j^+)=(1+\alpha_{ij})V_i(t_j)\ \ j\in I_\infty\end{aligned}\tag{12}$$

must have a oscillatory solution, which means the solution is either eventually positive or eventually negative. This contradicts with Lemma 4.3. The proof is complete.

5.2 Strong Oscillation

Nextly, in view of the definition of strong oscillation in section 4, we also establish three theorems which provide sufficient conditions for strong oscillation of problem (1) and (2).

By Lemma 4.1, it is easy to see the following strong oscillation result of problem (1) and (2) is parallel to Theorem 5.1.1

Theorem 5.2.1. Assume that

(1) $-1\le\prod\limits_{t-\theta\le t_k<t}(1+\alpha_{ij})^{-1}k_i(t)\le 0$,

(2) $\lim\limits_{t\to\infty}\inf\int_{t-\lambda}^{t}\prod\limits_{s-\lambda\le t_k<s}(1+\alpha_{ij})^{-1}q_i(s)ds>\frac{1}{e}$. Then every solution of problem (1) and (2) strongly oscillates in G.

Proof. Suppose that there exists a solution $u(x,t)$ of problem (1) and (2) which does not strongly oscillate in the domain G. In view of the definition of strong oscillation, we know that at least there is one of its nontrivial components is nonoscillatory. For convenience, we let $u_{i_0}(x,t)$ denote this nontrivial component. Through the proof of Theorem 5.1.1, it is easy to see that, in fact, $u_{i_0}(x,t)$ will satisfy the following system

$$\begin{aligned}&(Q_i(t)+z_i(t)Q_i(t-\theta))'+q_i(t)Q_i(t-\lambda)\le 0\ \ t\ge T,\ t\ne t_j,\ j\in I_\infty\\&Q_i(t_j^+)=(1+\alpha_{ij})Q_i(t_j)\ \ j\in I_\infty\end{aligned}\tag{13}$$

However, According to Theorem 5.1.1, we also know the above system has no nonoscillatory solutions, which contradicts with our assumption. This completes the proof.

Furthermore, from Lemma 4.2, we can achieve the following strong oscillation result of problem (1) and (2) which is parallel to Theorem 5.1.2

Theorem 5.2.2. Assume that

(1) $\prod\limits_{t-\theta\leq t_k<t}(1+\alpha_{ij})^{-1}k_i(t)\leq -1$,

(2) $\lim\limits_{t\to\infty}\inf\int_t^{t+\theta-\lambda}\prod\limits_{s\leq t_k<s+\tau-\lambda}(1+\alpha_{ij})\frac{q_i(s)}{-k_i(s+\theta-\lambda)}ds>\frac{1}{e}$.

Then every solution of problem (1) and (2) strongly oscillates in the domain G.

Proof. Similarly, we assume that there exists a solution $u(x,t)$ of problem (1) and (2) which does not strongly oscillate in the domain G. Therefore, at least there exists one of its nontrivial components is nonoscillatory. For convenience, we let $u_{i_0}(x,t)$ denote this nontrivial component. By the proof of Theorem 5.1.2, we can see that, in fact, $u_{i_0}(x,t)$ will satisfy the system of the form.

$$\begin{array}{l}(H_i(t)+z_i(t)H_i(t-\theta))'+q_i(t)H_i(t-\lambda)\leq 0\ \ t\geq T,\ t\neq t_j,\ j\in I_\infty\\ H_i(t_j^+)=(1+\alpha_{ij})H_i(t_j)\ \ j\in I_\infty\end{array}$$

But, by Theorem 5.1.2, it follows that system (13) has no nonoscillatory solutions, which contradicts with our assumption. The proof is complete.

According to Lemma 4.3, we obtain the following strong oscillation result of problem (1) and (2) at last, which is parallel to Theorem 5.1.3.

Theorem 5.2.3. Assume that

(1) $\prod\limits_{t-\theta\leq t_k<t}(1+\alpha_{ij})^{-1}k_i(t)\equiv\bar{p_i}$ is a positive constant,

(2) $\prod\limits_{t-\lambda\leq t_k<t}(1+\alpha_{ij})^{-1}q_i(t)$ is a positive periodic function as to θ,

(3) $\lim\limits_{t\to\infty}\inf\int_{t-(\lambda-\theta)}^{t}\prod\limits_{s-\lambda\leq t_k<s}(1+\alpha_{ij})^{-1}q_i(s)ds>\frac{1+\bar{p_i}}{e}$.

Then every solution of problem (1) and (2) strongly oscillates in G.

Proof. Suppose that there exists a solution $u(x,t)$ of problem (1) and (2) which does not strongly oscillate in the domain G. Hence, there must exist one of its nontrivial components which is nonoscillatory. For convenience, we let $u_{i_0}(x,t)$ denote this nontrivial component. By the proof of Theorem 5.1.3, it is not difficult to see that $u_{i_0}(x,t)$ satisfies the following system.

$$\begin{array}{l}(S_i(t)+z_i(t)S_i(t-\theta))'+q_i(t)S_i(t-\lambda)\leq 0\ \ t\geq T,\ t\neq t_j,\ j\in I_\infty\\ S_i(t_j^+)=(1+\alpha_{ij})S_i(t_j)\ \ j\in I_\infty\end{array}$$

While, through Theorem 5.1.2, we know system (13) has no nonoscillatory solutions, which contradicts with our assumption. This completes the proof.

6 Conclusions

In this paper, we use some analytical techniques to study the oscillation and the strong oscillation for impulsive neutral parabolic differential systems with several delays. A set of criteria are given for the oscillation and the strong oscillation of impulsive neutral parabolic differential systems with several delays. Moreover, if adding appropriate modification to the conditions of the above theorems, we can also achieve some useful corollaries.

References

1. Lakshmikantham, V., Bainov, D.D., Simeonov, P.S.: Theory of Impulsive Differential Equations. World Scientific, Singapore (1989)
2. Bainov, D.D., Simeonov, P.S.: Systems with Impulsive Effect: Stability, Theory and Applications. Wiley, New York (1989)
3. Bainov, D.D., Kamont, Z., Minchev, E.: Monotone Iterative Methods for Impulsive Hyperbolic Differential Functional Equations. J. Comput. Appl. Math. **70** (1996) 329–347
4. Cui, B.T., Liu, Y., Deng, F.: Some Oscillation Problems for Impulsive Hyperbolic Differential Systems with Several Delays. Appl. Math. Comput. **146** (2003) 667–679
5. Bainov, D.D., Minchev, E.: Estimates of Solutions of Impulsive Parabolic Equations and Applications to The Population Dynamics. Publ. Math. **40** (1996) 85–94
6. Erbe, L.H., Freedman, H.I., Liu, X., Wu, J.H.: Comparison Principle for Impulsive Parabolic Equations with Applications to Models of Single Species Growth. J. Austral. Math. Soc. Ser. **B32** (1991) 382–400
7. Fu, X., Liu, X., Sivaloganathan, S.: Oscillation Criteria for Impulsive Parabolic Systems. Appl. Anal. **79** (2001) 239–255
8. Fu, X., Liu, X., Sivaloganathan, S.: Oscillation Criteria for Impulsive Parabolic Equations with Delays. J. Math. Anal. Appl. **268** (2002) 647–664
9. Gao, W., Wang, J.: Estimates of Solutions of Impulsive Parabolic Equations under Neumann Boundary Condition. J.Math. Anal. Appl. **283** (2003) 478–490
10. Luo, J.: Oscillation of Hyperbolic Partial Differential Equations with Impulses. Appl. Math. Comput. **133** (2002) 309–318
11. Zhang, L.: Oscillation Criteria for Hyperbolic Partial Differential Equations with Fixed Moments of Impulse Effects. Acta. Math. Sinica. **43** (2000) 17–26
12. Yan, J.R.: The Oscillation of Impulsive Neutral Differential Equations with Delays. Chinese Annals of Mathematics. **25(2)** (2002) 95–98

Blind Estimation of Fast Time-Varying Multi-antenna Channels Based on Sequential Monte Carlo Method

Mingyan Jiang and Dongfeng Yuan

School of Information Science and Engineering, Shandong University,
Jinan 250100, China
State Key Lab. on Mobile Communications, Southeast University,
Nanjing 210096, China
{jiangmingyan, dfyuan}@sdu.edu.cn

Abstract. In this paper Monte Carlo Method (MCM) is used for tracking slow or fast fading in one antenna communication channel and in multi-antenna channels with space-time block coding (STBC), we compare it with Kalman filter tracking method, discuss its tracking ability when the system has carrier frequency offset, Simulation shows that MCM can be used as a blind method for channel tracking, many research hot spots of MCM are given in the end.

1 Introduction

In the applications of communication channel estimation, the non-linearity of the system and non-Gaussian noise become the main factor that baffles us to obtain accurate estimation. In order to get the sub-optimal estimation, we can use extended Kalman filter (EKF) method, but the EKF method is only suitable for the situation that the filtering error and predicting error are both small, the initial estimation covariance declining too rapidly will result in tracking instability, even algorithm divergence and final tracking failure. The sequential importance sampling (SIS) MCM (Particle Filter)[1] can be used to realize recursive Bayesian filter, this method is applicable for any non-linear system expressed by state-space model and traditional Kalman filter, the precision of the tracking estimation can approach the optimum[1]. Actually the sequential MCM principle is that adopting iteration operation of random variable to approach the unknown distribution, such as Minimum Mean Square Error (MMSE) and Maximum A Posterior (MAP) estimation[2]. Recently, due to the continuous improvement of computer's operation speed, the scheme based on the MCM described above has drawn great attention and becomes a new research hot spot, successful applications have been made in many fields: target tracking, navigation, digital communication channel blind deconvolution[3], joint channel estimation and survey in Rayleigh channel[4], time-varying spectrum estimation[5], the estimation of sequential signals in unknown models[6].

[1] This work is supported by the National Natural Science Foundation of China (No. 60372030) and the Key Project of Provincial Scientific Foundation of Shandong (No.Z2003G02).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 482 – 491, 2005.

2 Sequential Monte Carlo Method

Consider the non-linear/non-Gaussian random state-space model, given by:

$$\mathbf{x}_k = f(\mathbf{x}_{k-1}) + \mathbf{v}_{k-1} \ . \tag{1}$$

$$\mathbf{z}_k = h(\mathbf{x}_k) + \mathbf{n}_k \ . \tag{2}$$

where $\mathbf{x}_k \in \mathbf{R}^n$ is the state sequence at time *k*, $\mathbf{z}_k \in \mathbf{R}^p$ is the observation of the state $\mathbf{x}_k$, $k \in N$ is the set of natural numbers, $\mathbf{v}_{k-1} \in \mathbf{R}^n$, $\mathbf{n}_k \in \mathbf{R}^p$ are independently identically distributed (i.i.d) process noise sequence and measurement noise sequence, and $f : \mathbf{R}^n \to \mathbf{R}^n$, $h : \mathbf{R}^p \to \mathbf{R}^p$ are non-linear mappings that have boundaries respectively. Assume the initial probability density function (pdf) $p(\mathbf{x}_0 \mid \mathbf{z}_0) \equiv p(\mathbf{x}_0)$, the tracking problem is to recursively estimate the posterior pdf $p(\mathbf{x}_k \mid \mathbf{z}_{1:k})$ of the system state by the measurements with two stages: prediction and update. Suppose that the required pdf $p(\mathbf{x}_{k-1} \mid \mathbf{z}_{1:k-1})$ at time *k-1* is available, the prior pdf of the state at time *k* can be obtained via the Chapman-Kolmogorov equation:

$$p(\mathbf{x}_k \mid \mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} \mid \mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1} \ . \tag{3}$$

Eq.(3) use the condition that $p(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{z}_{1:k-1}) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1})$, the one order Markov process described in Eq.(1). At the time *k*, by Bayes' rule:

$$p(\mathbf{x}_k \mid \mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k \mid \mathbf{z}_{1:k-1})} \ . \tag{4}$$

where the normalizing constant $p(\mathbf{z}_k \mid \mathbf{z}_{1:k-1}) = \int p(\mathbf{z}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{z}_{1:k-1}) d\mathbf{x}_k$ is determined by Eq.(2) and by the likelihood function $p(\mathbf{z}_k \mid \mathbf{z}_{1:k-1})$ with the known statistics of noise $\mathbf{n}_k$.

2.1 Particle Filter

The key idea of MCM is to approach the actual posterior density function by a set of random samples with associated weights and then estimate it based on these samples and weights. If the number of samples is enough large, this estimation will approximate the posterior pdf. Let $\{\mathbf{x}_{0:k}, w_k^i\}_{i=1}^{N_s}$ denote random particles that characterize the posterior pdf $p(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k})$, where $\{\mathbf{x}_{0:k}^i, i = 0, \ldots, N_s\}$ is a set of states with associated weights $\{w_k^i, i = 1, \ldots, N_s\}$ and $\mathbf{x}_{0:k} = \{\mathbf{x}_j, j = 0, \ldots, k\}$ is the set of all states up to time *k*. The weights are normalized as $\sum_i w_k^i = 1$. The posterior density at time *k* is approximated as [7]

$$p(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_{0:k} - \mathbf{x}_{0:k}^i) \quad . \tag{5}$$

suppose $p(x) \propto \pi(x)$ is a probability density from which it is difficult to get samples, but $\pi(x)$ can be evaluated, let $x^i \sim q(x), i = 1,..., N_s$ be samples that are easily obtained from a proposal $q(\cdot)$ called importance density function. Then a weighted approximation to the density $p(\cdot)$ is

$$p(x) \approx \sum_{i}^{N_s} w^i \delta(x - x^i) \quad . \tag{6}$$

where

$$w^i \propto \frac{\pi(x^i)}{q(x^i)} \quad . \tag{7}$$

Eq.(7) is the normalized weight of the *i*-th particle. If the sample $\mathbf{x}_{0:k}^i$ can be drawn from an importance density function $q(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k-1})$, then Eq.(5) and Eq.(7) can define as

$$w_k^i \propto \frac{p(\mathbf{x}_{0:k}^i \mid \mathbf{z}_{1:k})}{q(\mathbf{x}_{0:k}^i \mid \mathbf{z}_{1:k})} \quad . \tag{8}$$

the importance density function $q(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k})$ is expressed as

$$q(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k}) = q(\mathbf{x}_k \mid \mathbf{x}_{0:k-1}, \mathbf{z}_{1:k}) q(\mathbf{x}_{0:k-1} \mid \mathbf{z}_{1:k-1}) \quad . \tag{9}$$

then Eq.(4) can be derived as follows

$$\begin{aligned} p(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k}) &= \frac{p(\mathbf{z}_k \mid \mathbf{x}_{0:k}, \mathbf{z}_{1:k-1}) p(\mathbf{x}_{0:k} \mid \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k \mid \mathbf{z}_{1:k-1})} \\ &= \frac{p(\mathbf{z}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{0:k-1}, \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k \mid \mathbf{z}_{1:k-1})} \end{aligned} \tag{10}$$

$$\propto p(\mathbf{z}_k \mid \mathbf{x}_k) p(\mathbf{x}_k \mid \mathbf{x}_{k-1}) p(\mathbf{x}_{0:k-1} \mid \mathbf{z}_{1:k-1}) \tag{11}$$

By substituting Eq.(9) and Eq.(11) into Eq.(8), the weight updating equation can be shown as

$$\begin{aligned} w_k^i &\propto \frac{p(\mathbf{z}_k \mid \mathbf{x}_k^i)p(\mathbf{x}_k^i \mid \mathbf{x}_{k-1}^i)p(\mathbf{x}_{0:k-1}^i \mid \mathbf{z}_{1:k-1})}{q(\mathbf{x}_k^i \mid \mathbf{x}_{0:k-1}^i, \mathbf{z}_{1:k})q(\mathbf{x}_{0:k-1}^i \mid \mathbf{z}_{1:k-1})} \\ &= w_{k-1}^i \frac{p(\mathbf{z}_k \mid \mathbf{x}_k^i)p(\mathbf{x}_k^i \mid \mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i \mid \mathbf{x}_{0:k-1}^i, \mathbf{z}_{1:k})} \quad . \end{aligned} \tag{12}$$

If $q(\mathbf{x}_k \mid \mathbf{x}_{0:k-1}, \mathbf{z}_{1:k}) = q(\mathbf{x}_k \mid \mathbf{x}_{k-1}, \mathbf{z}_k)$, then the importance density function is only dependent on $\mathbf{x}_{k-1}$ and $\mathbf{z}_k$. The modified weight is then

$$w_k^i \propto w_{k-1}^i \frac{p(\mathbf{z}_k \mid \mathbf{x}_k^i)p(\mathbf{x}_k^i \mid \mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k^i \mid \mathbf{x}_{k-1}^i, \mathbf{z}_k)} \quad . \tag{13}$$

and the posterior density $p(\mathbf{x}_k \mid \mathbf{z}_{1:k})$ can be approximated as

$$p(\mathbf{x}_k \mid \mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i) \tag{14}$$

where the weights are determined by Eq.(13), it can be show that as $N_s \to \infty$ the approximation Eq.(14) can approach the true posterior probability density $p(\mathbf{x}_k \mid \mathbf{z}_{1:k})$ .A common problem of the particle filter is the degeneracy phenomenon, it is discussed in [8].

2.2 Selection of Importance Density Function and Resampling

The Eq.(15) shows the method that selects $q(\mathbf{x}_k \mid \mathbf{x}_{k-1}^i, \mathbf{z}_k)$ which can minimize the variance of the true weights,

$$q(\mathbf{x}_k \mid \mathbf{x}_{k-1}^i, \mathbf{z}_k)_{opt} = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}^i, \mathbf{z}_k) = \frac{p(\mathbf{z}_k \mid \mathbf{x}_k, \mathbf{x}_{k-1}^i)p(\mathbf{x}_k \mid \mathbf{x}_k^i}{p(\mathbf{z}_\mathbf{k} \mid \mathbf{x}_{k-1}^i)} \tag{15}$$

substitute Eq.(15) into Eq.(13) ,then

$$w_k^i \propto w_{k-1}^i p(\mathbf{z}_k \mid \mathbf{x}_{k-1}^i) = w_{k-1}^i \int p(\mathbf{z}_k \mid \mathbf{x}_\mathbf{k}^{'}) p(\mathbf{x}_k^{'} \mid \mathbf{x}_{k-1}^i) d\mathbf{x}_k^{'} \tag{16}$$

It is difficult to work out Eq.(16) directly, generally we choose the importance density function as the prior pdf of the system state transfer, that is

$$q(\mathbf{x}_k \mid \mathbf{x}_{k-1}^i, \mathbf{z}_k) = p(\mathbf{x}_k \mid \mathbf{x}_{k-1}^i) \tag{17}$$

substitute Eq.(17) into Eq.(14), the weights iteration formula in common is.

$$w_k^i \propto w_{k-1}^i p(\mathbf{z}_k \mid \mathbf{x}_k^i) \tag{18}$$

The resampling step involves generating a new set $\{\mathbf{x}_k^{i*}\}_{i=1}^{N_s}$ by resampling N_s times to posterior pdf $p(\mathbf{x}_k \mid \mathbf{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i)$, so that $\Pr(\mathbf{x}_k^{i*} = \mathbf{x}_k^j) = w_k^j$. The resampling obeys i.i.d in fact, and so the weights are now reset to $w_k^i = 1/N_s$.

3 Tracking of Sequential MCM over Time-Varying Channels

The channel tracking methods based on adaptive filter are important to obtain the channel state information, but they have some problems[9][10][11]. With the increase of computer's operation speed, more attentions have been paid to the Sequential MCM relying on computation. Assume that the expression of time-varying channels[12] is

$$h(n) = \alpha h(n-1) + (1-\alpha)v(n) \ . \tag{19}$$

where the noise $v(n)$ is complex Gaussian distribution, mean is zero and variance is $\sigma^2{}_v$ which is statistically independent of channel state $h(n-1)$, the estimating method of coefficient α is detailed in [13]

$$\alpha = J_0(2\pi f_d T_s)\exp(j2\pi f_0 T_s) \ . \tag{20}$$

where $J_0(\cdot)$ is the zero order Bessel function, f_d is the maximum Doppler shift , f_0 is carrier frequency offset , and T_s is the information symbol duration.

In order to investigate the tracking performance of the Sequential MCM, we set up a simple communications system as

$$y(n) = h(n)s(n) + u(n) \ . \tag{21}$$

where $y(n)$ is the receive signal , $s(n)$ is the BPSK modulation signal and $u(n)$ is additive Gaussian noise with zero mean and σ_u^2 covariance. $s(n)$ can be obtained by least square estimation

$$\hat{s}(n) = (h(n)^H h(n))^{-1} h(n)^H y(n) \ . \tag{22}$$

Eq.(19) and Eq.(21) are the simplified form of system state space equations of Eq.(1) and Eq.(2), and the estimated tracking value of channel $h(n)$ can be obtained by the Sequential MCM. The procedure of the scheme is as follows:

(1) From Eq.(19), initialize the particle set $\{h_0^i : i = 1...N\}$, set the weights w_0^i to 1/N, and get a new particle set h_n^i of $h(n)$ with h_{n-1}^i.

(2) Calculate the particle weights using the likelihood function

$$lik_n^i = f(y(n) \mid h_n^i) = (\sigma_u \sqrt{2\pi})^{-1} exp(-(2\sigma_u^2)^{-1}(y(n) - h_n^i \hat{s}(n))^{2} \qquad (23)$$

.

(3) Normalize weights

$$w_n^i = \frac{lik_n^i}{\sum_{j=1}^{N} lik_n^j} \quad . \qquad (24)$$

(4) If w_n^i is smaller than $\hat{N}_T$, then resample[14], or take the mean of w_n^i as the estimation of $h(n)$.

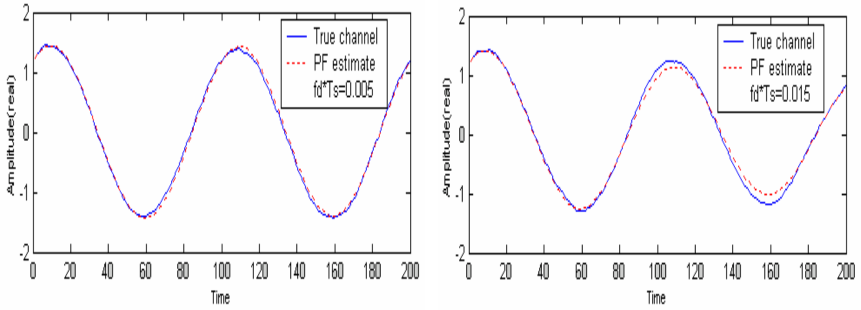


Fig. 1. Sequential MCM tracking performance over time-varying channels

In the simulation of the single-input single-output time-varying channel, the channel is initialized to be h_0=1.2-0.8i, and the total number of particles N=100. If commercial crystal oscillators are adopted, a normalized frequency offset will be introduced in the simulation [15], suppose $f_0 T_s$ =0.01, set Doppler normalized coefficients $f_d T_s$ to 0.005 and 0.015 respectively (an equivalent of moving velocity 36 kilometers and 108 kilometers respectively in the IS-136 communication environment). In the procedure (4), if w_n^i is smaller than the threshold $\hat{N}_T$ =2N/3, then resample, the channel tracking performances are shown in Figure.1 respectively. It is easy to get the conclusion that Sequential MCM can track the time-varying channel with good performance.

4 The Channel Tracking of Multi-antenna with STBC

4.1 Space Time Block Coding

To two transmit antenna and one receive antenna, the STBC's scheme is: when the time interval is *2n*, the antenna 1 transmits the symbol $s(2n)$, and the antenna 2 transmits $s(2n+1)$, when interval is *2n+1*,they transmit $-s^{*}(2n+1)$ and $s^{*}(2n)$ respectively. The channels are assumed flat fading in [16], differently we consider here time-selective fading channels i.e. time-varying channels. Let $h_i(n), i=1,2$ denote the channel response from the *i*-th transmit antenna to the receive antenna over the time-varying channel, the received symbols are $y(2n)$ and $y(2n+1)$, the expressions are

$$y(2n)=h_1(2n)s(2n)+h_2(2n)s(2n+1)+w(2n) \quad . \tag{25}$$

$$y(2n+1)=-h_1(2n+1)s^{*}(2n+1)+h_2(2n+1)s^{*}\left(2n\right)+w(2n+1 \quad . \tag{26}$$

where the noise $w(n)$ is complex Gaussian distributed with zero mean and $\sigma^{2}{}_{w}/2$ variance per dimension.

The STBC's decoding scheme is:

$$\boldsymbol{r}(n)=\boldsymbol{H}^{H}(n)\boldsymbol{Y}(n) \quad . \tag{27}$$

where *H* denotes conjugate transpose, the decoding symbol $\boldsymbol{r}(n)=[r(2n),r(2n+1)]^{T}$, and the channel matrix ***H***(n)

$$\boldsymbol{H}(n)=\begin{bmatrix} h_1(2n) & h_2(2n) \\ h^{*}{}_{2}(2n+1) & -h^{*}{}_{1}(2n+1) \end{bmatrix} \quad . \tag{28}$$

the receive ***Y***(n)

$$\boldsymbol{Y}(n)=[y(2n),y^{*}(2n+1)]^{T}=\boldsymbol{H}(n)\boldsymbol{s}(n)+\boldsymbol{w}(n) \quad . \tag{29}$$

where $\boldsymbol{w}(n)=[w(2n),w^{*}(2n+1)]^{T}$, $\boldsymbol{s}(n)=[s(2n),s(2n+1)]^{T}$, the final $\boldsymbol{s}(n)$ is obtained by maximum likelihood decoding based on Eq.(27) [16].

4.2 Channel Tracking Based on MCM in Multi-antenna System

STBC transmits two codes in two time intervals as one group, and the channel is considered to be invariable during the two intervals, i.e $h^{i}(2n-1)=h^{i}(2n), i=1,2\ n=1,2...$ .Now we suppose the channel is time-varying , extended the Eq.(19)

$$h_{2n}^{i} = \alpha h_{2n-1}^{i} + (1-\alpha)v^{i}(n),\ i = 1,2 \quad . \tag{30}$$

where $v(n)$ is a variable obeying the complex Gaussian $(0, \sigma_v^2)$ distribution.

When the channel is slow time-varying ($f_d T$ <0.01), $\boldsymbol{H}_n$ can keep its orthogonality, the maximum likelihood decoding can be adopted. When the channel h is fast time-varying, by least square decoding we can get

$$\hat{\boldsymbol{s}}(n) = (\boldsymbol{H}_n^H \boldsymbol{H})^{-1} \boldsymbol{H}_n^H \boldsymbol{Y}(n) \quad . \tag{31}$$

from the state equation Eq.(30) and Eq.(29) , we can estimate the change of channel's response based on the MCM, the procedures are as follows:

(1) Get the rough estimation $\bar{h}$ of the present channel using the prior channel $\hat{h}$ with $\bar{h}_{2n} = \alpha \hat{h}_{2n-1}$ and $\bar{h}_{2n+1} = \alpha^2 \hat{h}_{2n-1}$. From Eq.(28) and Eq.(31) we can get the rough estimation of transmit signals by decoding scheme $\hat{\boldsymbol{s}}(n) = [\hat{s}(2n), \hat{s}(2n+1)]^T$.

(2) Estimate $\hat{h}_{2n}$ using $\hat{h}_{2n-1}$ with MCM mentioned in above section, then estimate $\hat{h}_{2n+1}$ using $\hat{h}_{2n}$.

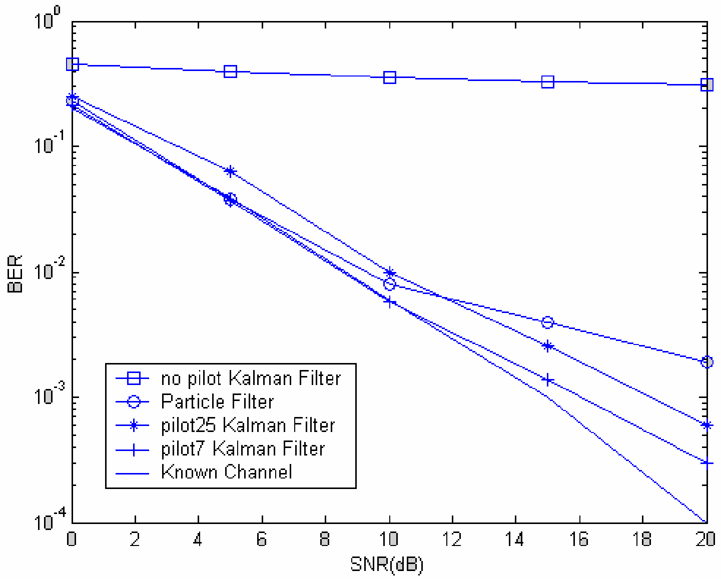


Fig. 2. Performance comparison of particle filter and Kalman filter

(3) Obtain the estimation $\hat{\boldsymbol{H}}_n$ of Eq.(28) using $\hat{h}_{2n}$ and $\hat{h}_{2n+1}$.

(4) Decode the transmit symbol $\boldsymbol{s}(n) = [s(2n), s(2n+1)]^T$ using $\hat{\boldsymbol{H}}_n$ and $\boldsymbol{Y}(n)$ by Eq.(31).

A simulation example: two transmit antennas, one receive antenna over time-varying channel, the channel response of antenna 1-1 is initialized to h_0^1= 1.3+0.8j, and the one of antenna 2-1 is set to h_0^2= -0.9+1.4j. The total number of particles N=100, carrier frequency offset coefficient $f_0 T$ =0.01, and the Doppler normalized coefficient $f_d T$ =0.01, the performance (Bit Error Rate and Signal-to-Noise Ratio) of channel tracking is shown in Figure.2.

Blind tracking performance can be obtained based on particle filter scheme, it has the advantage of no need for pilots and then can save frequency resource, but it has to pay more computational cost. Kalman filter scheme has better performance, the figure shows the simulations of no pilots, inserting one pilot symbol in every 7 and 25 symbols, but this scheme costs more frequency resource.

5 Conclusions

In this paper, the channel estimation and tracking based on MCM for multi-antenna system with STBC which has carrier frequency offset is discussed. Compared with Kalman tracking method, MCM has heavier computational burden, Kalman filter method needs inserting pilots and known noise variance, while MCM does not need these, so it can save frequency resource. In practice, the total number of particles is chosen from 100 to 200, but the farther increase number of particles contributes little to the performance of the method. How to improve the MCM has become a research hot spot, it focuses on the following aspects: the comparison of different particle filter methods, the design of importance function, the resampling method to decrease the computational complexity, the sub-optimal algorithm to decrease the computational complexity, the problem of convergence, the comparison of adopting different resampling procedures, estimating errors and computational complexities. Furthermore, the special processor integrated hardware and software better will have more practical value in the tracking and navigation, etc.

References

1. Carpenter, J., Clifford, P., Fearnhead, P.: Improved Particle Filter for Nonlinear Problems. IEEE proc. Radar,Sonar,Navig. 1 (1999) 146
2. Doucet, A., Godsill, S.J., Andrieu, C.: On sequential Monte Carlo Sampling Methods for Bayesian Filtering. Statistics and Computing, (2000) 197-208
3. Clapp, T.C., Godsill, S.J.: Fixed Lag Smoothing Using Sequential Importance Sampling,in Bayesian Statistics 6 Eds.J.M.Bernardo, J.O.Berger, A.P.Dawid, and A.F.M.Smith, Oxford: Oxford University Press, (1999) 743-752

4. Chen, R., Wang, C., Liu, J.S.: Adaptive Joint Detection and Decoding in Flat-fading Channels via Mixture Kalman Filtering, IEEE Transactions on Information Theory, 46 (2000) 2079-2094
5. Doucet, A., Godsill, S., West, M.: Monte Carlo Filtering and Smoothing with Application to Time-varing Spectral Estimation. Proceedings of the International Conference on Acoustics Speech and Signal Processing, Istanbul,Turkey. (2000)
6. Djuric, P. M.: Sequential Estimation of Signals Under Model Uncertainty in Sequential Monte Carlo Methods in Practice. A. Doucet, N.de.Freitas and N.Gordon,Eds,New York:Springer Verlay. (2001)
7. Sanjeev Arulampalam, Simon Maskell,Neil Gordon, Tim Clapp.: A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking. IEEE Trans on
8. Liu, J.S., Chen, R.: Sequential Monte Carlo Methods for Dynamical Systems. Journal of the American Statistical Association, 93 (1998) 1032-1044.
9. Taylor, D.P, Vitetta,G.M., HartB, D.etal.: Wireless Channel Eqaulization. Euro. Trans. Telecommun. (1998) 117-143
10. Haykin, S.: Adaptive Filter Theory, The Third Edition, Beijing: Publishing House of Electronics Industry. (1998)
11. Omidi, M.J., Pasupathy, S., Galuk, P.G.: Joint Data and Kalman Estimation for Rayleigh Fading Channels, Wireless Personal Communications, 3 (1999) 319-339
12. Chin, W.H., Ward, D.B., Constantinides, A.G.: Channel Tracking for Space-time Block Coded Systems Using Particle Filtering[C], Proc. DSP. (2002) 671-674
13. Tsatsanis, M.K., Giannakis, G.B., Zhou G.: Estimation and Equalization of Fading Channels with Random Coefficients[J]. Signal Processing 53 (1996) 211-228
14. Doucet, A., de Freitas, J. F. G., Gordon, N. J.: An introduction to sequential Monte Carlo methods in Sequential Monte Carlo Methods in Practice, Springer-Verlag: New York (2001)
15. Kannan, A., Krauss, T.P., Zoltowski, M.D.: Separation of Co-channel Signals Under Imperfect Timing and Carrier Synchronization, IEEE Trans. on Veh.Technol. 50 (2001) 79-96
16. Alamouti, S.M.: A Simple Transmit Diversity Technique for Wireless Communications[J]. IEEE J.Select.Areas Com. 16 (1998) 1451-1458

Locating Human Eyes Using Edge and Intensity Information

Jiatao Song[1,2], Zheru Chi[2], Zhengyou Wang[3], and Wei Wang[1]

[1] College of Electronic and Information Engineering, Ningbo University of Technology,
Ningbo 315016, P. R. China
sjt6612@163.com
[2] Centre for Multimedia Signal Processing, Department of Electronic and Information
Engineering, the Hong Kong Polytechnic University, Hong Kong
enzheru@eie.polyu.hk
[3] School of Information Technology, Jiangxi University of Finance & Economics,
Nanchang 330013, China

Abstract. In this paper, a new eye detection method is presented. The method consists of three steps: (1) extraction of binary edge image (BEI) based on the multiresolution analysis of wavelet transform; (2) extraction of eye region and segments from BEI, and (3) eye localization using light dot or intensity information. An improved face region extraction algorithm and a light dot detection method are proposed to improve eye detection performance. Experimental results show that our approach can achieve a correct eye detection rate of 98.7% on 150 Bern images with variations in view and gaze direction and a rate of 96.6% on 564 AR images with different facial expressions and lighting conditions.

1 Introduction

In face image analysis, one fundamental problem is to automatically detect human eyes since localizing eyes is a necessary step for face image normalization, face detection and facial feature extraction [1-4]. Eyes can provide very useful information for face discrimination [1] than other face components such as nose and mouth. Phillip et al. [5] used whether the centers of eyes can be automatic located as a criterion to classify a recognition system into fully automatic or partially types.

A number of methods have been proposed for eye detection. Yuille et al. [6] used a deformable template to perform this task. The advantage of their method is that it can provide not only the location of an eye but also the contour of the eye. But the method is computationally expensive. Another commonly used method is the eigenspace approach proposed by Pentland et al. [7]. The method is easy to implement, but it requires the normalization of a face image in size and orientation. In addition, the method can only correctly detect those eye patterns that are similar to the sample eye images used as eye models. Brunelli and Poggio [8], and Beymer [9] located eyes using template matching. The method also faces the same problem as the eigenspace method. Chow and Li [10] employed the Hough transform to locate eyes.

Although some progress has been made, the problem of automatic eye detection is still far from being fully solved due to the problem complexity. Many factors, such as

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 492 – 501, 2005.

facial expression, face rotation in plane and depth, occlusion, lighting conditions, and so on, will affect the performance of eye detection algorithms. Unfortunately, most of the existing methods mainly focus on eye detection from frontal-view face images without specifically taking into consideration of the above-mentioned factors that affect the eye detection. Kawaguchi and Rizon [11] tried to develop an eye detection method with good robustness to gaze directions and reflected light dots. They used images from two databases, the Bern database [12] and the AR database [13], to evaluate the validity of their algorithm, achieving a correct iris detection rate of 95.3% for 150 Bern face images and 96.8% for 63 AR images. But there are two drawbacks with their method. Firstly, they did not show how to automatically detect the light dot in the iris. In other word, whether there is a light dot in the iris cannot be determined by the system. Secondly, the face region extraction method used is image dependent. For the images without light dots, such as Bern images, the simple integral projection based technique used works well, but for images with light dots, such as AR images, the technique is not applicable, and a color based method has to be adopted.

In this paper, a new eye detection method is presented. We address the problem of eye detection in images with variations in pose, gaze direction, face expression, and lighting condition. Furthermore, because there are often reflected light spots in the face images captured using camera with flash and these light spots often locate near the center of the iris, we try to use the light dot for eye localization. A light dot detection algorithm is therefore proposed in our study.

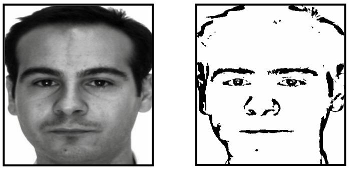


Fig. 1. Example of BEI. Left: a grayscale image; Right: BEI

The remaining of this paper is organized as follows. In Section 2, a method for the extraction of the binary edge image (BEI) is briefly discussed. The proposed eye detection method is presented in Section 3. In Section 4, experimental results on the Bern [12] and AR [13] face database are reported and discussed. Finally, concluding remarks are drawn in Section 5.

2 Extraction of Binary Face Image Edges

The first step of our proposed eye detection method is to segment the face image. It is realized by firstly producing the binary edge image (BEI) from a grayscale face image using the algorithm presented in [14]. Making use of the multi-resolution analysis property of Wavelet Transform, the BEI extraction algorithm includes two binarization steps and a noise removing step. The generated BEIs are shown in Figure 1. We can see clearly various face components including eyebrows, eyes, nose and mouth in BEIs. These components were all extracted with sufficient details and without touching neighboring face components in most cases. Furthermore, the BEI is of good robustness to light conditions. The BEI is suitable for face component segmentation and the extraction of some key feature points in the face image.

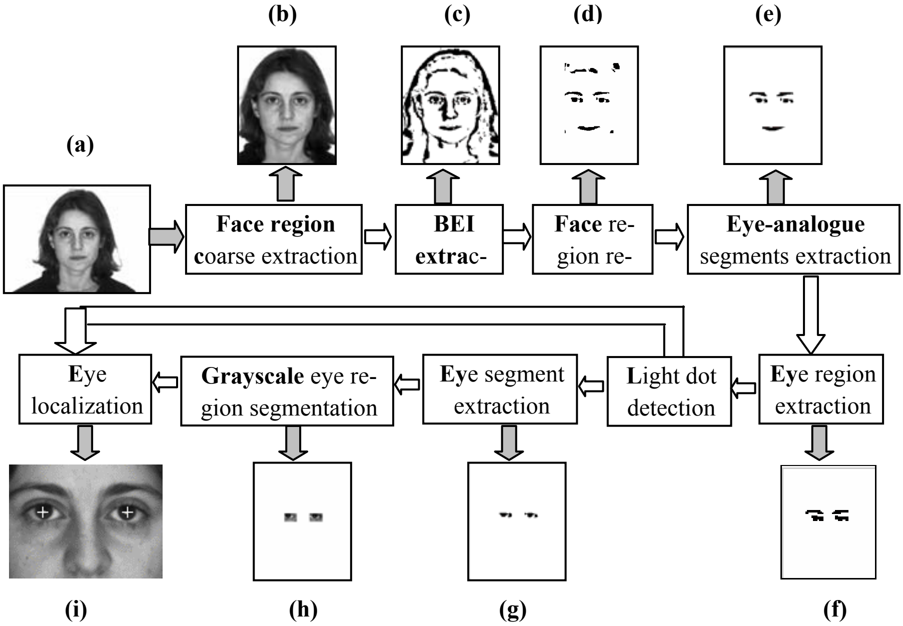


Fig. 2. Flowchart of our proposed eye detection method

3 Eye Localization Method

The images under investigation are supposed to be head and shoulder images with plain background. The flowchart of our method is illustrated in Fig. 2.

3.1 Face Region Extraction

Firstly we use the method proposed in [11] to roughly locate the face region and normalize the cropped face image into the size of $M*N$ (Figure 2 (b)). Here M and N denotes the height and width of the normalized face image, respectively. Obviously, when hair occupies large regions in both sides of the face, the left, right and lower boundaries of the face determined using this method may be far from the actual face boundaries. In this case, hair, neck, and even shoulder regions may be included in the extracted face region, increasing the difficulty in eye block detection. We propose a face region refinement algorithm to solve this problem.

The face boundaries are refined using the information containing in the BEI. We first define two measures to characterize the property of a foreground pixel in a BEI, **Horizontal Connection Length (HCL)** and **Vertical Connection Length (VCL)**.

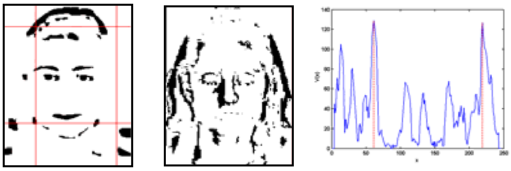
The **HCL** of pixel P, HCL_P, is defined as the maximum number of pixels which are horizontally connected to pixel P. The **VCL** of pixel P, VCL_P, is defined similarly except for pixels vertically connected. Furthermore, we define

$$\lambda_P = \frac{VCL_P}{HCL_P} \tag{1}$$

It is obvious that in a BEI, the λ values for most pixels in eyebrow segments, eye segments and the mouth segment is less than 1, while most pixels on the left and right boundaries of the face have a λ value of greater than 1.By using this feature, the left and right face boundaries can be easily removed from a BEI by iteratively removing those pixels which satisfy the following condition until none of them exists.

$$\lambda(k) = \frac{VCL(k)}{HCL(k)} \geq 3 \tag{2}$$

where *VCL(k)* and *HCL(k)* denotes the *VCL* and *HCL* after the *kth* removing operation steps.



(a) (b) (c)

Fig. 3. Refinement of face boundaries. (a) The BEI after the boundary pixel removing operation; (b) the removed pixels; (c) the vertical projection of (b)

Fig. 3(a) shows the BEI after the pixel removing operation. It can be seen that most pixels on the vertical face boundaries have been removed, while those horizontal face components, especially eye and eyebrow segments, have little change. Fig. 3(b) shows the removed pixels in which the vertical face boundaries are clearly seen. Thus, by projecting the removed BEI vertically and finding the locations of the two peaks at the left half and right half of the projection curve, the refined left and right face boundaries, denoted by x_L and x_R , can be determined. The upper face boundary y_U and the lower boundary y_L can be refined using a method similar to that described in [11]. The refined face boundaries are shown in Fig. 3(a). By removing those pixels outside the new face boundaries, a refined BEI, or **RBEI**, can be obtained (Fig. 2(d)).

3.2 Extraction of Eye-Analogue Segments

Eye segments are extracted from the RBEI. Before doing this, it is helpful to remove those blocks from a RBEI which are unlikely to be eye segments because of a small size or small maximum-*HCL*. The BEI after this processing is termed **PESBEI** which contains possible eye segments.

Eye-analogue segments are some large horizontal segments. They are extracted from PESBEI. The BEI containing eye-analogue segments only is termed eye-analogue segments BEI or **EASBEI**, as shown in Figure 2(e). It mainly contains eye, eyebrow or mouth blocks.

3.3 Detection of Eye Regions

An eye region refers to the area which contains eyes and eyebrows only. It is extracted by making use of a prior knowledge about the layout of the face components on a face. The knowledge includes (1) the vertical distance between an eye and the nose is normally greater than that between an eye and an eyebrow and that between the nose and the mouth; (2) there are two eyes and two eyebrows, but only one mouth and one nose on a face. Based on these two assumptions, the eye region can be extracted using horizontal and vertical integral projections. The BEI with eye and eyebrow segments only is termed eye region BEI or **ERBEI** as shown in Fig. 2(f).

3.4 Detecting Light Dots in the Eye Region

Using light spots for eye localization is one main contribution of our proposed eye detection method. Light dots are automatically detected from the face image. Light dots are some small holes in a BEI, i.e., some connected background pixels enclosed by foreground pixels. These holes can be easily labeled and located.

However, besides light dot holes, there are possible other holes in a BEI. In order to distinguish light dot holes from other holes, the distinct feature of a true light dot is used. In a grayscale face image a striking intensity contrast shows between a light dot and its surroundings. For each hole in a BEI, the intensity contrast, denoted by *C*, is defined as

$$C(i) = \frac{\max\limits_{(x,y)\in H_i} \{f(x,y)\}}{AI(i)} \quad (i = 1,2,\cdots,N_H) \tag{3}$$

where H_i represents hole *i*, N_H is the number of holes in a BEI, *f(x,y)* is the pixel intensity at location (*x,y*) of a grayscale face image, and *AI(i)* denotes the average pixel intensity of a small region around hole *i*. Because the intensity of the region around a light dot is very low, the contrast of a light dot hole is often higher than that of non-light dot holes. So, in our method, those holes in the eye region with an intensity contrast satisfying the following constraint are considered to be possible light dots:

$$C \geq \beta \tag{4}$$

where β is a threshold which was set to 2.5 in our experiments. Light dots in the left and right eye regions are detected separately. The possible light dot with the largest contrast is considered to be the true light dot.

3.5 Locating Eyes Using Intensity Information

If no light dot is detected in the left or right eye region, intensity information is used for eye localization. In order to do so, the left or right eye block is extracted from the ERBEI utilizing the horizontal integral projection.

After an eye block is extracted, the upper, lower, left and right boundaries of that eye can be determined. The boundary information is then used to guide the extraction of an eye from the grayscale face image (Figure 2(h)). By finding the lowest pixel intensity, denoted by *LPI*, from each grayscale eye region and scanning all pixels included in the following point set, denoted by *PS*:

$$PS = \{(x,y) \mid f(x,y) \le (\alpha \times LPI),\ (x,y) \in \text{the left or right eye region}\} \tag{5}$$

where *α* is a constant (>1) determined by experiments and *f(x,y)* is the same as that in Eq.(3) , the average position of *PS* is considered as the final eye location.

4 Experimental Results and Discussion

Two face databases used by Kawaguchi [11] are employed to evaluate the performance of our method. They are Bern and AR face databases. The Bern images we used include 150 face images without spectacles, consisting of 10 views of 15 persons. The AR images we used are all the 564 images of the 94 persons without spectacles in the first four CD-ROM of the AR database, representing three expressions, i.e., neutral, smile and anger under natural illumination environment, and three other lighting conditions, i.e., left light on, right light on and all side lights on, with neutral expression. The AR database is aimed to evaluate the robustness of our eye detection method to variations in eye appearance and lighting conditions, and to demonstrate the role of the reflected light dots in automatic eye localization.

In our experiments, the relative error proposed by Jesorsky [15] is used to evaluate the performance of our eyes detection method. The relative error is defined as

$$err = \frac{\max(d_l, d_r)}{d_{lr}} \tag{6}$$

where d_l is the disparity between the manually determined left eye location and the automatically detected left eye position, d_r is the right eye disparity, and d_{lr} the Euclidean distance between the manually determined left and right eye locations. The criterion for a successful eye detection we adopted is *err*<0.125. Because d_{lr} is about 8 times of the radius of an iris, denoted by *r*, our criterion approximates to

$$\max(d_l, d_r) < r \tag{7}$$

that is, the maximum disparity is small than the radius of an iris.

Figure 4 shows some examples of Bern and AR images of which eyes are correctly located by our proposed method. The detailed experimental results on 150 Bern

Table 1. Results of eyes detection on 150 Bern images

Algorithm	Correct locating rate(%)
Proposed algorithm	98.7
Kawaguchi's algorithm [11]	95.3
Template matching [11]	77.9
Eigenspace method, 50 training samples [11]	90.7
Eigenspace method, 100 training samples [11]	93.3

images are summarized in Table 1. It shows that our method can achieve a higher correct locating rate than Kawaguchi's method, template matching method, and the Eigenspace method. We found that our method is more robust to variations in view and gaze direction.

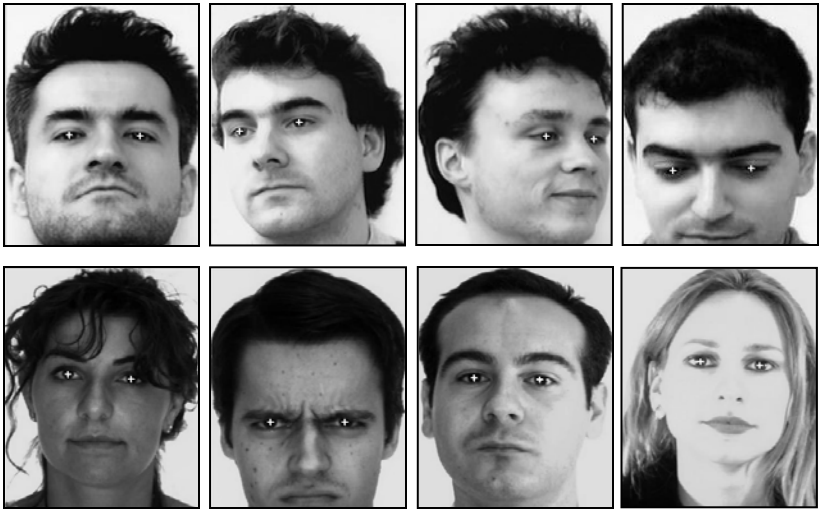
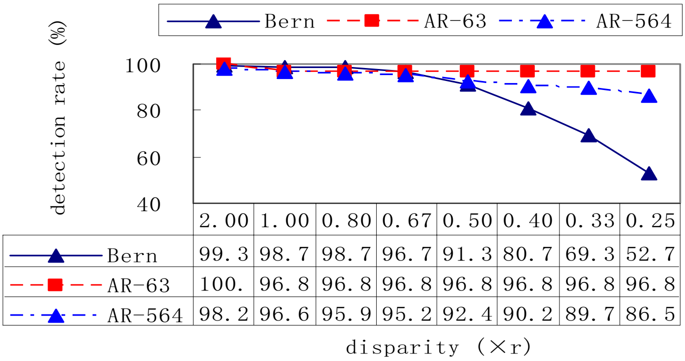


Fig. 4. Samples of images of which two eyes are correctly detected by our proposed method. (Upper row: Bern images; lower row: AR images)

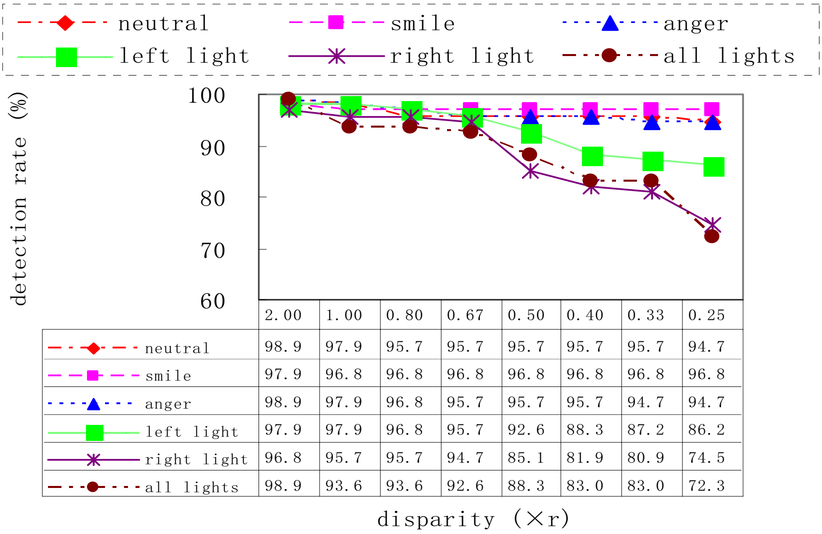
For AR images, we first used the subset AR-63 for experiment. This subset includes the 63 images used by Kawaguchi [11] and is from 21 people (12 men and 9 women) without spectacles, representing three expressions (neutral, smile and anger). A correct eye locating rate of 96.8% is achieved, which is the same as that reported by Kawaguchi [11]. Figure 5 shows a more detailed result of our experiments. From this Figure we can see that when the eye locating criteria, denoted by the maximum *disparity* of two eyes, which is measured in the radius of an iris r, is decreased to $0.25r$, the correct eye locating rate for AR-63 can still hold at 96.8%, indicating that our method achieves not only a high locating rate but also a high eye locating accuracy.

We also tested AR-564 images. This subset of 564 AR images is divided into six groups, representing three different expressions and three lighting conditions, and all groups are tested separately. Figure 5 shows that when the criterion defined in Eq. (7) is adopted, for all the 564 images tested a correct eye detection rate of 96.6% is achieved. Figure 6 shows that when the locating criterion is higher, the eye detection rate only degrades slightly for image groups with expression changes, while for image groups with illumination changes, the eye detection rate degraded significantly. This indicates that (1) our proposed method can achieve both a high eye detection rate and a high eye location accuracy. Even in the worst case, i.e., all side lights on, the eye detection rate reaches 93.6%; (2) The influence of variations in facial expressions on the eye detection rate is less than that caused by illumination change.



	2.00	1.00	0.80	0.67	0.50	0.40	0.33	0.25
Bern	99.3	98.7	98.7	96.7	91.3	80.7	69.3	52.7
AR-63	100.	96.8	96.8	96.8	96.8	96.8	96.8	96.8
AR-564	98.2	96.6	95.9	95.2	92.4	90.2	89.7	86.5

Fig. 5. The correctly eye detection rate for Bern, AR-63 and AR-564 images using different criteria



	2.00	1.00	0.80	0.67	0.50	0.40	0.33	0.25
neutral	98.9	97.9	95.7	95.7	95.7	95.7	95.7	94.7
smile	97.9	96.8	96.8	96.8	96.8	96.8	96.8	96.8
anger	98.9	97.9	96.8	95.7	95.7	95.7	94.7	94.7
left light	97.9	97.9	96.8	95.7	92.6	88.3	87.2	86.2
right light	96.8	95.7	95.7	94.7	85.1	81.9	80.9	74.5
all lights	98.9	93.6	93.6	92.6	88.3	83.0	83.0	72.3

Fig. 6. The eye detection rate for the AR-564 using different criteria

The high performance on AR images can be explained by a high light dot detection rate. Experiments show that for AR-63, there are 61 images (96.8%) of which two eyes can be located using reflected light dots, while for the neural group, smile group, anger group, left-light-on group, right-light-on group, and all-side-light-on group of AR-564, their light dot detection rates are 93.6%, 96.8%, 94.7%, 84.0%, 70.2%, 67.0%, respectively.

It is worth pointing out that by making use of light dots for eye detection, eyes partially occluded by hair (see the first image in the lower row of Fig. 4) and eyes closely connected with the eyebrows (see the second image in the lower row of Fig. 4) can be correctly located by our proposed method, which certainly increases the overall eye locating rate.

5 Conclusion

In this paper, a new method for human eyes localization has been presented. Our method consists of three steps: generation of a binary edge image (BEI) from the grayscale face image, extraction of the eye region and eye segments from the BEI, and eye localization using light dots or intensity information. Experimental results show that a correct eye detection rate of 98.7% for 150 Bern images without spectacles and 96.6% for 564 AR images is achieved, indicating the effectiveness of our method in dealing with variations in view, gaze direction, expression and lighting condition.

The information we used for the eye detection includes image intensity, light dots, the shape of an eye segment, and the configuration of face components. Obviously, all these are intrinsic attributes of a face image and eyes, and have nothing to do with image size. So, unlike conventional template matching and Eigenspace based approaches, our method needs no normalization in image size. Our proposed method can also tolerate a large image rotation in plane and depth.

Based on our proposed eye detection method, much more work can be carried out in the future. For example, to study the lighting conditions in which reflected light spots may form, and to develop a light dot based eye detection system is a very promising research direction. The system will be very similar to the active IR based eye detection and tracking system [16], but does not require special image capturing, so it will be more applicable.

Acknowledgements

The work described in this paper was partially supported by a research grant from the Hong Kong Polytechnic University (G-T768), a grant from Zhejiang Provincial Natural Science Foundation of China (602118), and a grant from the Scientific Research Foundation for the Returned Overseas Chinese Scholars of Zheiang Province (72).

References

1. Chellappa, R., Wilson, C. L., and Sirohey, S.: Human and machine recognition of faces: A survey. Proc. IEEE. Vol.83, No.5(1995) 705-741
2. Zhao, Z. Q., Huang, D.S. and Sun, B.-Y.: Human face recognition based on multiple features using neural networks committee. Pattern Recognition Letters. Vol.25, No.12(2004)1351-1358
3. Guo, L. and Huang, D.S.: Human face recognition based on Radial basis probabilistic neural network. Int. Joint Conf on Neural Networks (IJCNN2003), Portland, Oregon, July 20-24, (2003) 2208-2211
4. Yang, M.-H., Kriegman, D. J., Ahuja, N.: Detecting faces in images: A survey. IEEE Trans. on PAMI. Vol.24, No.1(2002) 34-58
5. Phillips, P. J., Moon, H., Rizvi, S. A., et al: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. on PAMI. Vol.22, No.10(2000) 1090-1104
6. Yuille, A. L., Hallinan, P. W., and Cohen, D. S.: Feature extraction from faces using deformable template. Int. J. Computer Vision. Vol.8, No.2(1992)99-111

7. Pentland, A., Moghanddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. IEEE conference on CVPR, Seattle, June (1994) 84-91
8. Bruneli, R., Poggio T.: Face Recognition: features versus templates. IEEE Trans. on PAMI. Vol.15, No.10(1993) 1042-1052
9. Beymer, D. J.: Face recognition under varying pose. Proc. of IEEE conference on CVPR, Seattle, June (1994) 756-761
10. Chow, G., Li X.: Towards a system for automatic facial feature detection. Pattern Recognition. Vol.26, No.12(1993) 1739-1755
11. Kawaguchi, T., Rizon, M.: Iris detection using intensity and edge information. Pattern Recognition. Vol. 36, No.2 (2003) 549-562
12. Achermann, B.:The face database of University of Bern. http://iamwww.unibe.ch/~fkiwww/staff/achermann.html. Institute of Computer science and Application Mathematics, University of Bern, Switzerland (1995)
13. Martinez, A. M. and Benavente, R.: The AR Face Database. CVC Technical Report #24, June (1998)
14. Song, J., Chi, Z., Liu, J. and Fu H.: Extraction of Face Image Edges with Application to Expression Analysis, Proceedings of the 8th International Conference on Control, Automation, Robotics and Vision (ICARCV), Kunming, China, December (2004) 804-809
15. Jesorsky, O., Kirchberg, K. J., and Frischholz R.W.: Robust Face detection using the Hausdorff distance. Proc. of the Third International Conference on Audio- and Video-based Biometric Person Authentication, Halmstad, Sweden, 6-8 June (2001) 90-95
16. Zhu, Z., Fujimura, K., Ji, Q.: Real-time eye detection and tracking under various light conditions. http://www.ecse.rpi.edu/homepages/qji/papers/Acmpaper.pdf

A Nonlinear Adaptive Predictive Control Algorithm Based on OFS Model

Haitao Zhang [1,2], Zonghai Chen[2], Ming Li[2], Wei Xiang[2], and Ting Qin[2]

[1] Department of Control Science and Engineering, Huazhong University of Science and Technolog, Wuhan, 430074, P.R.China
zht@mail.ustc.edu.cn
[2] Department of Automation, University of Science and Technology of China, Hefei, 230027, P.R.China

Abstract. Firstly, a method is introduced which uses Volterra series deploying technique to construct a nonlinear model based on OFS model. Then an improved novel incremental mode multiple steps adaptive predictive control strategy is brought forward, which can import more information about the system's dynamical characteristics. Experiments of constant water pressure equipment's control prove that this proposed algorithm can effectively alleviate system's oscillation when used to control a plant with severe nonlinearity, and that this algorithm shows good robustness for outer disturbances. So it is suitable to be generalized to the design of complex industrial process controller.

1 Introduction

Severe nonlinearities universally exist in Complex industrial processes, such as distillation process, pH neutralization process, etc. The control of severe nonlinearities is an urgent task in the field of process control all along [8].

The standard predictive control algorithm based on OFS model[2,10] has many advantages such as non-sensibility of system structure and time-delay variances, low computational complexity, and robustness of parameter variances, etc., therefore, when dealing with linear or mildly nonlinearity, this algorithm can give satisfying control performance[2,3,11]. However, this standard algorithm is based on linear OFS model, so for severe nonlinear plant, this algorithm's performance is not good enough. Reference [1] gives an improved nonlinear control algorithm based on OFS model, but the adaptive effect is not satisfying when treating severe nonlinearity. Reference [1] combines Volterra Series with OFS model to identify nonlinear plant, and gains comparatively good control performance in the control simulation on CSTR (continuous stirring tank reactor) plant, but the computation load is too heavy, and real-time performance is not satisfying. Therefore, design an effective adaptive control algorithm for severe nonlinearities is an urgent task.

In this paper, firstly, the conception of Volterra Series is summarized. Then, the method to construct nonlinear hybrid model is introduced which based in OFS model combined with Volterra series. Based on this hybrid model, we propose an improved incremental mode nonlinear adaptive predictive control algorithm, which extend the

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 502 – 511, 2005.

control horizon from 1 step to 2 steps by using Gröbner basis function. The pump's output water pressure control experiment performances validate that this proposed algorithm can gain satisfying dynamic and steady-state performances. Moreover, compared with former nonlinear Laguerre predictive algorithm whose control horizon is one [1, 6] , this one has better robustness under outside disturbances, besides, the vibrations of the control and controlled signals can be effectively decreased. Thus, the feasibility and superiority of this proposed algorithm are validated.

2 Volterra Series of Nonlinear Systems

Theorem 1[9] If the nonlinear system $F[u(t)_{-\infty}^{0}]$ fulfills the following 4 assumptions: 1)causal; 2)time-invariant; 3)input signal's power is finite; 4)the corresponding function of the system is continuous;
then, the system can be discomposed to the serial connection of a set of dynamic linear block and a memoryless nonlinear block.

Here, the linear block is $u(t)=\sum_{i=1}^{N} a_i P_i(t)$, $a_i = \int_{-\infty}^{0} u(t) P_i(t) dt$,and $\{P_i(t)\}_{i=1}^{\infty}$ is set of complete orthonormal bases. In practice, we always choose the set of impulse response of this dynamic linear block $g_i(t) = P_i(-t)$, then

$$a_i(t) = \int_0^{\infty} g_i(\tau) u(t-\tau) d\tau \tag{1}$$

The memoryless nonlinear system is defined as

$$F_N(a_1,\cdots,a_N) \overset{\Delta}{=} F\left[\sum_{i=1}^{N} a_i P(t)\right] \tag{2}$$

Theorem2 (Weierstrass Theorem) [5]

Any continuous function $f(x)$ defined can be approximated uniformly to arbitrary accuracy on any finite, closed interval $[a,b]$ by polynomial of finite degree. ■

Definition1. If $F_N(a_1,\cdots,a_N)$ can be deployed to [2,11]

$$F_N(a_1,\cdots,a_N) = \alpha + (\beta_1\alpha_1 + \beta_2\alpha_2 + \cdots + \beta_N\alpha_N) + (\gamma_{11}\alpha_1^2 + \gamma_{12}\alpha_1\alpha_2 + \cdots + \alpha_N^2) + \cdots \tag{3}$$

then, define $h_0 = \alpha$ as the 0 order Volterra kernel, $h_1(\tau) = \sum_{i=1}^{N} \beta_i g_i(\tau)$ as the 1 order Volterra kernel, $h_2(\tau_1,\tau_2) = \sum_{i,j=1}^{N} \gamma_{ij} g_i(\tau_1,\tau_2)$ as the 2 order kernel. ■

So, from (3), we have

$$y(t)=F[u(t)]\approx F_N(a_1,\cdots,a_N)=\sum_{n=0}^{\infty}y_n(t) \tag{4}$$

here,

$$y_n(t)=\int_0^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}h_n(\tau_1,\tau_2,\cdots,\tau_n)\prod_{i=1}^{n}u(t-\tau_i)d\tau_i \tag{5}$$

which is the Volterra series expression of nonlinear system for $u(t)$.

3 Incremental Mode Nonlinear Adaptive Control Algorithm

Theorem 3[12] The incremental model linear control algorithm[12] based on OFS model can eliminate steady-state error under output disturbances. ■

Theorem 3 has proved that the incremental OFS adaptive predictive control algorithm can trace set point curve without error, so it is superior to whole mode algorithm [2,6]. So, in order to enhance the closed-loop steady-state performance, we combine incremental mode OFS model with Volterra series, and construct a nonlinear adaptive algorithm based on this hybrid model to deal with server nonlinearities' control.

For convenience, we ignore the Volterra kernels whose orders are equal to or larger than 3, and use N order Laguerre functional series, which is the most elegant one of the OFS family[10], $\varphi_i(\tau)$, $i=1,\cdots,N$, to approximate the 1 order and 2 order Volterra kernels which belong to the space $\mathbf{L}^2(R^+)$ (square integrabel space on $[0,\infty)$), then

$$h_1(\tau_1)=\sum_{i=1}^{N}c_i\varphi_i(\tau_1) \tag{6}$$

$$h_2(\tau_1,\tau_2)=\sum_{n=1}^{N}\sum_{m=1}^{N}c_{nm}\varphi_n(\tau_1)\varphi_n(\tau_2) \tag{7}$$

here, c_i and c_{nm} are all constant coefficients, so

$$y(t)=h_0+\sum_{i=1}^{N}c_il_i(t)+\sum_{n=1}^{N}\sum_{m=1}^{N}c_{nm}l_n(t)l_m(t) \tag{8}$$

here,

$$l_j(t)=\int_0^{\infty}\varphi_j(\tau)u(t-\tau)d\tau \;\; j=1,\cdots,N \tag{9}$$

is the Laguerre functional series.

Nonlinear Laguerre model's state space expression is stated as follows

$$L(k+1)=AL(k)+bu(k) \tag{10}$$

$$y_m(k)=c_0+C^T L(k)+L^T(k)DL(k) \tag{11}$$

The definitions of A,b can refer to reference [2]. $c_0=h_0$, $C^T=[c_1,\cdots,c_N]$,
In the k th sampling period
$L(k)=[l_1(k),\cdots,l_N(k)]^T$ is the Laguerre state vector, $u(k)$ is the input.
$D=(c_{ij})_{N\times N}$ $(i=1,\cdots,N;j=1,\cdots,N)$
$y_m(k)$ is the output of Laguerre model.

From (13) and extend the prediction time, we can get that

$$y_m(k+i)=f\left(L(k+i)\right)=c_0+C^T L(k+i)+L^T(k+i)DL(k+i) \tag{12}$$

and

$$L(k+i)=A^i L(k)+\overline{A}_i bu(k) \tag{13}$$

Based on the reference [6], we have made the following 3 improvements.

1) import the output anti-disturbance amendment mechanism.
2) use RLS to update the coefficients matrices C,D to implement adaptive mechanism.
3) use Gröbner based function algorithm to extend the control horizon from 1 to 2, which can contain more information about the controlled plant's dynamic characteristics, and thus can yield better control performances. Its superiority will be validated in section 4.

In detail, let

$$y_m(k+i)=e_0(k,i)+e_1(k,i)u(k)+e_2(k,i)u^2(k) \tag{14}$$

here

$$e_0(k,i)=c_0+C^T A^i L(k)+L^T(k)(A^i)^T DA^i L(k) \tag{15}$$

$$e_1(k,i)=C^T\overline{A}_i b+L^T(k)(A^i)^T DA_i b_i+b_i^T(\overline{A}_i)^T DA^i L(k)$$

$$e_2(k,i)=b_i^T\overline{A}_i^T D\overline{A}_i b_i$$

$$\overline{A}_i=A^{i-1}+\cdots+A+I \tag{16}$$

Let $Y_m(k+1)=[y_m(k+1),\cdots,y_m(k+P)]^T$, then we have

$$Y_m(k+1) = E_0 + E_1 u(k) + E_2 u^2(k) \tag{17}$$

here,

$$E_j(k) = \left[e_j(k,1), \cdots, e_j(k,p)\right]^T \ (j = 0,1,2) \tag{18}$$

Now import **output amending mechanism** and let

$$Y_P(k+1) = Y_m(k+1) + [1,\cdots,1]_{P\times 1}^T (y(k) - y_m(k)) \tag{19}$$

and $Y_r(k+1) = [y_r(k+1), \cdots, y_r(k+P)]^T$, here $y_r(k+i) = \alpha^i y(k) + (1-\alpha^i)W(k)$, α is soften factor, $W(k)$ is set point. Let, $u(k) = u(k-1) + \Delta u(k)$.

The optimization control index is

$$J = \left\|Y_r(k+1) - Y_P(k+1)\right\|^2 + r\Delta u^2(k) \tag{20}$$

Let $Y_e(k+1) = E_0(k) + E_1(k)u(k-1) + E_2(k)u^2(k-1) + [1,\cdots,1]_{P\times 1}^T (y(k) - y_m(k))$, $Y_a(k+1) = Y_r(k+1) - Y_e(k+1)$, $E_r(k) = E_1 + 2E_2 u(k-1)$ then

$$\frac{\partial J}{\partial \Delta u(k)} = s_0(k) + s_1(k)\Delta u(k) + s_2(k)\Delta u^2(k) + s_3(k)\Delta u^3 \tag{21}$$

$$s_0(k) = -E_r^{\ T}(k)Y_a(k+1) - Y_a^{\ T}(k+1)E_r(k) \tag{22}$$

$$s_1(k) = 2[E_r^T(k)E_r(k) - E_2^T(k)Y_a(k+1) - Y_a^T(k+1)E_2(k) + r] \tag{23}$$

$$s_2(k) = 3[E_r^{\ T}(k)E_2(k) + E_2^{\ T}(k)E_r(k)] \tag{24}$$

$$s_3(k) = 4E_2^{\ T}(k)E_2(k) \tag{25}$$

Apply Newton method to solve the equation $\frac{\partial J}{\partial \Delta u(k)} = 0$ (the Root function of MATLAB), and then can get control law $\Delta u(k)$ which minimizes J. Noted that the order of equation (21) is an odd number, equation (21) must have a real root at least.

From equations (10) and (13), we have that the output of the nonlinear Laguerre model is

$$y_m(k)=\theta^T\phi(k) \tag{26}$$

here

$$\theta^T=[c_0,\cdots,c_N,c_{11},c_{12},\cdots,c_{22},\cdots,c_{NN}], \tag{27}$$
$$\phi^T(k)=[1,l_1,\cdots,l_N,l_1^2,l_1l_2,\cdots,l_2^2,\cdots,l_N^2]$$

This is a linear regression formulation, so θ^T can be identified by RLS online.

$$P(k)=\frac{1}{\lambda}\left[I-\frac{P(k-1)\phi(k)\phi^T(k)}{\lambda+\phi^T(k)P(k-1)\phi(k)}\right]P(k-1) \tag{28}$$

$$\hat{\theta}(k)=\hat{\theta}(k-1)+P(k)\phi(k)[y(k)-\phi^T(k)\hat{\theta}(k-1)] \tag{29}$$

Now, we **extend the control horizon** from 1 to 2, then the equation (10) is improved into

$$L(k+1)=\begin{bmatrix} A & \bar{A}_1B \\ A^2 & \bar{A}_2B \\ \vdots & \vdots \\ A^{P-1} & \bar{A}_{P-1}B \end{bmatrix}\begin{bmatrix} L^T(k) & u(k-1)\end{bmatrix}^T+\begin{bmatrix} \bar{A}_1B & 0 \\ \bar{A}_2B & \bar{A}_1 \\ \vdots & \vdots \\ \bar{A}_{P-1}B & \bar{A}_{P-2} \end{bmatrix}\Delta U(k) \tag{30}$$

Accordingly, the optimization index (20) is changed into

$$J=\left\|Y_r(k+1)-Y_P(k+1)\right\|^2+\left\|\Delta U(k)\right\|_R^2 \tag{31}$$

here, $\Delta U(k)=[\Delta u(k),\Delta u(k+1)]^T$, $R=rI_{2\times2}$ (I is a unit matrix)

From $\frac{\partial J}{\partial\Delta U(k)}=0$, the equation (21) is changed into

$$\begin{aligned} &s_{m30}\Delta u^3(k)+s_{m20}\Delta u^2(k)+s_{m10}\Delta u(k)+s_{m00}+s_{m21}\Delta u^2(k)\Delta u(k+1)+s_{m01}\Delta u(k+1)+ \\ &s_{m11}\Delta u(k)\Delta u(k+1)+s_{m02}\Delta u^2(k+1)+s_{m12}\Delta u(k)\Delta u^2(k+1)+s_{m03}\Delta u^3(k+1)=0 \end{aligned} \tag{32}$$

here, the polynomial equation coefficients $s_{mij}(m=1,2;0\le i\le3;0\le j\le3)$ are determined by $Y_r(k+1)$, $Y_P(k+1)$, $L(k),u(k-1)$ and R.

Lemma 1[1,7] Gröbner basis function can be used to solve the following function

$$f_1(\Delta u(k),\Delta u(k+1))=f_2(\Delta u(k),\Delta u(k+1)) \tag{33}$$

here f_1,f_2 is the polynomial defined in the ring $Q^2(\Delta u(k),\Delta u(k+1))$. ■

Lemma 2[1,7] Assume $k[x_1,\cdots,x_n]$ is a polynomial ring, if $f_1,\cdots,f_n \in k[x_1,\cdots,x_n]$, then $\langle f_1,\cdots,f_n \rangle$ is an ideal of $k[x_1,\cdots,x_n]$, we defined it as the ideal caused by $f_1,\cdots,f_n$. ■

Theorem 4[1] (Buchberger algorithm) let $\Im = \langle f_1,\cdots,f_n \rangle \neq \{0\}$ is a polynomial ideal, then a Gröbner basis of $\Im$ can be constructed in finite steps. ■

From Lemma1,2 and Theorem4, we can use Buchberger algorithm to calculate Gröbner basis, and then can solve the equation (32), the advantage of this algorithm is the ability to confirm the maximum order which can be reduced. Therefore, repeating Buchberger algorithm in this way, then, (32) can be rewritten as

$$v_{19}\Delta u^9(k) + \cdots + v_{11}\Delta u(k) + v_{10} = 0 \tag{34}$$

$$\Delta u(k+1) = \frac{v_{20}(\Delta u(k))}{v_{21}(\Delta u(k))} \tag{35}$$

here, the coefficients $v_{1i}, 0 \le i \le 9$ are determined by $Y_r(k+1), Y_P(k+1), L(k)$, $u(k-1)$ and R, while v_{20}, v_{21} are determined by by $Y_r(k+1), Y_P(k+1), L(k)$, $u(k-1)$, R and $\Delta u(k)$. Then we can use the root function of MATLAB online to compute the real root as the incremental model control law. The detailed deduction of (34), (35) can be referred to reference [7].

4 Experiment Results

As Fig.1 shown, this is an water output pressure control system, the executer is the Siemens MM440-typed converter, and the controlled plant is the Grundfos CHI-2 typed centrifugal pump (see the subfigure 2 of Fig.1). The signal flow is: the pressure sensor (see the subfigure 3 of Fig.1) transfers the output pressure into standard voltage signal (1-5V), and after A/D transmission, the signal is sent to PC serial port. The controller in PC calculates the control signal. After D/A transmission, the control signal is transferred into standard current signal (4-20mA), and implemented in the Siemens converter (see the subfigure 1 of Fig.1) to adjust the pumping frequency of Grundfos centrifugal pump.

In wide working area, the mechanism of this pump determined that it has severe nonlinearity. As fig.2 shown, the relationships of the output pressure and the output water current under different rotating frequencies show server nonlinearity. So this pressure control system is a severe nonlinear one. However, in stable working area, this system fulfills the 4 assumptions of **theoreml**, so it can be decomposed to a set of dynamic linear block and a memoryless nonlinear block. Consequently, we can use our proposed algorithm to solve this problem.

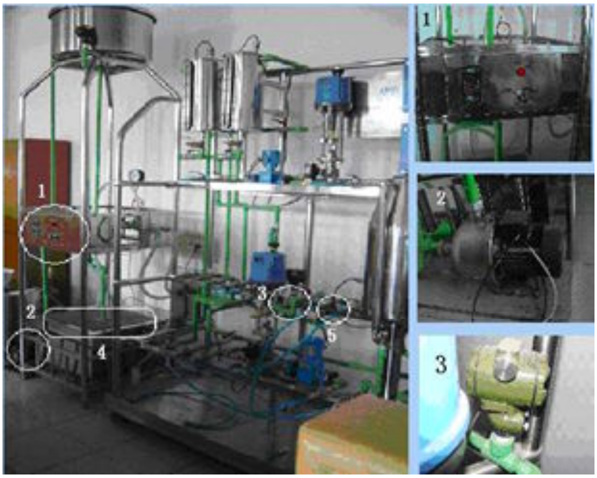


Fig. 1. Water pressure control system

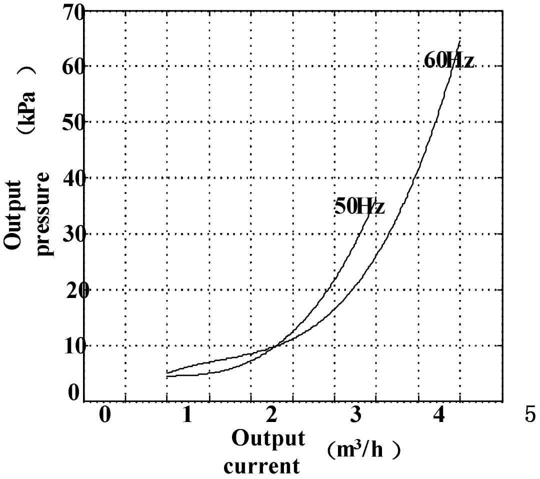


Fig. 2. Characteristic curves of Grundfos CHI-2 typed centrifugal pump

The control performances of the two nonlinear Laguerre predictive control algorithms whose control horizon is 1 (traditional algorithm) or 2(our improved algorithm) to trace 30.0 Kpa , 35.6 Kpa and 43.4 Kpa are shown in Fig.3 and Fig.4 respectively. In the 160th sampling period (sampling period is 1 second), the opening of hand valve (see the subfigure 5 of Fig.1) is increased by 20%, which imports an outside disturbance to test the robustness. The parameters of the traditional algorithm are: prediction horizon $P=6$, control horizon $M=1$, Laguerre series order $N=7$, time scaling factor $p=1.5$, forgetting factor $\lambda=0.7$, control weighting factor $r = 0.9$; $\alpha = 0.2$ the parameters of the improved algorithm are: $P = 6$, $M = 2$, $N = 7$, $p = 1.5$, $T = 2$, $\lambda = 0.2$, $\alpha = 0.2$, $r = 0.9$. From Fig.3 and Fig.4, we can conclude that these two algorithms both have certain robustness, and can eliminate steady-state error. The differences are that the improved algorithm can effectively decrease the vibrations of the control and controlled variables when compared with the traditional one. Besides, the overshooting is decreased remarkably under outside disturbance by our improved control algorithm. Thus, the control results validate the superiority of this proposed algorithm.

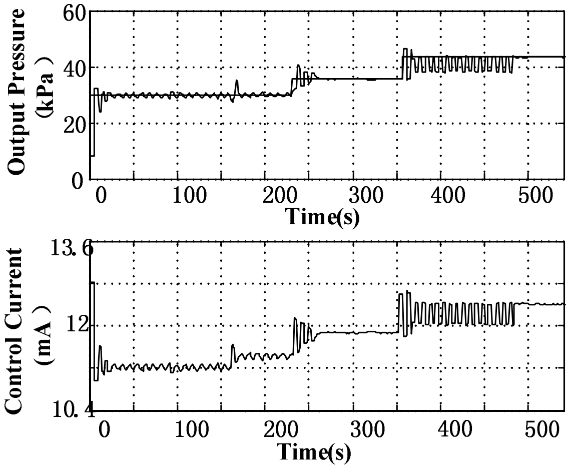


Fig. 3. Control performances of traditional control algorithm

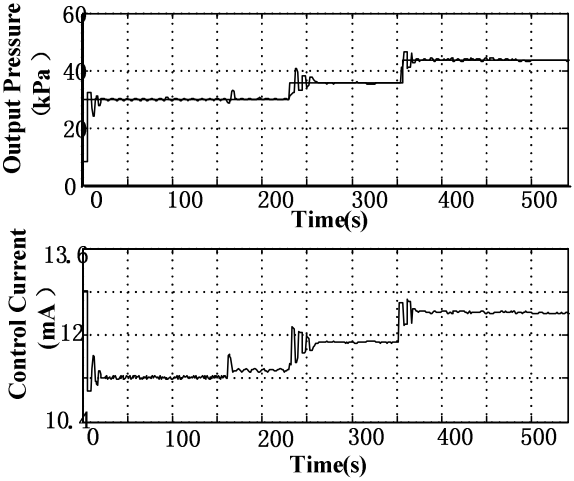


Fig. 4. Control performances of our improved control algorithm

5 Conclusion

This paper extends the control horizon of the adaptive control algorithm based on OFS-Volterra hybrid model from 1 to 2, and imports output anti-disturbance mechanism combined with online RLS OFS spectrum coefficients identification mechanism. In contrast to the traditional control algorithm whose control horizon is 1, this improved algorithm can effectively enhance the ability to solve server nonlinear plants, greatly decreases the vibrations and overshootings of the control and controlled variables. Besides, this algorithm also shows good robustness for outside disturbance, so is suitable for the design of complex industrial process controllers.

Acknowledgement

This paper is funded by Chinese National '985' High Level University Construction Fund (KY2701), and Hefei Key Scientific Research Project Fund (He Ke 2002-45).

References

1. Buchberger,B. Grobner bases: An algorithmic method in polynominal ideal theory. In N.K.Bose, editor, Multidimensional Systems Theory, D.Reidel Publishing Company, Dorderrecht, (1985) 184-232
2. Dumont,G.A., Zervos, C.C., et al.: Laguerre-based adaptive control of PH in bleach plant extraction stage. Automatica, 26(4), (1990) 781-787.
3. Li, S.F., Li,Y.Q., Xu. Z.F., Chen, Z.B.,: An extension to Laguerre model adaptive predictive control algorithm, Journal of China University of Science and Technology, 31(1), (2001) 92-98
4. Marmarelis V.Z., et al.: Volterra models and three-layer perceptrons, IEEE Transaction on neural networks, 8(6), (1997) 1421-1430
5. Michael, A.H., Dale, E.S.: Nonlinear Process Control, Prentice Hall PTR, New Jersey, (1997), 40-41
6. Parker, R.S., et al.: Nonlinear model predictive control of continuous bioreactor at near optimum conditions, Proceedings of American Control Conference, Philadelphia, Pennsylvania, (1998) 2549-2553
7. Parker, R.S., Nonlinear model predictive control of a continuous bioreactor using approximate data-driven models. Proceedings of automatic control conference. Anchorage, AK May 8-10, (2002) 2885-2890
8. Sbarbaro,D., Bassi,D.: A nonlinear controller based on self-organizing maps, IEEE,(1995) 1774-1777
9. Seretis, C, et al.: Nonlinear dynamical system identification using reduced Volterra models with generalized orthonormal basis functions. Proceedings of American Control Conference, Albuquerque, New Mexico, June (1997) 3042-3046
10. Wang, L.P.: Discrete model predictive controller design using Laguerre functions, Journal of Process Control, 14(2), (2004) 131-142
11. Zhang, H.T., Xu, Z.F., Li, S.F.: An adaptive predictive control based on Laguerre Function Model for diffusion furnace, System Engineering and Electronics 24(4),(2002) 54-57
12. Zhang, H.T., Chen Z.H., Qin, T.: An adaptive predictive control strategy for water recycling irrigation system of cropper, Journal of Information and Control of China 32(7), (2003) 586-590

Profiling Multiple Domains of User Interests and Using Them for Personalized Web Support

Hyung Joon Kook

Department of Computer Engineering, Sejong University, Seoul, Korea
kook@sejong.ac.kr

Abstract. As people's web usage is growing bigger, personalized support for web browsing is in great demand. Furthermore, the diversity of a user's interests demands an appropriate methodology for profiling multiple domains of user interests. To comply with such demands, we propose one feasible design approach to support personalized web usage, in which a *web user agent* takes over the task of learning and profiling the multiplicity and the changeability of user interests. To evaluate the advantages of this approach, we have constructed a personalized web supporting system, in which an autonomous agent, namely the *web guide agent*, utilizes the information gathered by the web user agent for adaptation, e.g., selective retrieval and re-ranking of web links, and automatic delivery of specific web pages. Compared to other design alternatives, the proposed scheme is operationally simple, while producing acceptably reliable outcome.

1 Introduction and Overview

With the growing web activities of users, the web navigation supports adapted to each person are increasingly on demand. However, in adaptive web research, user adaptation has been considered a hard problem due to the ill structuredness and the diversity involved in the user's web navigation behaviors. Some notable techniques developed in this area include employing machine learning and adaptive agents to support personalized web browsing [1,7,9]. Since it is hard to adapt to a user without proper modeling of the user, previous studies in this area have concentrated on building an adequate computational framework for learning and profiling user interests [3,4,5,6]. It is only very recent that research in this area began paying attention to support of personalized browsing and searching based on user's multiple interests. WebMate [2] and Web Personae [8] track the user's web activities to recognize changes in the user's interests and to update the user's multiple profiles or interest vectors, which are utilized for further navigation and search supports.

With the goal of building an adaptive web support system, this research proposes one feasible approach to profile users' multiple interests and to use the information for guiding the user's web navigation. As an important step to the research goal, this paper first focuses on the design of a *web user agent,* which specializes in the tracking and modeling of the user's changing interests while the user browses and searches the

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 512 – 520, 2005.

web. Then, we present an implementation of a navigation support system, in which a *web guide agent* uses the information gathered by the web user agent for supporting the user's web browsing activities.

This particular research aim is similar to that of WebMate or Web Personae, but we take a different approach to achieve the same goal. The agents in our research may be characterized as *reliable* enough to provide an adequate level of personalization, and *simple* enough to execute at an acceptable level of performance. In the following section, we present the information structure of the user profile, the web user agent's operational mechanism and discussions on some design issues. In the next section, we will explain how the user profile is used by the web guide agent to provide a personalized web navigation support. In the final section, a summary is given.

2 User Profile and The Web User Agent

It is not unusual that people have multiple interests in several domains (e.g., bears, cars, weather, geography), and even within a domain, they often have more detailed domains of interests (e.g., glacier, monsoon, hurricane, Greenland). In modeling such multiple interests in their web activities, a simple but feasible way to represent each domain is using a word vector, i.e., a set of word-weight pairs, where the words are the keywords from the web pages relevant to a user's interests and the weights are the relative importance of corresponding words. Since this vector represents a user's interest in a domain, we call it an *interest vector* to distinguish it from a *term vector* retrieved from a web page. Initially, the profile of a user is empty, i.e., contains no interest vector. As the user browses and searches web continuously, his interests are recognized and updated incrementally in the form of multiple interest vectors in his profile. An example of an interest vector is shown below. It consists of the number of updates, the date and time of last update, the number of keywords, and the keyword-weight pairs in descending order of weights.

```
(INTEREST VECTOR  IV25
    (NUMBER_OF_UPDATES  12)
    (LAST_UPDATE  2005 04 22 09:50)
    (NUMBER_OF_WORDS  6)
    (VECTOR
          (bear 37) (polar 19) (weather 13) (life 12) (food 9) (greenland 7))
)
```

During a user's web browsing, the *web user agent* tracks his/her browsing activity; an agent specialized for the task of recognizing the user's interests and learning the profile. The task is performed as follows. When a web page is known to be interesting to the user (e.g. by a positive feedback), the agent retrieves the term vector of the web page. A term vector is a set of word-weight pairs listed in the descending order of weights, where the words are the keywords occurring in the web page and the weights are the frequencies of corresponding words. A term vector may either be computed in real-time or be obtained from other sources, e.g., a separate agent specialized for web page analysis task (e.g. using TF-IDF). The user's interests may also be recognized from a search query, for which the agent creates a term vector with weights distributed evenly over the words in the query.

Once a term vector is obtained, the agent compares it with each interest vector in the user profile to find the one with the greatest similarity that exceeds predefined threshold. If such an interest vector is found, it is updated by merging with the term vector, which reflects further specification of that particular domain of interest. If none of the interest vectors has a greater similarity than threshold, a new interest vector is copied from the term vector and added to the profile, which reflects the expansion of the user's domains of interests. As a final touch, the interest vector is subjected to a trimming procedure, i.e., a procedure for selecting more relevant keywords to keep. Then, the interest vector is set as the *current* interest vector to serve as the basis for context-sensitive adaptation. The following two subsections give details on the evaluation of similarity and the vector trimming procedure.

2.1 Evaluation of Similarity

For a term vector obtained, the web user agent decides whether to use it to update an existing interest vector or to use it to create a new vector. The decision depends on how similar it is to an interest vector, which is determined by the *similarity*, which is a measure of how much a term vector is similar to an interest vector. It is computed as follows.

$$\begin{aligned} \mathit{Similarity} = &\sum(\mathit{weight\ of\ words\ in}\ \{T_i \cap T_t\}) \\ &- \sum(\mathit{weight\ of\ words\ in}\ \{T_i \cup T_t\} - \{T_i \cap T_t\}) \end{aligned}$$

where T_i = *words in interest vector whose weights are over* $W_{i\text{-}avg}$,
T_t = *words in term vector whose weights are over* $W_{t\text{-}avg}$,
$W_{i\text{-}avg}$ = *average weight of the words in interest vector, and*
$W_{t\text{-}avg}$ = *average weight of the words in term vector.*

According to the equation above, the computation of the similarity begins with retrieval of the words with over-average weights from each vector. Then the sum of the weights of the common words (i.e., the words occurring in both vectors) and the sum of the weights of the non-common words are computed, respectively. The similarity is obtained by subtracting the latter from the former.

Since evaluating similarity between a term vector and an interest vector is a subjective task, there may be various design alternatives to the proposed method, as summarized below.

Method 1 Take the weights of all, rather than the over-average words in each vector.

Method 2 Subtract the average weight of non-common words from the average weight of common words.

Method 3 Multiply the sum by the number of words.

Of the three alternatives above, Method 1 suggests, rather than to evaluate on the basis of only those words with over-average weights, to base the calculation on the weights of *all* words in each vector so that the similarity is computed by subtracting the sum of the weights of all non-common words from that of all common words. In case when there exists a common word whose weight is high in one vector but fairly low in the other, however, this method is likely to give a distorted similarity biased to similar. Another design alternative is to take the size of the vectors, i.e., the number of words in each vector, into consideration. There are two ways to do this. Method 2 is one way: It subtracts the average weight of the non-common words from the average weight of the common words. But this would be unnatural because, the less there are common words, or equivalently, the more there are non-common words, the similarity would be inflated more. Method 3 is the other way of employing the vector size: It multiplies the sum by the number of words. But then, the words with low weights would influence the computation of the similarity to a considerable extent. Therefore, with the reservation that the choice of the best alternative still remains a research issue, we expect that the proposed method is a close-to-optimal choice for evaluating the similarity while minimizing bias and unnaturalness.

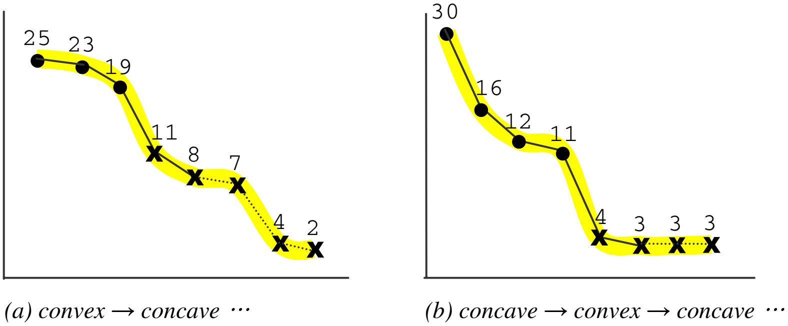
2.2 Vector Trimming Procedure

Literally merging two vectors whenever an update is performed would result in the growth of some vectors to unnecessarily large size. To avoid this, in the last stage of updating or creating an interest vector, the agent selects only those words with relatively high weights from the keywords in the vector. Doing so keeps the vector size from growing infinitely by eliminating relatively less relevant keywords from the vector.

Difficulty arises from two conflicting demands; one for making the list short enough to reflect the concentration of the user's interest by eliminating as many less relevant keywords as possible, and the other for making the list long enough to accommodate the diversity of the user's interest by keeping as many keywords as possible. To resolve this conflict, we employed a simple but acceptable heuristic. By examining the value changes in the weights of the keywords arranged in non-ascending order of their weights, the agent detects the first point where a weight value drops drastically, then it trims the vector at that point, i.e., the keywords thereafter are eliminated from the interest vector.

Speaking in terms of a 2D-plane plot where keywords are plotted in non-ascending order of weights, the first concave point is detected, and the keywords in the preceding, convex section are taken as valid. In applying this selection procedure, there are a few exceptional cases that demand special handling. First, if the plot never becomes concave, i.e., convex throughout, all keywords are taken as valid. A second exceptional case arises when the plot starts concave from the beginning. If we apply the general selection procedure described above, only the first keyword will be selected as valid, failing to model the user's diverse interests in this domain. To fix it, all keywords in the first concave section are taken as valid and the selection procedure resumes at the following section of the plot. Third, if the plot is entirely concave, i.e., has no convex section, all keywords are taken as valid. Fourth, a straight section is taken to be a continuation of immediately preceding (convex or concave) section. The last exceptional case is when the plot starts straight. In this case, all keywords in the straight section are taken as valid and the selection procedure resumes at the following section.

In Figure 1, example plots of two interest vectors are shown with different sets of weights. The points in the figure denote the keywords and the values denote their weights. In both plots, the selection procedure detects the drastic drop point by examining the sections indicated as solid line segments, and selects the keywords marked '●', leaving the keywords marked '**X**' unselected. Note that, once the drastic drop point is found, the remaining sections (indicated as dotted line segments) are never examined, thus saving processing time.



(a) convex → concave ⋯ *(b) concave → convex → concave ⋯*

Fig. 1. Vector Trimming Examples

Splitting the keywords into two parts of, namely, more important ones and less important ones is also a subjective matter. Therefore, various design alternatives are possible. We have experimented with a few of them, as summarized below.

Method 1 Select as many words as the average number of the words in two vectors.

Method 2 Select words in proportion to the update count.

Method 3 Select all common words regardless of their weights.

Method 4 select words up to the point where the weight sum reaches a limit, e.g., 100.

Method 5 Select words up to the most drastic drop point.

Method 6 Select words up to the first drastic drop point.

Of the six alternatives, Method 1 seems a fair, balanced method, since it takes the average of the two. But the problem is that the words selected as valid by this method may include those words that are not relevant to user's interest. Method 2 is proposing to place more reliability on more frequently updated interest vectors, allowing more words in them. But the problem of limiting the growth of the vector size still remains to be solved. The problem with Method 3 is that if common words with low weights are preserved in an interest vector, the vector's reliability is damaged and its size will have to be treated, as it gets larger. Method 4 can contribute to computing efficiency.

As in Methods 1-3, however, by depending mainly on arithmetic aspect, this method is likely to produce a result far from human intuition. Method 5 is directly comparable to the method we adopted. This method splits the words into two groups, i.e., more important ones and less important ones, and it selects the first group. Although this method is likely to produce the most reliable results, it takes more time to compute the most drastic weight drop point because all drop points must be examined. In contrast, Method 6, the method we propose, splits the words at the first drastic drop point, thus saving time for examining the drop points in the remaining parts of the vector. Overall, considering tradeoffs between reliability and performance, the proposed design may be regarded as an optimal choice.

3 Web Navigation Support by the Web Guide Agent

Manually browsing webs through direct interaction with the system often puts an enormous cognitive burden on the user, and sometimes reduces motivation for and interest in web browsing. In the system we propose, an automated mechanism is provided for personalized web navigation support. This is performed by an independent agent, the *web guide agent*. The web guide agent is activated and deactivated by user's request (by clicking on the "Enable/Disable Agent" button on the browser window). Once activated, the agent starts guiding the user's web navigation, performing as a background process, i.e. independently of the user's normal web browsing behavior.

3.1 The Web Guide Agent's Working Mechanism

The user preference information gathered in the user profile is used for adapting to the user's web navigation activities. Adaptation is performed in a couple of modes, namely *link suggestion* and *automatic recommendation of web pages*. Link suggestion is made when responding to the user's keyword search. Given a search result obtained in the form of a list of links, the agent reorganizes it to fit to the user's interest. The reorganization involves re-ordering of links, the elimination of less interesting links, etc.

While the link suggestion mechanism reflect the user preference to some extent, this method is still somewhat inconvenient for some users who are young-aged or less skilled in the use of the computers. For such users, a more automated form of navigation support would be desirable. To do this, the web guide agent examines the web pages in the suggested links to select one that is recommendable to the user. The selection is based on the information about the current web page the user is browsing, the current interest vector, and the information about the web pages in the suggested links. In the deactivated state, the agent stops the selection process, and it only maintains the base information for the selection, in preparation for the next activation. There are four occasions when the activated agent resumes its selection process. They are: (1) when it is invoked, (2) when the current interest vector is changed, (3) when the current web page is changed, and (4) when the list of the suggested links is changed. In other words, the agent's selection process is usually started by the first occasion (1), and resumed by any of the latter three occasions.

The selection process proceeds in the following way. First, the agent computes the *current affinity*, a relationship measure between the current web page and the current interest vector. After that, the term vectors from the web pages in the list of suggested links are extracted in turn, and the affinity between each term vector and the current interest vector is computed. If a term vector is found to have a higher affinity than the current affinity, the corresponding web page is recommended to the user. After updating the current affinity to the higher value just found, it continues the selection process for the remaining web pages in the suggested links. The process is repeated until the web page with the highest affinity is found and recommended. As described, the recommendation is made on the web page with an affinity higher than the current affinity, *not* the one with the highest affinity. On behalf of the user, this strategy contributes to minimize the agent's response time. The affinity is computed as follows:

$$Affinity = \sum(W_i \times min(W_{i,,}\ W_t\ )) - \sum(W)$$

where

(1) $W_{i,}$ *and* W_t *are the weights in the current interest vector and the weight in the term vector, respectively, of t in* $\{T_i \cap T_t\}$*,*

(2) *W is the weight of t in* $\{T_i\ \cup T_t\} - \{T_i \cap T_t\}$*,*

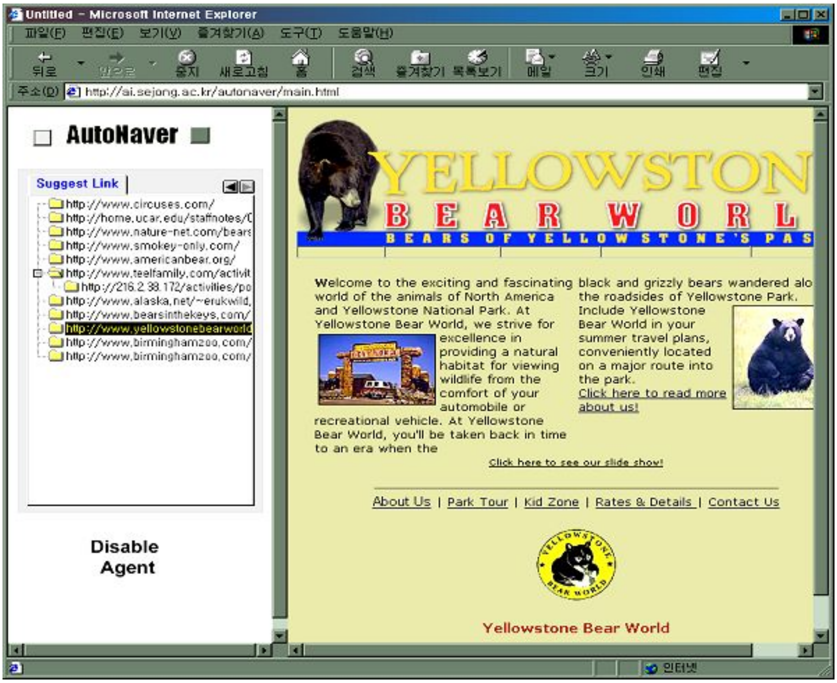
(3) T_i *are the words in the current interest vector, and*

(4) T_t *are the words in the term vector.*

In other words, the affinity between a web page and the user interests is computed in the following way. First, for each pair of the words that are common in the web page's term vector and the current interest vector, their weight products are accumulated. The affinity is obtained by subtracting the weight sum of the non-common words from the accumulated sum. There is one thing to note in the multiplication of weights. If a word's weight in the term vector is greater than the same word's weight in the current interest vector, the smaller weight is applied for multiplication. This is to avoid the inflation of the affinity by a common word, which has a high weight in the web page but not as much high weight in the current interest vector. In contrast, if a common word has a lower weight in the term vector than in the current interest vector, no adjustments are made to both weights for multiplication.

3.2 Delivery of a Web Page

When a web page is selected for recommendation, the web guide agent draws attention of the user by displaying progressive flashing bars in the corner of the browser window, indicating that a recommendation has been made and will be delivered onto the browsing window in a few seconds (5 seconds interval is given in the present implementation). At this point the user has two options, "disable-agent" or "do-nothing". If the user doesn't want the recommendation to be delivered for some reason (e.g., he



http://ai.sejong.ac.kr/autonaver/main.html

http://www.circuses.com/
http://home.ucar.edu/staffnotes/C
http://www.nature-net.com/bears
http://www.smokey-only.com/
http://www.americanbear.org/
http://www.teelfamily.com/activit
http://216.2.38.172/activities/po
http://www.alaska.net/~erukwild,
http://www.bearsinthekeys.com/
http://www.yellowstonebearworld
http://www.birminghamzoo.com/
http://www.birminghamzoo.com/

Welcome to the exciting and fascinating world of the animals of North America and Yellowstone National Park. At Yellowstone Bear World, we strive for excellence in providing a natural habitat for viewing wildlife from the comfort of your automobile or recreational vehicle. At Yellowstone Bear World, you'll be taken back in time to an era when the black and grizzly bears wandered alo the roadsides of Yellowstone Park. Include Yellowstone Bear World in your summer travel plans, conveniently located on a major route into the park.
Click here to read more about us!

Click here to see our slide show!

About Us | Park Tour | Kid Zone | Rates & Details | Contact Us

Fig. 2. A Recommended Web Page Delivered

may want to stay with the current web page), he simply clicks on the flashing bars to deactivate the agent. Otherwise, the current web page is replaced by the recommended web page immediately after the preset time delay is over. Figure 2 shows the navigation window when a recommended web page has just been delivered. Since there is some preset time delay (e.g., 5 seconds) for the user to decide, it is possible that the agent, during this time interval, may locate one or more web pages with higher affinity than that of the web page it is just about to deliver. In this case, the agent revises its recommendation to the last web page, i.e., the one with the highest affinity among the web pages examined during the time interval.

4 Summary

In this paper, we have proposed a simple, but acceptably reliable design principle for supporting personalized web surfing. User interests in multiple domains are represented as multiple interest vectors in the user profile. Changing, widening and deepening of a user's interests can be reflected in his profile by a web user agent in an incremental and continuous way. We have tested the appropriateness of the proposed design in an adaptive web navigation support system, in which the web guide agent assists the user's navigation in the background. It responds to a user's search query with a list of web page links, filtered and re-ranked on the basis of the user interests. A more active and automatic mode of adaptation is also supported by the agent, by using the user profile for making automatic delivery of specific web pages that seem the most relevant to user's current interests.

References

1. Armstrong, R., Freitag, D., Joachims, T., Mitchell, T.: WebWatcher: A Tour Guide for the World Wide Web. In IJCAI-97 (1997)
2. Chen, L., Sycara, K.: WebMate: A Personal Agent for Browsing and Searching. In Agents-98 (1998)
3. Chin, D.: Intelligent Interfaces as Agents. In Sullivan, J. W., Tyler, S. W.: Intelligent User Interfaces. Academic Press (1991), 177-206
4. Cook, R., Kay, J.: The Justified User Model: A Viewable, Explained User Model. In UM-94, (1994), 145-150
5. Dieterich, H., Malinowski, U., Kuhme. T., Schneider-Hufschmidt, M.: The State of Art in Adaptive User Interfaces. In Schneider-Hufschmidt, M., Kuhme, T., Malinowski, U. (eds.): Adaptive User Interfaces: Principles and Practice. North Holland (1993), 269-283
6. Hayes-Roth, B.: An Architecture for Adaptive Intelligent Systems. In Artificial Intelligence (1995), 72(1/2), 329-365
7. Maes, P.: Agent That Reduce Work and Information Overload. In CACM (2003), 37 no.7
8. McGowan, J. P., Kushmerick, N., Smyth, B.: What Do You Want to Be Today? Web Personae for Personalized Information Access. In Proceedings of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (2002)
9. Pazzani, M., Muramatsu, J., Billsus, D.: Syskill & Webert: Identifying Interesting Web Sites. In AAAI-96 (1996)

Analysis of SCTP Handover by Movement Patterns[+]

Dong Phil Kim, Seok Joo Koh, and Sang Wook Kim

Department of Computer Science, Kyungpook National University, Korea
{dpkim, sjkoh, swkim}@cs.knu.ac.kr

Abstract. Stream Control Transmission Protocol (SCTP) is a new end-to-end transport protocol, which can be used to support the mobility of mobile terminals. This paper describes a framework of SCTP handover and analyzes the handover latency for the single-homing mobile terminals. We then show the experimental results of the SCTP handover in terms of the handover latency and throughput for the two different movement patterns: linear and crossover patterns. For the linear movement pattern, it is shown that the SCTP handover latency may severely depend on the handover delay at the underlying link layer. In the case of the crossover movement pattern, we see that the throughput of the data transmission could be degraded, as the crossover movements occur more frequently.

1 Introduction

As the wireless access technology gets rapidly improved, the demand of mobility support has been more increased. With this trend, some protocols to support the mobility have so far been focused, which include Mobile IP (MIP) [1] and Session Initiation Protocol (SIP) [2] and Stream Control Transmission Protocol (SCTP) [3]. In particular, it is noted that the SCTP can be used to support the soft handover for mobile terminals with the help of the SCTP multi-homing feature [4].

In this paper, we analyze the SCTP handover algorithm in terms of handover latency. We perform the experimental analysis of the SCTP handover over Linux platforms for the single-homing mobile terminals (with a single network interface). In particular, we compare the performance of the SCTP handover for the two different movement patterns of the mobile terminals: linear and crossover patterns.

This paper is organized as follows. Section 2 describes an overview of SCTP handover and briefly compares the existing mobility protocols. In Section 3, we analyze the latency of SCTP handover theoretically. Section 4 describes some experimental results of SCTP handover that have been performed on the Linux platforms. Section 5 concludes this paper.

2 SCTP Handover

This section describes SCTP handover and compares the existing mobility protocols.

[+] This research was supported by the MIC, Korea, under the ITRC support program supervised by the IITA.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 521 – 529, 2005.

2.1 SCTP Handover Mechanism

Stream Control Transmission Protocol (SCTP) as defined in IETF RFC 2960 [3] is an end-to-end, connection-oriented transport layer protocol, next to TCP and UDP. The SCTP is featured by 'multi-streaming' and 'multi-homing'. In particular, the multihoming feature of SCTP can be used to provide the handover capability for the mobile terminals (MT) by adding a new IP address and deleting the old IP address during the active session [5].

Figure 1 sketches the SCTP handover for a mobile terminal (MT) between two different IP networks, where the MT is moving from Base Station (BS) A to B.

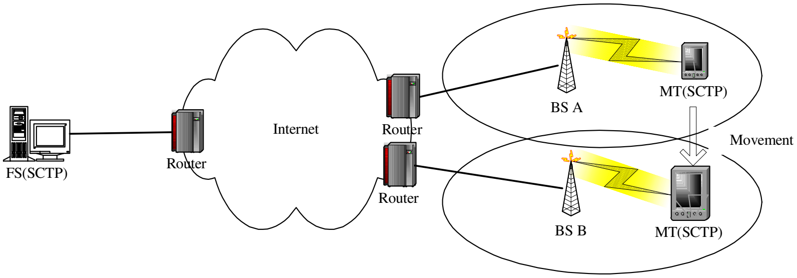


Fig. 1. SCTP Handover

In the figure, we assume that an MT initiates an SCTP association with a Fixed Server (FS). For the SCTP association, FS has 'IP address 1', whereas MT uses 'IP address 2'. Then, the overall SCTP handover procedures could be performed as follows:

(1) When the MT moves from BS A toward BS B, now it is in the overlapping region. In this phase, the MT obtains a new address 'IP address 3' from the BS B by using an address configuration scheme such as Dynamic Host Configuration Protocol (DHCP).

(2) The newly obtained IP address 3 will be informed by MT to FS in the tranport layer. This is done by sending an SCTP Address Configuration (ASCONF) chunk to FS. The MT receives the responding ASCONF-ACK chunk from the FS. This is called the 'Add-IP' operation.

(3) The MT is now in a dual homing state. The old IP address is still used as the primary address, until the new IP address 3 is set to be the "Primary Address" by the MT. Before the new primary address is set, IP address 3 is used as a backup path.

(4) As the MT further continues to move towards BS B, it needs to change the new IP address into its primary IP address according to an appropriate rule. Once the primary address is changed, the FS sends the outgoing data packets over the new primary IP address of MT (IP address 3). This is called the 'Primary-Change' operation.

(5) As the MT progresses to move toward B, it will delete the old IP address from the association. This is called the 'Delete-IP' operation.

The procedural steps described above will be repeated each time the MT moves to a new BS.

2.2 Comparison of the Existing Mobility Protocols

In this chapter, we review the existing IP-based mobility protocols that have so far been developed. More especially, we analyze and compare SCTP with the two existing protocols for supporting mobility: Mobile IP and SIP.

A. Mobile IP

Mobile IP (MIP) is a well-known protocol used to support IP mobility, which was standardized in the IETF. As per the associated IP version, MIP could be divided into MIPv4 [6] and MIPv6 [7]. MIP has been developed to support seamless Internet service against any change of the IP address, when the MT progresses toward the new IP region.

In this paper, we focus on the MIPv4 scheme. The specification of MIPv4 describes the protocol operations between the following entities: Mobile Terminal (MT), Home Agent (HA), Foreign Agent (FA), and Correspondent Node (CN). The basic protocol operations of MIPv4 are done as follows:

Step 1. When a MT moves into a new subnet, it can be assigned the CoA (Care of Address) from FA such as a router. The CoA could be the CoA of FA (IP address of FA) or the Co-located CoA (e.g., obtainted by Dynamic Host Configuration Protocol).
Step 2. After that, the MT registers its CoA with the HA.
Step 3. If the HA receives data packets destined for the MT from the CN, the HA will intercept these packets and forward them to the MT by using the Mobile IP tunneling.
Step 4. The FA (in case of using CoA of FA) will de-capsulate the packets received from HA and then deliver the original packets to the MT.

The basic specification of MIP cannot support fast handover for time-critical and loss-sensitive applications. To address this problem, some extensions of MIP are being developed in the IETF, such as Fast handover for MIP [8, 9] and Hierarchical MIP [10].

B. Session Initiation Protocol

The Session Initiation Protocol (SIP) has been made in the IETF for supporting the control of IP-based multimedia sessions as a signaling protocol [11]. The SIP is an application-layer protocol that can establish, modify, and terminate multimedia sessions for supporting user mobility.

It is noted that the SIP can support user mobility to provide the location management [12]. The following describes the SIP operations for mobility.

Step 1. When an MT moves into a new network, it will register its current location by sending an SIP REGISTER message to the SIP Registrar.
Step 2. The Registrar may deny or accept the request. In the acceptance case, the SIP server will update the location database with the new location information.

When the MT moves into a new network or system, the SIP registration procedures are repeated to update the location.

On the other hand, the SIP may be used to support handover using the SIP RE-INVITE message.

Step 3. When an MT moves to a new network region during a session, it will send a new RE-INVITE message to CN.

Step 4. The RE-INIVITE message must include a new IP address of the MT. When CN receives the RE-INVITE message, it replies with the SIP OK message. Now, the MT can directly communicate with CN.

It is noted that the SIP-based handover cannot provide seamless mobility, since the on-going TCP/UDP session will be terminated when the MT changes its IP address.

C. Comparison of the Existing Mobility Protocols

Table 1 summarizes the comparison of the existing mobility protocols: MIP, SCTP, and SIP.

Fist of all, the MIP operates at the IP network layer to support the mobility. The MIP needs the route optimization extension to avoid the so-called triangular routing problem. Furthermore, the MIP provides the location management but supports the limited handover with the help of the mobility agents such as Home Agent (HA) and Foreign Agent (FA). In order to support the fast and seamless handover, the MIP needs to be extended as the Fast Handover to MIP (FMIP).

Secondly, the SCTP can be used to provide the seamless handover in the transport layer. The SCTP does not support the location management, but it can be used along with the MIP or SIP for location management. On the other hand, the SCTP does not require any additional mobility agents. It intrinsically provides the route optimization for data transport

Finally, the SIP is an application layer signaling protocol. The SIP could provide the location management. It is noted that most of the next-generation network systems consider the SIP as a signaling protocol for IP-based multimedia services. However, the SIP could not support seamless handover.

Table 1. Comparison of the existing mobility protocols

	MIP	**SCTP**	**SIP**
Operation Layer	Network Layer	Transport Layer	Application Layer
Location Management	Provided	Not Provided (May be used with MIP)	Provided
Mobility Agents	HA, FA (MIPv4)	No need of mobility agents	SIP Servers (e.g. Registrar)
Route Optimization	Need an extension for route opt.	Intrinsically provided	Intrinsically provided
Handover Support	Limited handover by MIP, (FMIP as extension)	Provided	Limit handover in the application layer

3 Analysis of SCTP Handover of Handover Latency

In this section, we analyze the latency of SCTP handover for a mobile terminal (MT). The handover latency is defined as the gap between 'the time that the MT has received the last DATA chunk over the old IP address', and 'the time that the MT has received the first DATA chunk over the new IP address' [13, 14].

For handover analysis, we consider the single-homing MT that can only use a single network interface at a time. This scenario could be applied to the horizontal handover of an MT that is moving within homogenous networks. In this case, the "link-up" of a new link and "link-down" of the old link will occur at the same time in the underlying link and network layers. That is, the SCTP handover will occur together with the link-layer handover at the same time.

For the single-homing MT, the handover latency $T_{handover\text{-}latency}$ can be calculated by summing up the time T_{DHCP} (for the configuration of a new IP address from a DHCP server), the time T_{ASCONF} (for the Add-IP and Primary-Change and Delete-IP operations in the SCTP handover), and the time $T_{link\text{-}handover}$ (for the handover at the underlying link layer). Accordingly, the total handover latency of SCTP will be

$$T_{handover\_latency} = T_{ASCONF} + T_{link\_handover} + T_{DHCP} \tag{1}$$

In the equation, T_{ASCONF} corresponds to the Round Trip Time (RTT) for exchange of ASCONF and ASCONF-ACK chunks between MT and FS. It is noted that the RTT is proportional to the distance between two endpoints and also inversely proportional to the bandwidth of the link. We also note that the SCTP handover requires three times of exchanges of ASCONF and ASCONF-ACK chunks for ADD-IP, Primary-Change, Delete-IP, respectively. That is, we can rewrite T_{ASCONF} as

$$T_{ASCONF} = 3 \times \left(\frac{L \times D_{AR\_FS}}{BW_{wired}} + \frac{L \times D_{MT\_AR}}{BW_{wireless}} \right) \tag{2}$$

where L is the packet length of chunks, and $D_{AR\text{-}FS}$ is the distance AR (Access Router) and FS , and $D_{MT\text{-}AR}$ is the distance between MT and AR, and BW_{wired} is the bandwidth of the wired link, and $BW_{wireless}$ is the bandwidth of the wireless link.

From the equation (1) and (2), and by considering that the address configuration time T_{DHCP} can relatively be viewed to be a constant value, we may conclude that the SCTP handover latency $T_{handover\text{-}latency}$ depends on the handover delay of the underlying link layer $T_{link\text{-}handover}$ and the RTT for exchange of ASCONF chunks T_{ASCONF}.

4 Experimentation of SCTP Handover on Linux Platform

In this section, we describe some experimental results of the SCTP handover that have been performed over Linux platform.

4.1 Test Scenarios

To experiment the SCTP handover, we construct a small test network, which consists of one router and two hosts (FS and MT). Each host supports Linux Kernel 2.6.8 together with LK-SCTP [15] tools.

We consider the handover of MT that is moving between two IP networks and have experiment the following two test scenarios, as shown in Fig. 2.

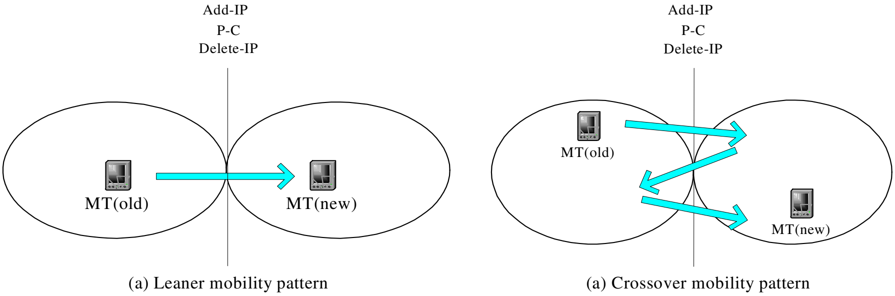


Fig. 2. Mobility pattern for SCTP Handover

Scenario A: linear mobility pattern (Fig. 2(a))

In the linear mobility pattern, an MT moves straightforward from an old IP region toward a new IP region linearly, without coming back to the old region. That is, an MT will not be affected by the so-called ping-pong effect. In this case, the handover of MT could occur only once.

Scenario B: crossover mobility pattern (Fig. 2(b))

In the crossover mobility pattern, an MT is moving forward and backward between the old IP region and the new IP region. That is, an MT is affected by the ping-pong effect. In this case, the handover of MT would occur several times.

For those two scenarios, we commonly applied the following handover procedures:

1) MT initiates an SCTP association with the FS. Initially, the MT uses the IP address 192.168.0.101, and the FS binds to the IP address 192.168.0.100. After initiation, two endpoints start to exchange the data packets.
2) MT adds a new IP address “192.168.0.102” to the association using *sctp_bindx()*.
3) MT requests the Primary-Change to the FS using *setsockopt()*.
4) MT deletes the old IP address from the association using *sctp_bindx()*.
5) When the data transport is completed, the MT shutdown the SCTP association.

For the analysis of the linear mobility pattern, we measured the ‘handover latency’ as a performance metric of SCTP. In addition, in the case of the crossover mobility pattern, we measured the ‘throughput for data transmission” as the total number of bytes transmitted during the association period”. The purpose of the throughput measurement is to identify whether the SCTP handover performance will be affected by the ping-pong effect in the crossover mobility pattern of MT.

4.2 Results and Discussion

From the table, we see that the FS and MT (old) establish an SCTP association through the 1st ~ 4th packets. At packets 10 and 11, the SCTP Add-IP operation is performed between MT and FS.

For the Primary-Change operation, the MT sends an ASCONF chunk to the FS at Packet 15, and the FS replies with ASCONF-ACK to the MT at Packet 19. It is noted that after the primary IP address is changed, the FS transmits all DATA packets to the new IP address of MT.

When the existing old link gets down, the MT sends the ASCONF chunk to FS for the Delete-IP operation, as shown at Packet 27. The FS then replies with ASCONF-ACK to MT at Packet 28.

From the table, we note that after the Add-IP operation is completed, the MT uses only the new IP address for all the SCTP packets. This is because the old link and IP address are not available any more since the "link-down" of the old link occurs at the same time together with the Add-IP operation for the single-homing MT.

Table 2 shows the result of SCTP handover for the single-homing MT.

Table 2. Results of SCTP Handover for the linear mobility pattern

NUM	TIME	PACKET	FROM	TO	NUM	TIME	PACKET INFO	FROM	TO
1	0.0000	INIT	MT (old)	FS	19	2.779674	ASCONF_ACK	FS	MT(new)
2	0.000221	INIT_ACK	FS	MT(old)	20	2.780051	DATA	MT(new)	FS
3	0.000430	COOKIE_ECHO	MT(old)	FS	21	2.780357	DATA	FS	MT(new)
4	0.000678	COOKIE_ACK	FS	MT(old)	22	2.780567	DATA	FS	MT(new)
5	0.002023	DATA	FS	MT(old)	23	2.780699	SACK	MT(new)	FS
6	0.002196	SACK	MT(old)	FS	24	2.780729	DATA	FS	MT(new)
7	0.002277	DATA	FS	MT(old)	25	2.805702	DATA	MT(new)	FS
8	0.021647	SCAK_DATA	MT (old)	FS	26	2.805919	SCAK	FS	MT(new)
9	0.021926	SACK	FS	MT(old)	27	2.806360	ASCONF	MT(new)	FS
10	0.023800	ASCONF	MT(old)	FS	28	2.806505	ASCONF_ACK	FS	MT(new)
11	0.223489	ASCONF_ACK	FS	MT(old)	29	2.806784	DATA	FS	MT(new)
12	2.463763	DATA	FS	MT(old)	30	2.806900	DATA	MT(new)	FS
13	2.463948	DATA	MT(old)	FS	31	2.807058	SACK	MT(new)	FS
14	2.778830	DATA	FS	MT(old)	32	2.807483	DATA	MT(new)	FS
15	2.779132	ASCONF	MT(old)	FS	33	2.807483	SACK	FS	MT(new)
16	2.779296	DATA	FS	MT(old)	34	2.807808	SHOUTDOWN	FS	MT(new)
17	2.779431	SACK	MT(old)	FS	35	2.807970	SHUTDOWN_ACK	MT(new)	FS
18	2.779466	SACK	FS	MT(old)	36	2.808103	SHUTDOWN_COMPLETE	FS	MT(old)

From the table, the SCTP handover latency is measured as the time gap between Packet 9 and 12, "2.463 - 0.022 = 2.441 (sec). In fact, this handover latency approximately corresponds to the time taken for processing link-down (old-link) and link-up

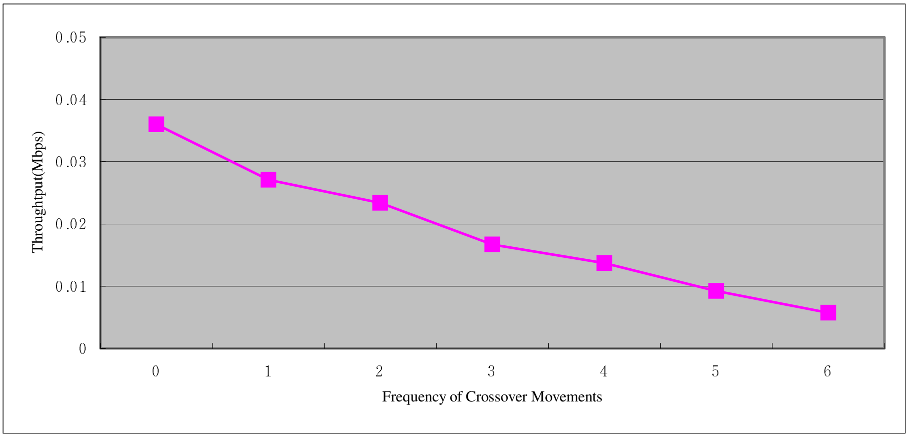


Fig. 3. Throughput for the crossover movements of MT

(new link) at MT (note that these events occur at Packet 11 and 12 in this experiment). Accordingly, we see that the handover latency for the single-homing MT is severely affected by the processing time of the link-down and link-up at the underlying layer.

This result can be interpreted from the formula (1) in Section 3. More specifically, the SCTP handover latency $T_{handover\text{-}latency}$ is linearly proportional to both $T_{link\text{-}handover}$ (the processing time of the link-down and link-up in the underlying layer) and T_{ASCONF} (RTT between MT and FS). However, we can see that the SCTP handover latency is more affected by $T_{link\text{-}handover}$ rather than the T_{ASCONF} for the single-homing MT.

On the other hand, Figure 3 shows the result of the throughput for the crossover mobility patterns of MT. In the figure, the measured throughputs are depicted for the different numbers of the crossover occurrences.

From the figure, it is shown that the throughput performance gets degraded, as the number of the crossover movements of an MT gets larger. This is because the total handover latency will cumulatively increase, as the MT moves across (handover) the different IP network regions more frequently. That is, the larger (cumulatively) handover latency will result in the lower throughput during the SCTP association period.

5 Conclusions

In this paper, we analyze the latency of SCTP handover for the single-homing mobile terminal theoretically and by experimentations over Linux platform. From the results, it is shown that the SCTP handover latency may severely depend on the handover delay at the underlying link layer. Instead, the RTT between two SCTP endpoints is negligible in the terms of the overall handover latency. On the other hand, in the case of the crossover movement pattern, the throughput of the data transmission could be degraded, as the MT performs the crossover movements more frequently.

References

1. Perkins, C.: IP Mobility Support for IPv4. IETF RFC 3344, August (2002)
2. Handly, M., et al.: SIP: Session Initiation Protocol. RFC 2543, Internet Engineering Task Force, March (1999)
3. Stewart, R., et al.: Stream Control Transmission Protoco. IETF RFC 2960, October (2000)
4. Snoeren, C., et al.: An End-to-End Approach to Host Mobility. Proceeding MobiCom, August (2000) 155-166
5. Stewart, R., et al.: Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration. IETF Internet Draft, draft-ietf-tsvwg-addip-sctp-08.txt, June (2004)
6. IETF RFC 3344: IP Mobility Support for IPv4. August (2002)
7. IETF RFC 3775: Mobility Support in IPv6. June (2004)
8. Gustafsson, E., Jonsson, A., and Perkins, C.: Mobile IPv4 Regional Registration. IETF draft draft-ietf-mip4-reg-tunnel-00.txt, November (2004)
9. Hesham Soliman, et al.: Hierachical Mobile IPv6 Mobility Management (HMIPv6). IETF Draft Draft-ietf-mipshop-hmipv6-04.txt, December (2004)
10. K.EL Malki, et al.: Low Latency Handoffs Handoffs in Mobile IPv4. IETF Draft Draft-ietf-mobileip-lowlatency-handoffs-v4-09.txt, June (2004)
11. IETF RFC 3261: SIP: Session Initiation Protocol. June (2002)
12. Elin, Wedlund et al.: Mobility Support using SIP. WoWMoM (1999) 76-82
13. Chang, M., et al.: Transport Layer Mobility Support Utilizing Link Signal Strength Information. IEICE Transaction on Communications, Vol. E87-B, No.9, September (2004) 2548-2556
14. Koh, S., et al.: mSCTP for Soft Handover in Transport Layer. IEEE Communications Letters, Vol. 8, No.3, March (2004) 189-191
15. Linux Kernel SCTP Project, Available from http://lksctp.sourceforge.net

A Fuzzy Time Series Prediction Method Using the Evolutionary Algorithm

Hwan Il Kang

Dept. of Information Eng., Myongji University, Yongin Geonggi province,
449-728, Republic of Korea
hwan@mju.ac.kr

Abstract. This paper proposes a time series prediction method for the nonlinear system using the fuzzy system and the genetic algorithm. At first, we obtain the optimal fuzzy membership function using the genetic algorithm. With the optimal fuzzy rules and the input differences, a better time prediction series system may be obtained. In addition, we may obtain the optimal fuzzy membership functions in terms of the evolutionary strategy and we obtain the time series prediction methods using the optimal fuzzy rules. We compare the time series prediction method using the genetic algorithm with that using the evolutionary strategy.

1 Introduction

The time series is a collection of the measured values $x_1, x_2, ..., x_n$ at the determined time. In general, the time series has the property of the chaotic signal or the irregular signal and it is difficult for the time series to be modeled as a function. Recently, the paper in [1] presents a modeling of the time series using the fuzzy rules and the appropriate membership functions. The modeling of the time series using the neural network and the fuzzy logic have been studied [2,3,4,5]. The paper in [6] generates the fuzzy rules through the learning and the authors in [7] use the IF-Then rules as a chaotic signal prediction. The authors in [8] develop the fuzzy system for the Shanghai stock index prediction. In addition, for the smart trading, the fuzzy rules are developed by [9]. In this paper, a better time series prediction method is obtained by the modified input method compared with the paper [10]. As the performance measure we select the root mean squared error (RMSE) and we use the Mackey-Glass time series as an input. For the optimal fuzzy rules, we use the genetic algorithm or the evolutional strategy. We want to compare the performance using the time prediction method based on the genetic algorithm with that using time prediction method based on the evolutionary strategy.

2 Fuzzy Inference Method

The fuzzy inference system finds out the fuzzy rule R such that

$$\hat{x}_{n+k} = R(x_1, x_{2,} ..., x_n) \tag{1}$$

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 530 – 537, 2005.

The equation (1) is used to predict the future in the short term. Instead of using the exact value, we can use the difference of the value as follow:

$$\hat{x}_n - \hat{x}_{n+k} = R((x_1 - x_1),(x_2 - x_1),...,(x_{n-1} - x_n)) \tag{2}$$

In this paper, the fuzzy rule determines where the fuzzy membership function belongs to with the modified input and output. For example, in the case of the value x_1 - x_2, we save the most frequent fuzzy membership function and the value of the membership function. When, for the same condition clauses, the different resultant clauses happen, we solve this illogical relation with the max-min method. That is, for all the illogical relations, the resultant clause is selected having the maximum value of the minimization of the values of the membership functions. The estimated value is the result from the Middle of Area(MOA), the centroid, or the Middle of the Maxima(MOM). To obtain the output, we may use the Mandami's Min-Max or Larson product [12].

3 Introduction to the Evolutionary Algorithm

The shapes of the fuzzy function may be composed of the same triangles. But, depending on the input and output, we modify the shapes of the fuzzy function with the evolutionary algorithm. At first, the optimal fuzzy shapes are determined by the genetic algorithm. The main operator of the genetic algorithm is the crossover. In the second hand, with the evolutionary strategy, the optimal fuzzy shapes may be determined. The main operator of the evolutional strategy is the mutation.

3.1 The Genetic Algorithm

The genetic algorithm is a repetitive search optimization method using the biological evolution and the survival of the fitness. The advantage of the genetic algorithm eliminates the possibility of searching the local optimal points instead of the global optimal points. The characteristic of the genetic algorithm uses the binary number associated with the given parameters. We use the two operators such as the crossover and the mutation. The process of the selection methods gives us the better set of solutions [11]. The roulette wheel, tournament, and the tournament with elitism belong to the category of the selection methods. The mathematical basics of the genetic algorithms are as follows: Each generation has a N number of parameters which are represented as binary numbers. In this paper, we use the proportional selection as the selection methods. As operators, we use the one point crossover and the bit mutation. Each individual of each generation consists of $S = \{0,1\}^l$. Each generation is described by

$$X = (x_1, x_2, ..., x_N)^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1l} \\ x_{21} & x_{22} & \cdots & x_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & & x_{Nl} \end{pmatrix} \tag{3}$$

$x_i \in S$ means the i-th individual of the matrix X and the x_{ij} is the j-th element of the vector x_i. The process of the genetic algorithm is classified as five steps.

Step 1: Setting to $k = 0$, and generate the initial $X(k)$.

Step 2: From the current generations, select the pairs form the generations.

Step 3: Generate the new N individuals by the crossover

Step 4: Operate the bit mutation to the N individuals

Step 5: If the stop condition is satisfied, stop. Otherwise, set to $k = k + 1$ and go to step 2

3.2 The Evolutionary Strategy

The operators used in the evolutionary strategy are the selection method and the mutation. At first we generate strings. The number of the strings are a and we select the strings b among the strings. We generate strings by adding the Gaussian random parameter to each string. The Gaussian parameter is determined by the random variance and the zero mean. The total number of strings is ab. We select the strings by the selection rule. The number of the strings are a. We continue this process until the stop condition holds [12].

4 An Application to Time Series Prediction

The Mackey-Glass time series prediction is described by

$$\frac{dx(t)}{dt} = \frac{0.2x(t - \tau)}{1 + x^{10}(t - \tau)} - 0.1x(t) \qquad (4)$$

where $\tau = 17, \ x(0) = 1.2, \quad x(t) = 0 \quad (t < 0)$.

We generate 1200 samples per each second. To determine the Fuzzy rule and the shape of the triangle, we use the subsequences from $x(117)$ to $x(616)$. We use the Max-Min method[10] and Larson product as a defuzzification scheme. To optimize the fuzzy membership function, we use the subsequences from $x(495)$ to $x(699)$. The fitness function is chosen as a reciprocal of the sum of the squared of the error.

4.1 The Definition of the Membership Function

Step 1: Generate the 50 random numbers. Theses 50 numbers are used as a chromosome. We can choose the smallest four numbers. We make a 50-bit binary number whose values are equal to zero except the positions of the four smallest numbers.

So the 50-bit binary number has four 1 bits.

Step 2: The six membership function is defined by

$$\begin{aligned}
\mu_1 &= \frac{(x - p_{\min} + \varepsilon)}{\varepsilon} && x < p_{\min} = b(0) \\
&= \frac{(b(1) - x)}{(b(1) - p_{\min})} && p_{\min} < \quad x < p_{\min} = b(1) \\
\mu_i &= \frac{(x - b(i - 2))}{b(i-1) - b(i-2)} && b(i-2) < x < b(i-1) \\
&= \frac{(b(i) - x)}{(b(i) - b(i-1))} && b(i-1) < x < b(i) \\
&(i = 2,3,4,5) \\
\mu_6 &= \frac{(x - b(4))}{p_{\max} - b(4)} && b(4) < x < p_{\max} = b(5) \\
&= \frac{(b(i) - x)}{(b(i) - b(i-1))} && p_{\max} < x < p_{\max} + \varepsilon
\end{aligned} \tag{5}$$

where $b(i) = p_{\min} + \dfrac{j(i)(p_{\max} - p_{\min})}{51} \quad i = 1,2,3,4$

$j(i)$ means that the j-th element is equal to 1 by counting the bits from the left to the right. In addition, from the first to j-th element, we meet 1 i times. For example, for the case of the binary number 0000000001000000000100000000010000000000 000000000000, we have $j(2) = 20$.

Step 3: Using the subsequences from $x(117)$ to $x(616)$, we search the fuzzy rules and we obtain the optimal membership shape using the subsequences from $x(495)$ to $x(699)$.

For the genetic algorithm, the number of generations is 20, the bit length is 50, the crossover rate is 0.25 and the mutation rate is 0.01. The number of each generation is 20 and we allow the one point crossover. We choose the proportional selection as the selection rule. For the evolutional strategy, the number of the initial string is 50. Among 50 strings, we choose the five strings having the largest five fitness numbers. From the five strings, the means of the random number is zero mean and its variance is 0.1.

5 Experimental Results

Mackey-Glass time series is Fig. 1 and the optimal membership functions are Fig. 2. In Fig. 3, the fitness functions are plotted for the generation by the genetic algorithm and by the evolutionary strategy, respectively. The estimated values and the real values are plotted in Fig. 4 by the genetic algorithm and in Fig. 5 by the evolutionary strategy. From the subsequences from $x(701)$ to $x(900)$, we performed the time series prediction using the fuzzy rules and the optimal membership functions obtained from the former subsequences. The root mean squared error is 0.0333 for the genetic algorithm. The root mean squared error is 0.0330 for the evolutionary strategy.

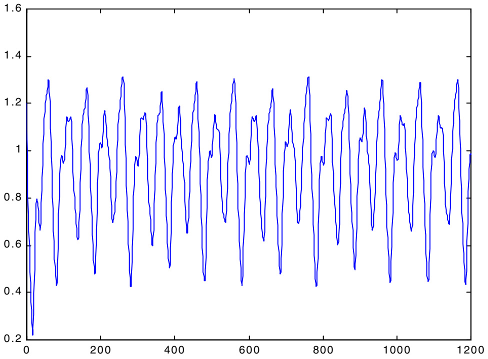
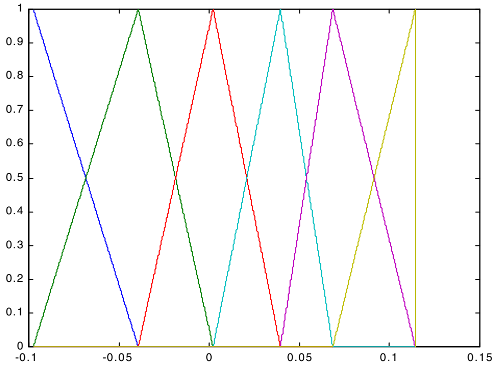
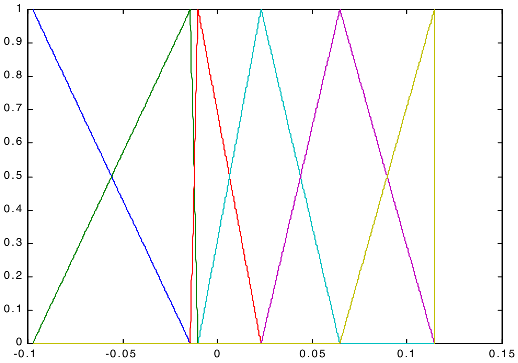


Fig. 1. Mackey-Glass time series

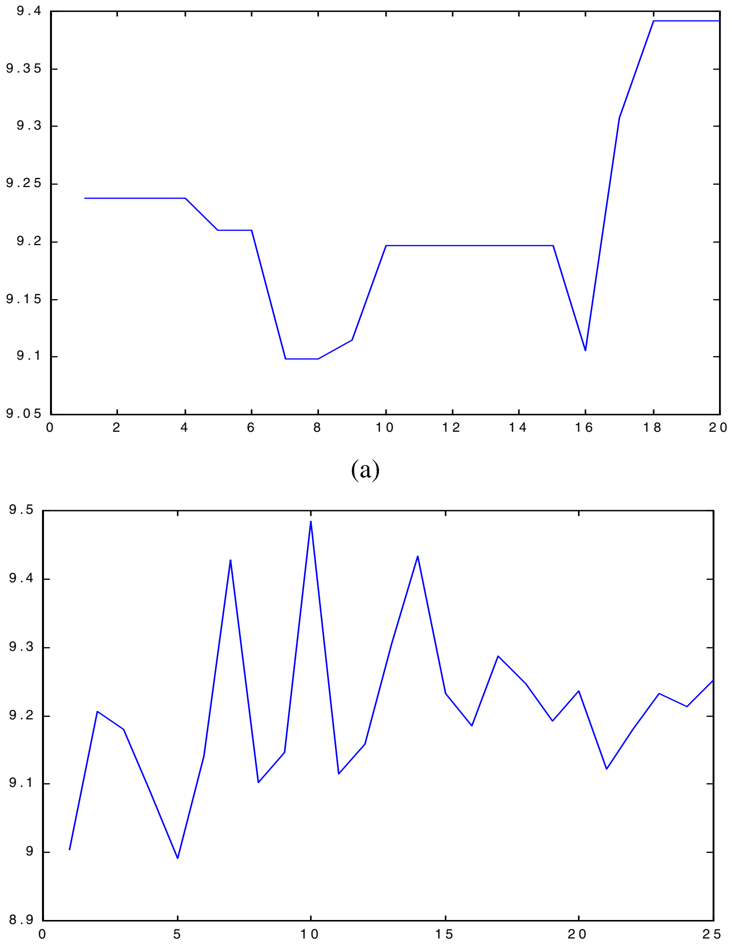


(a)



(b)

Fig. 2. (a) The optimal membership functions using the genetic algorithm, (b) The optimal membership functions using the evolutionary strategy



(a)

(b)

Fig. 3. (a) The fitness functions according to the generations, (b) The fitness functions according to the generations

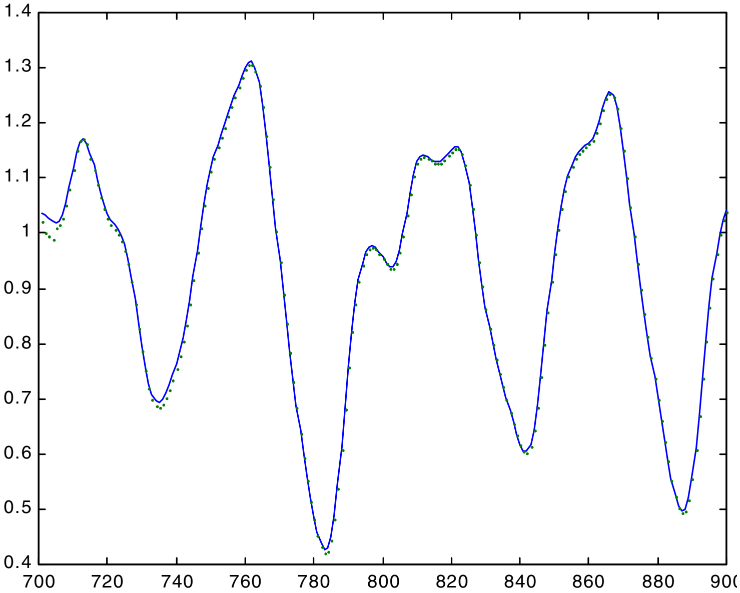


Fig. 4. The estimated values and the true values using the genetic algorithm

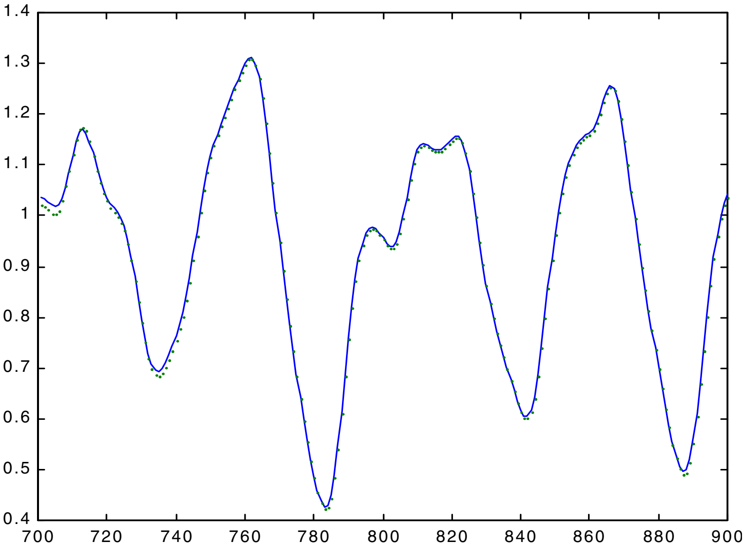


Fig. 5. The estimated values and the true values by the evolutionary strategy

6 Conclusions

In this paper, the optimal membership functions may be determined by the evolutionary algorithm such as the genetic algorithm and the evolutionary strategy. Using the optimal fuzzy rules and the optimal membership functions, we determine the time series prediction with the modified input.

References

1. Kim, Intack, Gong, Chang Wook: The Fuzzy Learning Algorithms for Time Series Prediction. The fuzzy and intelligent system journal 7 (3) (1997) 34-42
2. Tong, R. M.: The evaluation of Fuzzy Models derived from Experimental Data. Fuzzy sets and Systems 4 (1980) 1-12
3. Pedrycz, W.: Fuzzy Control & Fuzzy Systems, John Wiley & Sons (1989)
4. Hornik, K., Stinchcombe M., White H.: Multilayer Feedback Networks are Universal Approximators," Neural Network 2 (1989) 359-366
5. Wang, L. E.: Fuzzy Systems are Universal Approximators, Proc. IEEE International Conf. on Fuzzy Systems, San Diego, (1992) 1163--1170
6. Wang L. X., Mendel, J. M.: Generating Fuzzy Rules by Learning from Examples, IEEE Trans Syst., Man, Cybern., vol. 22 (1992) 1414-1427
7. Jang J. R., Sun C.: Prediction Chaotic Time Series with Fuzzy If_Then rules, Second IEEE International Conference on Fuzzy Systems, San Francisco (1993) 1079--1084
8. Ye Z., Gu L.: A Fuzzy System for Trading the Shanghai Stock Market, Trading on the Edge, Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets, G. J. Deboeck, Ed. New York: Wiley (1994) 207-214
9. Benachenhou D.: Smart Trading with (FRET), in Trading on the Edge, Neural, Genetic, and Fuzzy Systems for Chaotic Financial Markets, G. J. Deboeck, Ed.,New York: Wiley (1994) 215-242
10. Lee, Seong Rok,Kim, Intack: Study on New Fuzzy Time Series Prediction Method, Myongji University Technical journal 19 (1999) 565-569
11. Michalewicz, Zbigniwe, Genetic Algorithms+ Data Structures=Evolution Programs, Springer Verlag (1994)
12. Lim, Yong do, Lee, Sang D.: Fuzzy, Neural Networks and the Evolutionary evolution, Young and il publishing company, Seoul (1996) 215-126

Modeling for Security Verification of a Cryptographic Protocol with MAC Payload

Huanbao Wang, Yousheng Zhang, and Yuan Li

School of Computer Science & Information Technology,
Hefei University of Technology, Hefei Anhui, 230009, China
wanghuanbao@yahoo.com.cn
zhangyos@mail.hf.an.cn
li_fu_yuan@126.com

Abstract. We propose a new sub-term relation to specify syntax of messages with MAC (Message Authentication Code) payload for the cryptographic protocols in the strand space model. The sub-term relation was introduced to formal analysis of cryptographic protocols based on theorem proving, but some defects have been found in it. In the present paper, first, the *operator f* is defined to the extend sub-term relation, which is used to amend its original flaws. Second, a new ideal is constructed, and is used to expand the bounds on the penetrator's abilities. Third, the decidable theorem for honesty of ideals holds as it is described under the extended sub-term relation is proved. Fourth, we propose the theorem of the satisfiability for decidable conditions of honest ideals and annotate how invariant-sets generate, which is used to verify security properties of cryptographic protocols.

1 Introduction

Security protocols, namely cryptographic protocols are absolutely necessary components that mediate communication between Internet users. It is very crucial for designing a security protocol with both secure and efficient property, and proving its correctness in information security research field. Since the late 70's of 20th century, many formal analysis approaches of cryptographic protocols have been proposed; they include logic proving (e.g. [1], [2]), model checking (e.g. [3], [4], [5]), theorem proving (e.g. [6], [7]), and type checking (e.g. [8], [9], [10], [11], [12]). The existing methods however, are limited for verifying properties of cryptographic protocols, for instance, one can only reason the insecurity of cryptographic protocols with logic proving and model checking.

The strand space model [7], a typical method based on theorem proving, has attracted more attention; we may derive a proof of cryptographic protocols' security properties by using it. The ideal is defined as an algebraic structure of invariant-sets that represent the secret knowledge of cryptographic protocols' principals. And the bounds on the penetrator's abilities are described by using the properties of ideals. So the key procedure of formal analysis of cryptographic protocols is how to construct invariant-sets. The sub-term relation is defined inductively according to the mapping *inv*, *encr*, and *join*; and the literature [7] develops the decidable theorem for honest

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 538 – 547, 2005.

ideals, which are used to prove the correspondence and secrecy properties of cryptographic protocols. But the strand space model fails when it is applied to analyze terms with MAC payload, as there are no sub-term relations that can be used to analyze MAC payload in it. The literature [13] develops CPA (Cryptographic Protocol Algebra) for syntax of ordinary terms. But, as its operators' set is an infinite one, it is very difficult for verification algorithm for security properties of cryptographic protocols in the CPA model to be specified. In the present paper, we will introduce our works about the several key problems based on theorem proving in the strand space model.

In Section 2, we define the *f operator* as a special syntax of messages with MAC payload for cryptographic protocols to extend the original sub-term relation. The new sub-term relation covers four operators that include *inv*, *encr*, *join*, and *f operator*. So the extended sub-term relation can construct some complex messages among participants of cryptographic protocols. In Section 3, we propose some propositions about the sub-terms' equivalence and derive the same computational complexity of problems including only *encr* operator as one of problems including only *f operator*. Although our new representation of invariant-sets still holds in the style of F. Javier Thayer Fábrega, Jonathan C. Herzog, and Joshua D. Guttman, obviously, its semantics is different from the former. We also specify the attacking actions. In Section 4, we redefine the decidable theorem for honesty of ideals and prove the extended decidable theorem for honesty of ideals, which still holds as it is described under the extended sub-term relation. When, especially, given G_f. ($\subseteq$H.), we verify the theorem refined. In Section 5, we propose the theorem of the satisfiability for decidable conditions of honest ideals and a new algorithm for invariant-sets' construction, which is used to verify authenticity and secrecy of cryptographic protocols.

2 Sub-term Relation

Consider a term set A; its elements are all the possible messages that can be exchanged between principals in a cryptographic protocol. In this section we assume that only one sub-term relation is defined on the set A, in fact, where more algebraic structures exist, but they are independent of our immediate purposes.

2.1 Foperator

We will first specialize a set of terms A. In particular, we assume: a set $T\subseteq A$ of texts, which represents the atomic messages, a set $K\subseteq A$ of cryptographic keys disjoint from T, viz. $K\cap T=\varnothing$. The definition of *inv*, *encr*, and *join* operator sees the literature [7]. We will follow custom and write *inv*(k) as k^{-1}, *encr*(k, m) as $\{m\}_k$, and *join*(a, b) as $a\bullet b$. Moreover, we now define a new sub-term operator.

Definition 2.1 ***f operator:***
f: $A^n\rightarrow A$, is denoted by $f(m_1, m_2, \cdots, m_n)$, where A is a term set; $m_1, m_2, \cdots, m_n\in A$; n is a finite integer; and the function $f(m_1, m_2, \cdots, m_n)$ is one-way, then:

if given $m_1, m_2, \cdots, m_n$, then we can calculate $f(m_1, m_2, \cdots, m_n)$; whereas, if given $m_1, m_2, \cdots, m_n$, we can not calculate $m_1, m_2, \cdots, m_n$.

Particularly, if given $m_1, m_2, \cdots, m_{n-1}$, whereas, k is secret, then we can not calculate $f(k, m_1, m_2, \cdots, m_{n-1})$. We refer to $f(k, m_1, m_2, \cdots, m_{n-1})$ as a one-way function with a cryptographic key k, where $k \in K$, $m_1, m_2, \cdots, m_{n-1} \in B \subseteq A$. The mapping is written: f: $K \times B^{n-1} \rightarrow B$, where $B \subseteq A$.

For the sake of simplicity, it is denoted *f operator*.

2.2 Some Assumptions

In formal analysis of strand spaces, and many other approaches (e.g. [6]), the proofs of many agreements or theorems proving use two assumptions, and they are free encryption and perfect encryption. The free assumption guarantees the equivalence of sub-terms holds in the meaning of sub-term syntax, while the perfect encryption assumption can be used to encrypt sub-terms with absolutely safety.

Free Encryption:

$$\text{For } m, m' \in A, \text{ and } k, k' \in K, \{m\}_k = \{m\}_{k'} \Rightarrow m = m' \wedge k = k'$$

It means that a ciphertext can be only regarded as a ciphertext in just one way, namely that if two ciphertexts are equivalent to each other, we can reason their cryptographic keys and plaintexts equal to each other respectively.

The *inv*, *encr*, *join* and *f operator* can calculate terms that are exchanged between principals of a cryptographic protocol. For clarity of expositing sub-term relation, we make a stronger assumption here, namely,

Sub-term Constructed:

For $m_1, m_2, \cdots, m_{n-1}, m_1', m_2', \cdots, m_{n-1}' \in A$, and $k, k' \in K$, besides

$$m_1 \bullet m_2 = m_1' \bullet m_2' \Rightarrow m_1 = m_1' \wedge m_2 = m_2', m_1 \bullet m_2 \neq \{m_1'\}_{k'}, m_1 \bullet m_2 \notin K \cup T,$$

and $\{m_1\}_k \notin K \cup T$ in strand spaces [7], we will also assume:

1. $f(k, m_1, m_2, \cdots, m_{n-1}) \neq f(k', m_1', m_2', \cdots, m_{n-1}')$, means that if sub-terms and cryptographic keys are not same respectively, then the values calculated by *f operator* will not equal to each other, that is, if given $f(k, m_1, m_2, \cdots, m_{n-1})$, and $k', m_1', m_2', \cdots, m_{n-1}',$ then we can not calculate $f(k', m_1', m_2', \cdots, m_{n-1}')$ that equals to $f(k, m_1, m_2, \cdots, m_{n-1})$. Obviously,

$$f(k, m_1, m_2, \cdots, m_{n-1}) = f(k', m_1', m_2', \cdots, m_{n-1}') \\ \Rightarrow k = k' \wedge m_1 = m_1' \wedge \cdots \wedge m_{n-1} = m_{n-1}'$$

2. $f(k, m_1, m_2, \cdots, m_{n-1}) \neq \{m_1'\}_{k'}$, means that the value of sub-terms f-calculated cannot replace the value of sub-terms encrypted, namely that *f operator* and encryption are not equivalent.

3. $f(k, m_1, m_2, \cdots, m_{n-1}) \neq m_1 \bullet m_2$, means that the value of sub-terms f-calculated cannot replace the value of sub-terms concatenated, namely that *f operator* and concatenation are not equivalent.

4. $f(k, m_1, m_2, \cdots, m_{n-1}) \notin K \cup T$, means that the atomic sub-terms of messages or cryptographic keys cannot be generated by f-calculated of sub-terms.

2.3 Smallest Sub-term Relation

On the basis of the sub-terms' construction rules and its corresponding assumptions, a sub-term relation $\sqsubset$ is defined inductively, as the smallest relation such that:

Definition 2.2 Smallest Sub-term Relation:

1. $a \sqsubset a$;
2. $a \sqsubset \{g\}_k$ if $a \sqsubset g$;
3. $a \sqsubset g \bullet h$ if $a \sqsubset \mathrm{g} \vee a \sqsubset h$;
4. $a \sqsubset f(k, m_1, m_2, \cdots, m_{\text{n-1}})$ if $a \sqsubset k \vee a \sqsubset m_1 \vee a \sqsubset m_2 \vee \cdots \vee a \sqsubset m_{\text{n-1}}$.

Thus if $k \neq k'$, and $\{h\}_{k'} \sqsubset \{h\}_k$, then $\{h\}_{k'} \sqsubset h$; moreover, the function $f(k, m_1, m_2, \cdots, m_{\text{n-1}})$ is in just one-way by Definition 2.1, we may reason the relation $a \sqsubset k \vee a \sqsubset m_1 \vee a \sqsubset m_2 \vee \cdots \vee a \sqsubset m_{\text{n-1}}$ by the smallest relation $a \sqsubset f(k, m_1, m_2, \cdots, m_{\text{n-1}})$, but the concrete relation expression cannot be found out.

3 Invariant-Sets

We will get the equivalence properties of sub-terms operators and discuss how to denote invariant-sets under the extended sub-term relation.

3.1 Equivalence

According to the properties of a one-way function with trap-door parameter, if its trap-door parameter is known, we can calculate its inverse image, whereas, if its trap-door parameter is unknown, obviously we cannot do it.

Proposition 3.1 ***Equivalence*** **of** ***encr*** **and** ***f operators***

1. The inverse image of a one-way function with trap-door parameter known is calculable, which is equivalent for h of $\{h\}_k$ to be calculated by k^{-1} known.

2. The inverse image of a one-way function with trap-door parameter unknown is un-calculable, which is equivalent for $k, m_1, m_2, \cdots, m_{\text{n-1}}$ of $f(k, m_1, m_2, \cdots, m_{\text{n-1}})$ to be not calculated by $f(k, m_1, m_2, \cdots, m_{\text{n-1}})$ known.

3. The analysis of sub-terms architecture of one-way functions with trap-door parameter comes down to be the analysis of sub-terms architecture with *encr* and *f operators*.

PROOF. According to Definition 2.1, as the mapping f: $K \times B^{n\text{-}1} \rightarrow B$ is concerned, we can only calculate $f(k, m_1, m_2, \cdots, m_{\text{n-1}})$ only when given k and $m_1, m_2, \cdots, m_{\text{n-1}}$. Thus if k is secret, the function $f(k, m_1, m_2, \cdots, m_{\text{n-1}})$ can be kept unknown.

Thus the conditions for formal analysis of the sub-terms architecture with *f operator* already covered the ones for formal analysis of the sub-terms architecture of one-way functions with trap-door parameter.

Proposition 3.2 The image of a one-way function with a cryptographic key k known is calculable, which is equivalent for $\{h\}_k$ to be calculated by k known or for h of $\{h\}_k$ to be calculated by k^{-1} known.

3.2 Smallest *K*-Ideal

The invariant-set is denoted the smallest K-ideal in strand space model. In this paper, we will adopt it and describe the algebraic structure of invariant-sets with the smallest K-ideal, but which will be redefined under the extended sub-term relation.

Definition 3.1 ***Smallest K-ideal***
If $K \subseteq \mathcal{K}$, a K-ideal of the term set $\mathcal{A}$ is a subset I of $\mathcal{A}$ such that for all h, m_1, m_2, $\cdots$, $m_{n-1} \in I$, where n is one infinite integer, $g \in \mathcal{A}$ and $k \in K$:

1. $h \bullet g$, $g \bullet h \in I$;
2. $\{h\}_k \in I$;
3. $f(k, m_1, m_2, \cdots, m_{n-1}) \in I$.

The smallest K-ideal containing h is denoted $I_K[h]$.

If an attacker knows $g \bullet h$ or $h \bullet g$, then he/she can calculate h; moreover, if an attacker knows $\{h\}_k$ and k^{-1} , then he can calculate h. Thus if we want to keep h secret, then $g \bullet h$, $h \bullet g$, and k^{-1} at least should be secret; in addition, if we also keep $\{h\}_k$ secret, then an attacker cannot have any chance for a brute-force attack.

We will define the smallest set $I_K[h]$ containing h therefore, which is close for concatenation and encryption. Thus it can guarantee that $I_K[h]$ disjointed from the knowledge set of an attacker appears an empty set in the circulation of a crypto- graphic protocol.

Moreover, for $f(m_1, m_2, \cdots, m_n)$, we cannot calculate m_1, m_2, $\cdots$, m_n , because it is in just one way. So the value of $f(m_1, m_2, \cdots, m_n)$ reckoned in $I_K[h]$ can be kept itself secret in order to analyze the secret MAC [14] payload. Especially, for a one-way function with a cryptographic key k $f(k, m_1, m_2, \cdots, m_{n-1})$, if only we keep k secret, it can be kept secret. Thus the *f operator* being similar to *encr operator* will be reckoned in $I_K[h]$.

Definition 3.2 Specification of Extended Attacking Actions
Under the extended sub-term relation, attacking actions are classified into M. , F. , T. , C. , S. , K. , E. , D. , and H. , where

F. Flushing: $\langle -\{m_1, m_2, \cdots, m_{n-1}\} \rangle$, where $m_1, m_2, \cdots, m_{n-1} \in \mathcal{B} \subseteq \mathcal{A}$;

H. Message digest: $\langle -k, -\{m_1, m_2, \cdots, m_{n-1}\}, +f(k, m_1, m_2, \cdots, m_{n-1}) \rangle$.

The other actions' specification sees the literature [7].

4 Honest Ideals

A term is *simple* iff it is not of the form $a \bullet b$ for a, $b \in \mathcal{A}$. A term is *simple* iff it is either an element of $\mathcal{T}$, an element of $\mathcal{K}$, is of the form $\{h\}_k$, or of the form $f(k, m_1, m_2, \cdots, m_{n-1})$.

Proposition 4.1 Properties of Ideal

1. Suppose $k\in \mathcal{K}$; $S\subseteq A$; and for every $s\in S$, s is *simple* and is not of the form $\{g\}_k$ and $f(k, m_1, m_2, \cdots, m_{n-1})$. If $\{h\}_k\in I_K[S]$, then $h\in I_K[S]$; moreover, if $f(k, m_1, m_2, \cdots, m_{n-1})\in I_K[S]$, then $m_1, m_2, \cdots, m_{n-1}\in I_K[S]$.

2. Suppose $k\in \mathcal{K}$; $S\subseteq A$; and for every $s\in S$, s is *simple* and is not of the form $\{g\}_k$ and $f(k, m_1, m_2, \cdots, m_{n-1})$. If $k\in \mathcal{K}$ and $\{h\}_k\in I_K[S]$, then $k\in K\subseteq \mathcal{K}$; moreover, if $k\in \mathcal{K}$ and $f(k, m_1, m_2, \cdots, m_{n-1})\in I_K[S]$, then $k\in K\subseteq \mathcal{K}$.

3. Suppose $S\subseteq A$, and for every $s\in S$, s is *simple*. If $g\bullet h\in I_K[S]$, then either $g\in I_K[S]$ or $h\in I_K[S]$.

Proposition 4.2 Decidable Entry Points

Suppose C is a bundle over A. If m is the $\preccurlyeq_C$-minimal element in $\{m\in C$: $uns\_term(m)\in I\}$, then m is an *entry point* for I.

A set $I\subseteq A$ is *honest* relation to a bundle C, if and only if whenever a penetrator node p is an entry point for I, p is an M. node or a K. node.

Theorem 4.1 Decidable Theorem for Honesty of Ideals under the Extended Sub-term Relation:

Suppose C is a bundle over A; $S\subseteq T\cup \mathcal{K}$; $K\subseteq \mathcal{K}$; and $\mathcal{K}\subseteq S\cup K^{-1}$, where $K^{-1}=\{k^{-1}$: $k^{-1}=inv(k)\wedge k\in K\}$. Then $I_K[S]$ is *honest*.

PROOF: We will show the satisfiability for decidable conditions of *honest* ideals and discuss the extended decidable theorem under the extended sub-term relation.

The proof of the original theorem sees the literature [7]. Suppose n is a penetrator node and an *entry point* for I ($=I_K[S]$). We may consider the various kinds of strands on which a penetrator node can occur. By the proof of the original theorem, n node cannot be on a strand of F. , T. , C. , S. , E. , or D. . We consider now the remaining cases:

H. n belongs to a strand with trace $<-k, -\{m_1, m_2, \cdots, m_{n-1}\}, +f(k, m_1, m_2, \cdots, m_{n-1})>$. By assumption $f(k, m_1, m_2, \cdots, m_{n-1})\in I$, then $-\{m_1, m_2, \cdots, m_{n-1}\}\in I$, contradicting the definition of *entry point*.

Moreover, we especially consider two kinds of strands on which a penetrator node can occur. When there two attacks exist, we can check two cases respectively:

G. For C is a bundle over A, and a node G. $<+j>$ (M. $\cup$ K. $\subseteq G_P$, $j\in G_P$), if $B\backslash f(k, m_1, m_2, \cdots, m_{n-1})\subseteq T\subseteq A$ is in existence, the decidable conditions of *honesty* of $I_K[S]$ cover G. $<+j>$, where $j\in G_P\subseteq B$.

G_f. For C is a bundle over A, and a node G_f. $<-y=f(x_1, x_2, \cdots, x_n), -y_0, +(x_1, x_2, \cdots, x_n)>$ (G_f. $\subseteq$H.), where y_0 is one trap-door secrecy of a one-way function with trap-door parameter. On the basis of Term 1 in Proposition 3.1, it is contained in the extended decidable theorem for honesty of ideals. If a trap-door secrecy is unknown, then the inverse image of a one-way function with trap-door parameter cannot be calculated, it is equivalent for $k, m_1, m_2, \cdots, m_{n-1}$ of $f(k, m_1, m_2, \cdots, m_{n-1})$ to be not calculated by $f(k, m_1, m_2, \cdots, m_{n-1})$ known by Term 2 in Proposition 3.1. So the decidable theorem for honesty of ideals is extended by $B\backslash f(k, m_1, m_2, \cdots, m_{n-1})\subseteq T\subseteq A$ being given.

The only remaining possibilities are that n is on a strand of M. and K. as asserted.

By f: $K \times B^{n-1} \rightarrow B$ in Definition 2.1, if given $B \setminus f(k, m_1, m_2, \cdots, m_{n-1}) \subseteq T \subseteq A$, then Theorem 4.1 holds under the extended sub-term relation by Definition 2.2. In fact, the condition expression $B \setminus f(k, m_1, m_2, \cdots, m_{n-1}) \subseteq T \subseteq A$ is easily in existence. Although Theorem 4.1 in this paper is already expended under the extended sub-term relation, it has the same description as the original theorem (Theorem 6.11 [7]).

5 Invariant-Sets' Construction

If a principal sends message M in the circulation of a cryptographic protocol, then the process of authentication is divided into:

1. Construct an invariant-set $I_K[M]$ that contains M;
2. Check only regular *entry points* which are covered in $I_K[M]$;
3. Judge who sends the message M and in which step it is sent.

This authentication procedure has two key steps, constructing $I_K[M]$ that contains M and judging honesty of principals' *entry points*; and the later is often extremely difficult for multi-principals' cryptographic protocols. We will develop the method of authenticity checking based on the honesty of $I_K[M]$ in this paper.

By Theorem 4.1, if $S \subseteq T \cup K$, $K \subseteq K$, and $K \subseteq S \cup K^{-1}$, then $I_K[S]$ is *honest.* Obviously, the conditions, $S \subseteq T \cup K$, $K \subseteq K$, are easily contented. We will analyze the satisfiability for the decidable conditions, viz. $K \subseteq S \cup K^{-1}$ of honesty of $I_K[S]$ and develop a new algorithm for invariant-sets' construction, which exists based on the honesty of $I_K[S]$.

5.1 Satisfiability for Decidable Conditions of Honest Ideals

By the decidable conditions, viz. $K \subseteq S \cup K^{-1}$ of honesty of $I_K[S]$, where $S \subseteq T \cup K$, $K \subseteq K$, $K^{-1} = \{k^{-1}: k^{-1} = inv(k) \wedge k \in K\}$, we can get:

$$K = K_{\min} \cup K^{-1}_{\min}\text{ , where } K_{\min} \subseteq K,\ K^{-1}_{\min} \subseteq K^{-1};$$

$$K = K_{\max} = K^{-1}_{\max}\text{ , where } K_{\max} \subseteq K,\ K^{-1}_{\max} \subseteq K^{-1}.$$

For example, if given the sub-set S of $T \cup K$,

$S = \{t_1, t_2, t_3, k_1, k_2, k_3, k_4, k_5, k^{-1}_1, k^{-1}_2, k^{-1}_3 : t_1, t_2, t_3 \in T \wedge k_1, k_2, k_3, k_4, k_5 \in K \wedge k_4, k_5, k^{-1}_1, k^{-1}_2, k^{-1}_3 \in K^{-1}\} \subseteq T \cup K$,

we get

$$K_{\min} = \{k_1, k_2, k_3, k_4, k_5 : k_1, k_2, k_3, k_4, k_5 \in K\} \subseteq K;$$
$$K^{-1}_{\min} = \{k_4, k_5, k^{-1}_1, k^{-1}_2, k^{-1}_3 : k_4, k_5, k^{-1}_1, k^{-1}_2, k^{-1}_3 \in K^{-1}\} \subseteq K^{-1},$$
$$\text{where } k_4 = k^{-1}_4,\ k_5 = k^{-1}_5\ .$$

Theorem 5.1 Satisfiability for Decidable Conditions of Honest Ideals

Strong Satisfiability (*SS*): If $K = K_{\min} \cup K^{-1}_{\min}$, where $K_{\min} \subseteq K$, $K^{-1}_{\min} \subseteq K^{-1}$, then the *honesty* of $I_K[S]$ is *satisfiable.*

Weak Satisfiability (*WS*): If $K = K_{\max} = K^{-1}_{\max}$, where $K_{\max} \subseteq K$, $K^{-1}_{\max} \subseteq K^{-1}$, then the *honesty* of $I_K[S]$ is *satisfiable.*

5.2 Algorithm for Invariant-Sets' Construction

By Theorem 5.1, we develop a new algorithm for invariant-sets' construction, which are based on theorem proving and used to verify authenticity property of cryptographic protocols.

Algorithm for Invariant-sets' Construction:

Step 1: Initialization.

Select a bundle $C=<N_C,\ (\rightarrow_C \cup \Rightarrow_C)>$, where N_C is a node set, then $I_K[S]\equiv\emptyset$; Select a node $n=<s,\ i>\in N_C$, which denotes $n\in C$, where *index*(n)=i, *strand*(n)=s. Define *term*(n) to be $(tr(s))_i$, i.e. the ith signed term in the trace of strand s. Similarly, *uns_term*(n) is $((tr(s))_i)_2$, i.e. the unsigned part of the ith signed term in the trace of strand s.

Step 2: Analyze the structure of *uns_term*(n), then:

Step 2.1: if *uns_term*(n) is of the form $a\bullet b$, then it is reckoned in $I_K[S]$, $\{m\colon\ m\Box a\vee m\Box b\}$ is constructed, and $I_K[S]=I_K[S]\cup\{m\colon\ m\Box a\vee m\Box b\}$ is calculated. If a or b is not of an atomic sub-term and is of the form $\{m\}_k$, then switch to *Step 2.2*; If a or b is not of an atomic sub-term and is of the form $f(k,\ m_1,\ m_2,\ \cdots,\ m_{n-1})$, then switch to *Step 2.3*.

Step 2.2: If *uns_term*(n) is of the form $\{m\}_k$, then it is reckoned in $I_K[S]$, $\{t\colon\ t\Box m\}$ is constructed, and $I_K[S]=I_K[S]\cup\{t\colon\ t\Box m\}$ is calculated. If m is not of an atomic sub-term and is of the form $a\bullet b$, then switch to *Step 2.1*; If m is not of an atomic sub-term and is of the form $f(k,\ m_1,\ m_2,\ \cdots,\ m_{n-1})$, then switch to *Step 2.3*.

Step 2.3: If *uns_term*(n) is of the form $f(k,\ m_1,\ m_2,\ \cdots,\ m_{n-1})$, $\{m\colon\ m\Box k\}$ is constructed, and $I_K[S]=I_K[S]\cup\{m\colon\ m\Box k\}$ is calculated.

Step 3: All nodes $n=<s,\ i>\in N_C$ of $C=<N_C,\ (\rightarrow_C \cup \Rightarrow_C)>$ are went through, and at last $I_K[S]$ is got.

Step 4: Check if $I_K[S]$ is accordant to *SS* (*Strong Satisfiability*), then judge if the correspondence of a cryptographic protocol and the secrecy of its parameters.

Example. We will formalize the Internet Exchange Key (IKE) protocol [14] according to the strand space model and show its terms' architecture with MAC payload. *Quick Mode*, in which IKE runs, is defined as follows.

Strands with *trace Init* and *Resp* for IKE to run in *Quick Mode*:

1. *Init* is the set strands $s\in\Sigma$, whose trace is $<+M_1,\ -M_2,\ +M_3>$, where

$$M_1=HDR^*|HASH(1)|SA|N\mathrm{i}|[,\ KE]|[,\ ID\mathrm{ci},\ ID\mathrm{cr}],$$
$$M_2=HDR^*|HASH(2)|SA|N\mathrm{r}|[,\ KE]|[,\ ID\mathrm{ci},\ ID\mathrm{cr}],$$
$$M_3=HDR^*|HASH(3).$$

The principal associated with a strand $s\in<+M_1,\ -M_2,\ +M_3>$ is *Initiator*.

2. *Resp* is the set strands $s\in\Sigma$, whose trace is $<-M_1,\ +M_2,\ -M_3>$. The principal associated with a strand $s\in<-M_1,\ +M_2,\ -M_3>$ is *Responder*.

3. *Strand Space* is an infiltrated one Σ such that $\Sigma=Init\cup Resp\cup P$.

Consider now that the sets $Init\cap Resp=\emptyset$. And P Contains no strands of the same form as $Init\cup Resp$. So Σ over A is strand space for to run in *Quick Mode*. And given

HASH(1)=prf(SKEYID_a, M-ID|SA|Ni[|KE][|IDci|IDcr]),
HASH(2)=prf(SKEYID_a, M-ID|Ni_b|SA|Nr[|KE][|IDci|IDcr]),
and HASH(3)=prf(SKEYID_a, 0|M-ID|Ni_b|Nr_b),
then, HDR*, SA, Ni, Nr, KE, IDci, IDcr, SKEYID_a, M-ID $\in T$, prf $\in -f''$.

We can construct $I_K[S]$ for Σ over A according to our new algorithm for invariant-sets' construction in Section 5.2 and to Theorem 4.1 in Section 4. And we easily get the proof of authenticity for the IKE protocol to run in *Quick Mode*.

6 Conclusion

In the present paper, we have developed the extended sub-term relation by defining the *f operator* in the strand space model. In the mapping f: $K \times B^{n-1} \rightarrow B$, the term set B ($\subseteq A$) has no bounds, so the *f operator* is contributed to specify an ordinary sub-term relation. The operator set, which contains *inv*, *encr*, *join*, and *f operator*, is finite. Moreover, the propositions of the sub-terms' equivalence also show the same computational complexity of problems including only *encr* operator as one of problems including only *f operator*. The finite operator set guarantees the completeness of our new algorithm for invariant-sets' construction.

Obviously, the semantics of our new representation of invariant–sets is different from the former in the strand space model. In the definition of the smallest K-ideal, the bound condition $f(k, m_1, m_2, \cdots, m_{n-1}) \in I$ only satisfy to analyze the secret MAC payload. The new smallest K-ideal is reasonable according to Proposition 3.2. Of course, even if their computational complexity is equivalent, *encr* is different from *f operator*. When we add one operator to the operator set, what about the definition of the smallest K-ideal? We will further explore this key issue in modeling for security verification of cryptographic protocols.

We prove that the decidable theorem for honesty of ideals, which is already expanded, still holds as it is described under the extended sub-term relation. Even though the typical G. and G_f. attacks exist, it holds by checking the bound condition $B \setminus f(k, m_1, m_2, \cdots, m_{n-1}) \subseteq T \subseteq A$. Therefore, the modified strand space model in the present paper can be used to analyze complex sub-terms' architecture, such as one with MAC payload.

References

1. Michael Burrows, Martı́n Abadi, Roger Needham: A Logic of Authentication. Proceedings of the Royal Society, Series A, 426(1871) (1989) 233-271. Also appeared as SRC Research Report 39 and, in a shortened form, in ACM Transactions on Computer Systems 8, 1 (1990) 18-36
2. Kindred D.: Theory Generation for Security Protocols [Ph.D. Thesis]. Pittsburgh: Department of Computer Science, Carnegie Mellon University (1999)
3. Gavin Lowe: Breaking and Fixing the Needham-Schroeder Public-key Protocol Using FDR. In Proceedings of TACAS, Vol. 1055 of Lecture Notes in Computer Science, Springer Verlag (1996) 147-166

4. Millen J.: The Interrogator Model. In Proceedings of the 1995 IEEE Symposium on Security and Privacy. Los Alamitos: IEEE Computer Society Press (1995) 251-260
5. Meadows C.: The NRL Protocol Analyzer: An Overview. Journal of Logic Programming (1996) 26(2) 113-131
6. Lawrence C. Paulson: Proving Properties of Security Protocols by Induction. In 10th IEEE Computer Security Foundations Workshop. IEEE Computer Society Press (1997) 70-83
7. F. Javier Thayer Fábrega, Jonathan C. Herzog, Joshua D. Guttman: Strand Spaces: Proving Security Protocols Correct. Journal of Computer Security (1999) 7(2-3) 191-230
8. Abadi, M., Gordon, A.D.: A Calculus for Cryptographic Protocols: The Spi Calculus. Information and Computation (1999) 148 1-70
9. Gordon, A., Jeffrey, A.: Authenticity by Typing in Security Protocols. In 14th IEEE Computer Security Foundations Workshop. IEEE Computer Society Press (2001) 145-159
10. Gordon, A., Jeffrey, A.: Typing Correspondence Assertions for Communication Protocols. In Mathematical Foundations of Programming Semantics 17, Vol. 45 Electronic Notes in Theoretical Computer Science, Elsevier (2001)
11. Abadi, M.: Secrecy by Typing in Security Protocols. Journal of the ACM (1999) 46(5) 749-786
12. Abadi, M., Blanchet, B.: Secrecy Types for Asymmetric Communication. In Foundations of Software Science and Computation Structures, LNCS 2030. Genova, Italy: Springer-Verlag (2001) 25-41
13. Jinpeng Huai, Xianxian Li: Algebraic Model and Security of Cryptographic Protocols. Science in China (Ser. E) (2003) 33(12) 1087-1106
14. Harkins, D., Carrel, D.: The Internet Key Exchange (IKE), RFC2409 (1998). Available at http:// www.faqs.org/rfcs/rfcs2409.html

Bilingual Semantic Network Construction

Jianyong Duan, Yi Hu, Ruzhan Lu, Yan Tian, and Hui Liu

Department of computer science and engineering, Shanghai Jiaotong University,
200030, Shanghai, P.R. China
{duanjy, huyi516, lh_charles}@hotmail.com
ruzhan-lu@cs.sjtu.edu.cn, tianyan@sjtu.edu.cn

Abstract. This article proposes a neural network for building Chinese and English semantic resources connection. Abundant monolingual semantic information is stored into its bipartite graph structure respectively. Two hidden layers are also set in every part, word layer and concept layer. Every word associates with different concepts separately; every concept includes different vocabularies; and these two layers also independently connect to their counterparts through bipartite graph. These distributed characteristics in hidden layers meet the need of parallel network computing. The unsupervised method is used to train the network, and samples are translation lexicons, results of the bilingual word-level alignment algorithm. The training principle comes from the inspiration of bilingual semantic asymmetry. Every translational equivalent contains the unambiguous information by comparison between source and target languages. These translation lexicons are viewed as a kind of special context. They almost have definite meaning. Every input will activate and suppress various kinds of potential connections by the interaction of hidden layers, and modify their connective weights. Finally a demo test presents.

1 Introduction

In natural language processing field, further processing is dependent on the semantic knowledge database. A great number of monolingual semantic resources are built. WordNet[1], a famous online lexical reference system about English, nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. In this paper concept also refers to this kind of synonym set. Different relations link the synonym sets. And several languages are followed. EuroWordNet[2] is a multilingual database with WordNets for several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian). The WordNets are structured in the same way as the American WordNet for English in terms of synsets (sets of synonymous words) with basic semantic relations between them. Each WordNet represents a unique language-internal system of lexicalizations. In these occidental languages, WordNets can be connected to any other WordNet.

In orient languages, such as Chinese, Japanese, Thailand and Indonesia, some researchers attempted to connect these languages by handwork[3]. In their inter-lingual approach, a common set of concepts among languages is needed to represent the units of meaning, and relations to indicate the roles of the concepts in forming the

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 548 – 557, 2005.

intermediate representation. Words from each language were linked to the concepts by considering the provided description. It is time-consuming and boring work, and it also brings inconsistent problem. Semantic connections are complex even for monolingual. People cannot efficiently handle the reticulate relations by hands. Other groups also have made concept alignment across two different language hierarchies (Chinese and English). Dorr[4] described an approach to large-scale construction of a semantic lexicon for Chinese verbs. They leverage off of three existing resources, a classification of English verbs, a Chinese conceptual database, HowNet[5], and a large machine-readable dictionary. They use thematic role information to create links between Chinese concepts and English classes, and produced the concept-to-class correspondence. Chen[6] proposed a method to integrate five linguistic resources including English/Chinese sense-tagged corpora, English/Chinese thesauruses, and bilingual dictionary. Chinese words were mapped into WordNet. A Chinese-English WordNet derives by following the structure of WordNet. Their shortcoming is in need of sense tagged corpora. However these resources are scarce, and the robustness of sense tagged tools are limited in large-scale use. For example in Chinese part, it reported that the correct rate achieved 76.04%, and in English part, it reported a performance of 75%. But their integration results would be worse.

In this paper, we propose a machine learning method to build the bilingual concept connection[7,8]. Neural network is introduced into this field. The training principle comes from bilingual semantic asymmetry. Every translational equivalent contains the unambiguous information by comparison between source and target languages. In other words, the comparison of bilingual equivalent words is also a special context that has a suggestion for disambiguation[9,10]. These equivalent pairs are acquired by bilingual word alignment algorithm. The words from source and target languages maybe both have some senses. These senses from two parts interconnect and constitute candidate connection set. The potential connections that have stored in network will be activated among those candidate connections. While the other candidate connections will be inhibited at the same time. The potential connections will be turned into real connections with certain weight. Neural network has a built-in capability to adapt its weights to changes in the surrounding environment. The potential ability will emerge gradually in network.

In the following section we describe proposed method in details. Transformation for semantic resources in section 2, followed by Network construction Architectures in section 3, Initialization of Neural Network in section 4, its detailed learning process in section 5. Section 6 discusses a related demo test and we conclude with some thoughts on future research directions in section 7.

2 Semantic Resources Transformation

2.1 WordNet Concept Granularity Adjustment

In WordNet some senses are so close together that a separation line is hard to be drawn even by humans. These senses should be merged together as a synonym set. The fine granularity of senses defined in WordNet proves useless from an IR perspective. Other applications will also benefit from coarse granularity[11].

Rada[12] proposed a set of methods that enable the automatic transformation of WordNet into a coarse grained dictionary. On the one hand she proposed semantic principles to test ambiguity synsets and collapse them together into one single synset. On the other hand she adapted probabilistic principles to drop synsets with very low occurrence probability to decrease number of possible word senses. Based on her approach, we get the new version of WordNet with coarse senses. Every synset will contain more words. And every word will be less ambiguous.

2.2 Reorganization of Cilin and Hownet

There are two kinds of well-known semantic dictionaries in Chinese, Cilin[13] and Hownet. They have different mechanisms. Cilin is a synonym dictionary. It covers more than 70 thousands Chinese words. The words in Cilin are classified into different classes according to their meanings. There are 12 macro classes, 94 medium classes and 1428 minor classes altogether. Every word in Cilin is assigned a sense code that represents its meaning. HowNet is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. The knowledge structured by HowNet is a graph rather than a tree. These two kinds of dictionaries are different. Cilin is a hierarchical structure that every synset contains definite classification information about words. Sets of synonyms are the foundation stone of Cilin. It is similar to WordNet. However it is lack of abundant relations among synsets in Cilin while Hownet mainly focus on these features. The structure of Cilin is used as Chinese basic frame in this research. And the relations of Hownet are organized into this frame as a supplement.

3 Network Construction Architectures

3.1 Structure of Network

This network is bipartite graph structure. It consists of bilingual resources in two parts, Chinese part and English part. The monolingual semantic knowledge is stored into the two sub-networks separately. Every part keeps its own ontology. For example, Cilin as Chinese knowledge and WordNet as English knowledge are both implanted into this network. Prior information about monolingual semantic knowledge has been built in this network, thereby simplifying the network design by not having to learn them.

This is also a multilayer network including input layer, hidden layers and layers of output neurons. Hidden layers have two levels. The upper level in hidden layers is concept layer containing some relations about a synset, such as semantic hierarchy in source language and connecting information with counterpart neurons in target language. The lower level is word layer, includes its translational equivalents in target language and semantic classification information in source language. The two levels in hidden layers have communications through its semantic relation in monolingual knowledge. Every concept in concept layer is a synset. Concepts associated with words by their lexical semantic relations. The relations among these words are more than synonymy, it also can represent antonym and hyponymy through adding connecting link among them. Thus network also is a graph. Relations are expressed by these links.

Two sub-networks also connect each through two channels. One is translational equivalent in word layer. Two words from source and target languages will connect each other through correspondent translation. The other channel is potential concept connection in concept layer. The approach of concept connection is not by hand. Our approach is possible to capture the same idea as manual or supervised method in bilingual concept network construction, but avoiding the unreliability of human judge. Because the network has a natural ability for classification, this ability will be exploited directly through presenting a number of different training samples. Further research will be present in next section.

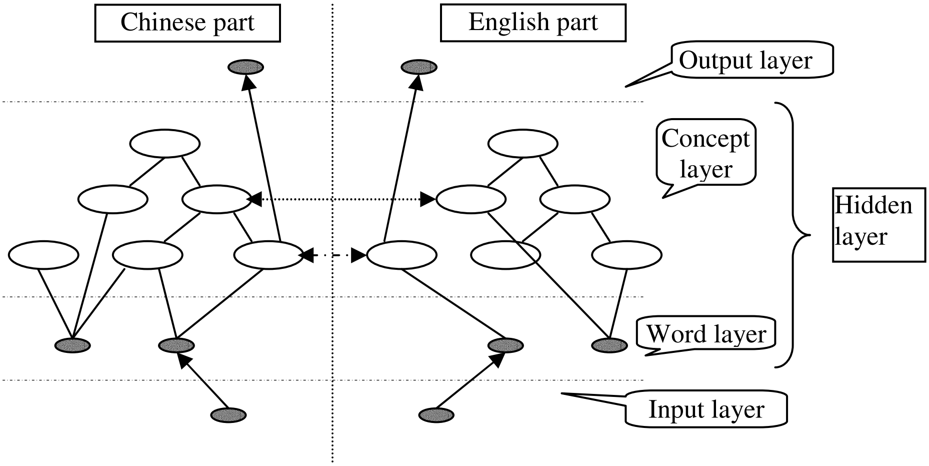


Fig. 1. Neural network with bipartite graph structure

4 Initialization of Neural Network

4.1 Bilingual Lexical Semantic Mapping

Different words can form special lexical semantic relations in a language. Words may make up of a semantic field under the domination of a common concept. Those words belonged to the same semantic field are interdependent and restrict each other. A word is meaningful and unambiguous in a given semantic field.

In different languages, semantic fields may be different because of their own ontology. English and Chinese have different vocabularies and semantic system respectively. Thus there are some difficulties in concept alignment. Such as concept composition, concept divergency and concept granularity problem. The strong connection links can’t be imposed on these concept pairs by hands. In this paper, we adopt soft concept alignment and allow uncertainty in the concept hierarchy, in order to combine diverse knowledge sources. Concepts from source language and target language can link each other through their words. These words are corresponding to each other through bilingual dictionary.

4.2 Communication Links Training

Communication links (CL) interconnect the bilingual concept nodes in the network, and they carry no weights. They merely provide some cross-language connection channels among those potential aligned concepts. These channels are not designated by hand. They are acquired through a training process with bilingual dictionary and monolingual semantic information.

The monolingual vocabulary has been divided into a lot of synonym sets. These words form different semantic fields. Elements in every semantic field have similar functions. The more correlative concepts are, the stronger their elements connect. And among those irrelevant concepts, the connection links will be disorderly and scattering. In the macroscopic, the potential aligned concepts will show stronger cross-language connections. The real connections between aligned concepts will emerge. Assume an English concept as:

E1 { plant, factory, manufactory, mill, shop, work…… }

Its element "plant" will have two equivalent translations which associate two Chinese concepts"工厂[gong chang]"in C1, and "植物[zhi wu]"in C2:

C1 {工厂、工业、产业、行业、车间……}
C2 {植物、蔬菜、菜、花、草…..}

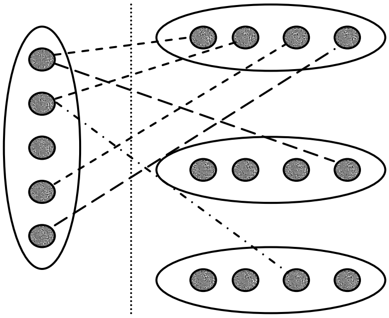


Fig. 2. Emerging of potential connection

Although every concept contains some ambiguous words whose translations are scattering in different concepts, its semantic field will focus on connections with few concepts in target language eventually. Strong connections will be activated, and weak connections will be suppressed. Then concept C1 has more word corresponding with E1, thus potential aligned concept pair (E1, C1) will have more steady connection than unrelated concept pair (E1, C2). An algorithm for mapping the related concepts is present as below:

Algorithm 1.

```
For each synonym set $y_k^s = \{x_1^s, x_2^s, \cdots x_i^s, \cdots\}$ {
   seek translational equivalents $T = \{x_1^t, x_2^t, \cdots x_i^t, \cdots\}$ }
```

```
$y_i^t \leftarrow y_i^t + \{x_j^t \mid x_j^t \in y_i^t\}$
For each ($x_i^t, y_j^t$) { $v(x_i^t, y_j^t) = \begin{cases} 1 & if \quad x_i^t \in y_j^t \\ 0 & if \quad x_i^t \notin y_j^t \end{cases}$ }
For each CL=$(y_k^s, y_j^t)$ { $score(y_k^s, y_j^t) = \sum_{i=1}^{n} v(x_i^t, y_j^t)$
If $score(y_k^s, y_j^t) > \varepsilon$ {Activated CL=$(y_k^s, y_j^t)$}}
```

5 An Unsupervised Learning Process

The property that is of primary significance for a neural network is the ability of the network to learn from its environment, and to improve its performance through learning. Interneuron connection strength to store the acquired knowledge can be easily learned through interaction in hidden layers. Every input as a stimulus will modify its topology. In this paper bilingual translational equivalents are used as training sample because bilingual resources acquired easier from Internet than sense-tagged bilingual corpus. Bilingual translational equivalent contains unambiguous information. The specified meanings of translational equivalents are also viewed as a kind of context because they have a hint for sense disambiguation. For example, "plant" is a ambiguous word in English, however in English-Chinese equivalent word pairs (plant, 工厂) and (plant, 植物), both "plant" have definite meaning.

The interaction of hidden layers can implement the unsupervised disambiguation among those candidate concept connections. Although there are $n \times m$ concept pair combinations for a translational equivalent, only one real connection will be activated among those candidate connections. The figure 3 shows the connection activation pattern among those candidate connections. Word token "a" in source language has three related concepts "A", "B", "C", and word token "b" in target language have two concepts "D", "E". In previous section we have proposed a communication links that have been stored in the neural network, such as a concept pair (B, D), that is the potential real connection between two concepts although the connection weight is zero. With the communicational connection constraint, the number of combinatorial connections will drop fast from candidate connection set. The real connection (B, D) will be activated and their connection weight will be modified.

A large number of translational equivalent pairs are needed for training our network because automatic word alignment tools prove useful for bilingual lexical extraction. Most word alignment approaches are automatic by statistical or dictionary based means. Here we adopt the class-based method to acquire translational equivalents from sentence-aligned corpus[14]. On the one hand it improves the coverage rate, its coverage exceed 80%. On the other hand it also guarantees the quality of translational equivalents, its precision exceed 93%. Through this method we get a large-scale of training samples.

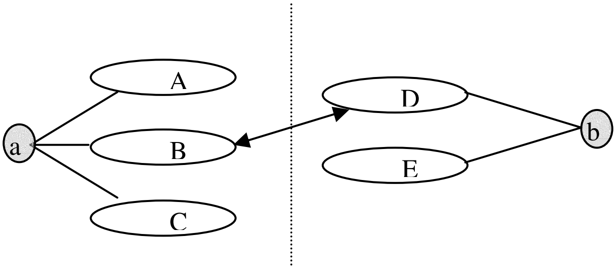


Fig. 3. Illustration of Activation Pattern

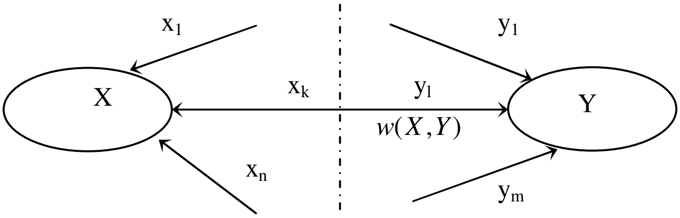


Fig. 4. Weight learning model

These translational equivalents input the neural network. Through the unsupervised disambiguation knowledge that stored in hidden layers as fig.3 shows, output layer will produce sense tagged pairs. The network also has a slight change. Interconnection modes among those related concepts are modified, as fig. 4 shows.

We consider cross-language links. Here x_k or y_l represents count of concept pair (X,Y) activated in the training procedure, the number is symmetrical, that is $x_k = y_l$. The counts of x_k and y_l will increase with every activation of concept pair (X,Y). And related $(m+n-1)$ concept pairs that contain X or Y also modify their weight distribution as equation (1).

$$w(X,Y) = \frac{x_k}{\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j - x_k} = \frac{y_l}{\sum_{i=1}^{n} x_i + \sum_{j=1}^{m} y_j - y_l} \tag{1}$$

Algorithm 2

```
Input: all the translational equivalents in
set $E = \left\{(x_1^s, x_1^t), \cdots (x_i^s, x_i^t) \cdots\right\}$
Repeat following processes until unable to continue
for each ( $(x_i^s, x_i^t) \in E$ ){
    $S \leftarrow S \bigcup \{y_i^s \mid x_i^s \in y_i^s\}$ and $T \leftarrow T \bigcup \{y_i^t \mid x_i^t \in y_i^t\}$
2 seek existed CLs between two sets T and E
3 if exist only one CL as $\left(y_a^s, y_b^t\right)$ {
   calculate $w(y_a^s, y_b^t)$ and related $(m+n-2)$ CLs}
4 if exist more CLs {
    select CL= $\arg\max w(y_a^s, y_b^t)$ ;
    calculate $w(y_a^s, y_b^t)$ and related $(m+n-2)$ CLs}
5 output sense tagged pair $(x_i^s, x_i^t)$, and E=E-{ $(x_i^s, x_i^t)$ } }
```

6 Results and Discussion

Only nouns are considered in our neural network. Originally WordNet 2.0 contains 114648 unique noun word tokens, 79689 synsets, and 141690 word-sense pairs. With Rada's method, we extract 16651 noun synsets excluding proper nouns and lower frequent words from WordNet. In Chinese part, 8672 nodes from Cilin and HowNet have been built in concept layer. We use 100 thousand bilingual sentence-aligned pairs to acquire the training samples. In table 1, we list vocabulary scales in different resources (Repeated tokens count only once).

Table 1. Vocabulary scales in different resources

Resources Language	Translational equivalent	Bilingual dictionary	Word layer stored
English	24,805	18,061	94,640
Chinese	16,791	13,702	37,740

6.1 A Demo Test

It is difficult to directly evaluate the effectiveness of network because the disambiguation ability comes from the interconnection among distributed nodes in hidden layers. One demo test is proposed to test the effectiveness of assumption in the network construction. We pick out some samples randomly, and designate their sense by hands according to the Cilin and WordNet respectively as correct sense labels. We view the network as a classifier and let it designate the meaning of these samples. Table 2 shows the results.

Table 2. A demo test result for the performance of network

Candidate links	Samples	Real connections	Precision	Coverage
Null	287	0	0.0%	23.7%
1	837	792	94.6%	69.2%
2	53	37	69.8%	4.4%
More	33	16	48.5%	2.7%
Total	1210	845	69.8%	100.0%
Without null link	923	845	91.5%	76.3%

In table 2, the results have confirmed the validity of the network model. The precision will reach to 91% without consideration of null link. It can ensure the effectiveness of learning process. It proves that the network has provided the most possible candidate links for training samples. The unsupervised learning that we proposed is an effective approach. But it still exist some problem. Such as the coverage of null link will be 23.7%. It means that many samples cannot be mapped in concept layers. We check every sample without communication links. Two major reasons are found. One is the sample quality problem that bad alignment pairs introduce noise. In other words, the learning process is also a filter procedure for word alignment results. The other is that semantic distances are far in samples. The problem of null link hopes to be improved through building the larger bilingual dictionary or acquiring a large-scale of translational equivalents.

7 Conclusions and Future Work

In this paper a valuable semantic network with disambiguation ability is acquired. This network has unsupervised training mechanism. In the input layer, it doesn't need the sense tagged samples. On the contrary, the training procedure also is the disambiguation procedure for sense tagging task. The disambiguation mechanism is the interconnection and constrain between word layer and concept layer.

We have obtained a distributed network with certain disambiguation ability. When translational equivalents input the network, the sense-tagged pairs will output as byproduct. It demonstrates that the network is efficient, and also can be used for further knowledge acquisition. Our future work will go on some aspects. First direction is to optimize the network model. We hope this network will more robust and also can mine its potential ability as much as possible. In this paper we only focus on nouns to build the network. Future work will let network cover all the vocabulary and hope this method to be used for other language pairs. The third direction is that we should perform this network into some broader application platforms. For example, these acquired bilingual sense-tagged pairs can be performed on cross-language information retrieval.

Acknowledgment

This research was supported by NSFC major research program, No.60496326. Basic Theory and Core Techniques of Non-canonical Knowledge and 863 Project of China, No.2001AA114210-11.

References

1. George, A. Miller, Christiane Fellbaum, etc: WordNet 2.0. Prinston University (2004)
2. Vossen, P.: EuroWordNet: Building a Multilingual Database with Wordnets for European Languages. ELRA News (1998)
3. Virach Sornlertlamvanich: Alignment of Concepts and the Hierarchies. Proceedings of the Third Meeting of Special Interest Group on AI Challenges (1999)

4. Bonnie Jean Dorr, Gina Anne Levow, Dekang Lin: Construction of a Chinese_English Verb Lexicon for Embedded Machine Translation in Cross Language Information Retrieval. Machine Translation (2002)
5. Zhendong Dong, Qiang Dong: HowNet. http://www.keenage.com (1999)
6. Hsin-Hsi Chen, Chi-Ching Lin, and Wen-Cheng Lin: Building a Chinese-English WordNet for Translingual Applications. ACM Transactions on Asian Language Information Processing (2002)
7. Wermter, S.: Hybrid Connectionist Natural Language Processing. Chapman and Hall, London, UK (1995)
8. Ma, Q., Zhang, M., Murata, M., Zhou, M., Isahara. H.: Self-Organizing Chinese and Japanese Semantic Maps. Conference on Computational Linguistics (2002)
9. Mona Diab and Philip Resnik: An Unsupervised Method for Multifingual Word Sense Tagging Using Parallel Corpora: A Preliminary Investigation. ACL-2000 Workshop on Word Senses and Multilinguality (2000)
10. Cong Li and Hang Li: Word Translation Disambiguation Using Bilingual Bootstrapping. Computational Linguistics (2004)
11. Martha Palmer, Olga Babko-Malaya, Hoa Trang Dang: Different Sense Granularities for Different Applications. Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding (ScaNaLU 2004) at HLT-NAACL (2004)
12. Rada Mihalcea: Turning WordNet into an Information Retrieval Resource: Systematic Polysemy and Conversion to Hierarchical Codes. International Journal of Pattern Recognition and Artificial Intelligence (2003)
13. Jiaju Mei, Yunqi Gao: Tongyi Cilin. Shanghai Dictionary Press (1983)
14. Ker Sur, J., and Jason, S. Chang: A Class-based Approach to Word Alignment. Computational Linguistics (1997)

Approaching the Upper Limit of Lifetime for Data Gathering Sensor Networks

Haibin Yu, Peng Zeng, and Wei Liang

Shenyang Institute of Automation Chinese Academy of Sciences
Shenyang, Liaoning, P.R.China
{yhb, zp, weiliang}@sia.cn

Abstract. Data gathering is a broad research area in wireless sensor network. In this paper, we consider the problem of routing between the base station and remote data sources via intermediate sensor nodes in a homogeneous sensor network. Sensor nodes have limited and unreplenishable power resources, both path energy cost and path length are important metrics affecting sensor lifetime. In this paper, we first explore the fundamental limits of sensor network lifetime that all algorithms can possibly achieve. Different from previous work, we explicitly consider the constraints of the limited energy and the limited end-to-end latency. We then model the formation of length and energy constrained paths and define the new composite metrics for energy-latency-optimal routing. We also design a distributed data gathering protocol called ELAG (Energy and Latency Aware data Gathering). This protocol balances energy consumption across the network by periodically determining a new optimal path consistent with associated energy distributions. Simulation results testify to the effectiveness of the protocol in producing a longer network lifetime.

1 Introduction

Wireless sensor network (WSN) of the future are envisioned to consist of hundreds of inexpensive nodes that can be readily deployed in physical environments to collect useful information in a robust and autonomous manner. However, there are several obstacles that need to be overcome before this vision becomes a reality. Such obstacles arise from the limited energy, computing capabilities and communication resources available to the sensors [1], [2], [3].

The basic operation of WSN is the systematic gathering of sensed data to be eventually transmitted to a base station for processing. The key challenges in such data gathering are: (i) energy efficiency: sensor nodes having limited and unreplenishable power resources. It needs to conserve the sensor energies, so as to maximize their lifetime; and (ii) latency awareness: events occurring in the environment being sensed may be time-sensitive. Therefore it is often important to bound the end-to-end latency of data dissemination.

An important problem in data gathering is to determine the lifetime that all algorithms can possibly achieve. Several authors have proposed several mathematical

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 558 – 567, 2005.

models for this purpose [4], [5], [6], [7]. However, to the best of our knowledge, none of these models providing upper bounds on the lifetime of sensor networks has explicitly considered bounding the lifetime under the end-to-end latency constraint. In this paper, we first explore the upper bound of sensor network lifetime that explicitly considers the constraints of the limited energy and the limited end-to-end latency. Then we propose a new composite metrics for energy-latency-optimal routing, and design a fully distributed routing protocol using the metrics to achieve a lifetime that is close to the derived upper bound. The protocol determines the optimal route by the guidance of limited global information on node energies obtained through user's command messages dissemination. While some control information piggybacked on command message arouses some extra overhead and consumes node energy, global information obtained through this process ensures a significant tradeoff in terms of node energy balancing and network lifetime.

We perform simulations to compare the derived lifetime upper bound and the actual lifetime achieved by our proposed protocol. The simulation results show that our protocol can achieve nearly 90% of the lifetime upper bound.

The rest of the paper is organized as follows. In section 2, we formulate the data gathering problem and define the upper limit of lifetime for data gathering sensor networks. In section 3, we describe a new composite metrics for energy-latency-optimal routing. In section 4, we present energy and latency aware data gathering protocol (ELAG). In section 5, we carry out simulations to compare the lifetime upper bound and the achieved lifetime using ELAG. In section 6 we conclude the paper.

2 The Data Gathering Problem

2.1 The Sensor Network Model

Consider a network of n sensor nodes and a base station node distributed over a region. As shown in Figure 1, the data from the nodes that detect the target needs to be collected and transmitted to the base station node. We assume that each sensor generates one data packet per time unit to be transmitted to the base station. For simplicity, we refer to each time unit as a round. For the sensor network is deployed in a big region, it needs to use multi-hop forwarding. Each sensor i has a battery with finite, unreplenishable energy E. Whenever a sensor transmits or receives a data packet it consumes some energy from its battery. The base station has an unlimited amount of energy available to it.

In the uniform distributed network, we partition the set of all sensor nodes V into subsets $S_0, S_1, \ldots, S_n$, satisfying $V = S_0 \cup S_1 \cup \ldots \cup S_n$, $S_i \cap S_j = \phi$ for all $i \neq j$ and no S_i is empty. S_i is the set of nodes that can be reached from the base station node B in i hops ($S_0 = \{B\}$), but not less than i hops. We call S_i the sphere of radius i around B.

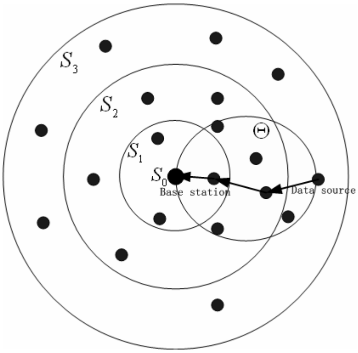


Fig. 1. Wireless sensor network for data gathering

2.2 Bounding Network Lifetime

A sensor network for data gathering can be in one of the following states:

1) Target present and network sensing while satisfying user dictated constraints (in this paper, we consider the limited end-to-end latency constraint). This state is termed "active".

2) Target present and network sensing but not satisfying user dictated constraints. This state is termed "quality failure".

3) End-to-end connectivity of the network is broken, no data can be sent to the base station node. This state is termed "connectivity broken".

In non-mission-critical applications, a reasonable definition of lifetime is the cumulative active time of the network until the connectivity broken (i.e. whenever the network is active its lifetime clock is ticking, otherwise not). In mission-critical applications, lifetime is defined as the cumulative active time of the network until the first quality failure. In this paper, we adopt this latter definition of lifetime.

According to the end-to-end latency constraint, path lengths will tend to be as small as possible. In a densely deployed sensor network, the shortest path should be in the straight line from the source node to the base station. Let Γ denote the latency constraint. Let $HopDelay_i^j$ denote the delay of forwarding a packet from node i to node j. In this paper, we assume the hop delay is the same along a path. The bound of the hop number of the data transmission paths is,

$$MaxLen = \frac{\Gamma}{HopDelay} \ . \tag{1}$$

Let Θ denote an ellipse. The diameter of Θ is the straight line from the source node to the base station, $|\Theta| = 2 \times MaxLen$,where $|\Theta|$ is the perimeter of Θ .To satisfy the latency constraint, all sensor nodes participating in routing should be located in Θ, as shown in Figure 1. When a sensor node in sphere S_n detects the target, it transmits exactly one packet in each round. A node in sphere S_{n-1} transmits the packets it receives from the data source node to another relay node in sphere S_{n-2} .

Corresponding to the spheres S , we introduce *balls* of radius i denoted B_i , with $B_i = S_0 \cup \ldots \cup S_i$. Further, we introduce $b_i = |B_i|$, $N = |V|$, r as the energy consumption for receiving one packet and t as the energy required to transmit one packet. Without loss of generality, we assume each sensor node in the network has the same probability $1/N$ to be a data source. Each sensor node in sphere S_i has the probability $P\Theta_i$ to participate in routing. Using these definitions, we set

$$m_i = \frac{1}{N} \cdot t + \frac{N - b_i}{N} \cdot P\Theta_i \cdot (r + t) \ . \tag{2}$$

In the equation above, $1/N$ is the probability of a node acting as a data source in sphere S_i . $N - b_i$ is the total number of nodes outside B_i and thus $N - b_i / N$ is the total number of packets that the set of nodes in sphere S_i receive and forward in each round. The best the routing algorithm can do is to balance the energy consumption for receiving and transmitting packets across all the nodes in S_i , therefore, $P\Theta_i$ is the same for all nodes in S_i . For example, Let Θ_n denote a ellipse derivated from a data source node in sphere S_n , which satisfies the latency constraint. Let $C_i = \Theta_n \cap S_i$, $c_i = |C_i|$. In this scenario, each sensor node in C_i has the same probability $P\Theta_j = 1/c_i$ to participate in routing. m_i provides a lower bound on the energy consumption (for receiving and transmitting packets) for the node in S_i during one round.

For most sensor networks, the energy consumption of the nodes in the bottleneck sphere for T rounds is $T \cdot \max\{m_1, m_2, \ldots, m_n\}$. In these cases, the traffic can be balanced evenly between the nodes in the bottleneck sphere. Hence, all nodes in the bottleneck sphere run out of energy during the same round, breaking connectivity.

For each node has the exact same amount of energy E , from the discussions above, it is obvious that the maximum number of rounds $T_{\max}$ a sensor network can perform before running out of energy under the given assumptions is bounded by the following expression:

$$T_{\max} \le \frac{E}{\max\{m_1, m_2, \ldots, m_n\}} \ . \tag{3}$$

This means, that whatever routing we use, the sensor network cannot perform more than $T_{\max}$ rounds before connectivity breaks.

3 Composite Metrics for Energy-Latency-Optimal Routing

The model described above does not make any assumption about how a route is found but instead provides an upper bound of the lifetime. In this section, we devise a new composite metrics for energy-latency-optimal routing that balances the energy consumption across all the nodes in the network to achieve the bound.

To find energy-optimal paths, we first define the following metrics which reflect dispersion or concentration of energy consumption across a network.

Average of energy level: The average of the energy level of all the nodes in the route. A high average indicates lower energy consumption at this route compared to others.

Variance of energy level: The variance of the energy levels of all the nodes in the route, which is the primary measure of dispersion. A high variance indicates higher energy consumption at some of the nodes compared to others in the route.

Let $l = \{n_1, n_2, \dots n_k\}$ be a route from a source node to the base station. The average of energy level of l is defined as:

$$E_{avg} = \sum_{i=1}^{k} E_{n_i} \Big/ k \ . \tag{4}$$

where E_{n_i} is the residual energy level of intermediate node n_i in l. k is the number of intermediate nodes in l.

The variance of energy level of l is defined as:

$$D_E = \sqrt{\sum_{i=1}^{k} (E_{n_i} - E_{avg})^2} \Big/ E_{avg} \ . \tag{5}$$

Now we define the composite metrics for energy-optimal routing:

$$C_E = E_{avg} \cdot (1 - D_E) \ . \tag{6}$$

where E_{avg} measures the residue energy level of the whole route. D_E is a percent, by which we avoid choosing a route has a node that is energy-deficient relative to other nodes.

To find latency-optimal paths, we use the bound of the hop number of the data transmission paths defined by Eq.(1). Let $D(l)$ be the length of l in terms of number of hops. We define the metrics for latency-optimal routing:

$$C_D = 1 - D(l) / MaxLen \ . \tag{7}$$

In this paper, we explicitly consider routing under both the constraints of energy efficiency and path length, and define the composite metrics for energy-latency-optimal routing:

$$C(l) = \alpha \cdot C_E + (1 - \alpha) \cdot C_D \ . \tag{8}$$

where $0 < \alpha < 1$. α should be chosen according to the specific requirement i.e. if $\alpha = 0$, it is for latency-optimal routing. If $\alpha = 1$, it is for energy-optimal routing.

4 Energy and Latency Aware Data Gathering Protocol

Energy and latency aware data gathering protocol (ELAG) uses the above composite metrics and implements energy-latency-optimal routing. For it is energy consuming to compute D_E which needs not only the average of energy level of the route but also the variance of energy level of every node in the route. To reduce the overhead of route selection, we redefine the variance of energy level of the route:

$$D_E = (E_{avg} - Min(E_{n_i}))/E_{avg} \quad . \tag{9}$$

where $Min(E_{n_i})$ is the minimum residual energy level.

The proposed protocol operates in two different phases: global information collection and path determination, global information refreshing as described below.

4.1 Global Information Collection and Path Determination

During this phase, user's command messages are transmitted from the base station to other sensor nodes through all possible paths by flooding. Each data packet potentially collects information about the energy consumption en route by keeping track of residual energy levels of nodes on the path. The fundamental steps of the command dissemination are as follows:

– User first disseminates the sensing task to WSN through the base station node by flooding. Each sensor node will receive the command message from all possible paths originated from the base station, and forward the command message to all its neighbors. Each command message is piggybacked with some information about the path it passed by, like the length of the path in terms of number of hops, denoted by Hop , the residual energy level of the path, denoted by E_{all}, and the minimum residual energy levels of the intermediate node in the path, denoted by $E_{\min}$ .The piggybacked information is used to find the energy-latency-optimal routing paths according to the new metrics.

– In the process of sensing task dissemination, each intermediate node stores the information of the energy-latency-optimal path according to the metrics seen so far. For example, node i uses two parameters C_i and P_i to record the information of path l with the highest $C(l)$ value. C_i is used to record the $C(l)$ value. P_i is used to record the neighbor node ID on l . Initially, C_i is set to zero, P_i is set to null. When node i receives the command message from one of its neighbors j , there exists a path l from i to the base station node through j . i takes out the piggybacked information and computes the overhead of the path according to the metrics as follow:

  First, updates the path information.

  $hop(l) = hop(l) + 1$, $E_{all}(l) = E_{all}(l) + E_i$.

  If $E_{\min}(l) > E_i$, then $E_{\min}(l) = E_i$.

  Second, calculates the $C(l)$ value using Eq.(4), Eq(8) and Eq.(9):

  $E_{avg}(l) = E_{all}(l)/hop(l)$.

  $D_E(l) = (E_{avg}(l) - E_{\min}(l))/E_{avg}(l)$.

  $C(l) = \alpha \cdot E_{avg}(l) \cdot (1 - D_E(l)) + (1-\alpha) \cdot (1 - hop(l)/H_{\max})$.

  Third, compares C_i with $C(l)$.

If $C_i < C(l)$, it means the path from i to the base station through j is more optimal than that through P_i, then update the path information stored in i, let $C_i = C(l)$ $P_i = j$, rebroadcast the command message piggybacked with the updated path information.

4.2 Global Information Refreshing

WSN is closely tied to the ever-changing physical world, the system will experience extreme dynamics. To keep high quality data gathering, it is necessary to refresh the path information stored in each intermediate node. The global information refreshing in ELAG has two paradigms:

1) Refreshing originated from the base station node

If the quality of received reports is poor (long delay, high loss rate, high bit error rate), the base station node starts the global information refreshing process by broadcasts a refreshing command which is the same as the process of user's command dissemination. Refreshing originated from the base station updates the path information stored in all sensor nodes.

2) Refreshing originated from sensor node

Each sensor node determines whether its energy level has fallen below the threshold th. If so, it starts the global information refreshing process by broadcasts a refreshing command piggybacked with the updated path information. The refreshing originated from sensor node only updates the global information in the nodes that have a path to sink through the originator.

The threshold value *th* plays a very important role in the global information refreshing phase since it is used to provide an approximate indication that the current optimal path has become obsolete. In this paper, *th* is defined as:

$$th = \beta \cdot E_{\min} \ . \tag{10}$$

where $0 < \beta < 1$ and $E_{\min}$ is the minimum energy level in the current optimal path. Since $E_{\min}$ changes with time, the threshold is recalculated in each round, consistent with the current energy distribution across the network.

5 Simulations and Analysis

In this section, we carry out several sets of simulations to investigate how much percent of the upper bound can be achieved by the protocol we proposed.

For our simulation, we created the 50-node network which was manually deployed in a 1000m by 1000m square. Some detailed parameters are described in table 1.

In our evaluation, we compare three protocols for energy efficiency: directed diffusion [8], GEAR [9] and the proposed ELAG. We present the following set of result: 1) system lifetime, and 2) control overhead.

Table 1. Simulation settings

Name	Parameter
MAC Layer	802.11(Simplified DCF□
Bandwidth	10 Kb/s
Packet size	160 bytes
Antenna reach	200 m
Average degree	6.9 neighbors
Average shortest path from the sensor nodes in sub-region 0 to sink	8 hop
Initial energy per node	0.02 J

5.1 System Lifetime

We assume that before the network starts any activity, all ordinary sensor nodes have the same energy level. Therefore, in the very beginning, energy distribution is uniform across the network.

We calculate the energy consumption for receiving and transmitting one packet based on the first order radio model described in [10], $t = 1.58 \times 10^{-4} J$, $r = 6.4 \times 10^{-5} J$. We divide the sensor network into 9 spheres according to the radio transmission range: $V = S_0 \cup S_1 \cup \ldots \cup S_9$, $S_0 = \{B\}$. For simplicity in the computation of the lifetime below, we assume all possible paths from the data source to the base station can satisfy the latency constraint, so all sensor nodes have the same chance to participate in routing. For data sources are located in sub-region 8, other sensor nodes will act as the relay nodes in this scenario. We find the bottleneck sphere S_3 using Eq.(2) ,and compute the upper bound of the lifetime using Eq.(3):

$$m_3 = 2.22 \times 10^{-4} J$$

$$T_{\max} \le \frac{E}{m_3} = 450.45s$$

When a network becomes active, the energy distribution across it gradually becomes non-uniform since nodes participating in a route inevitably consume more energy than other nodes. ELAG tries to adapt to the dynamically changing energy distribution and gradually uniforms the initial uneven energy distribution. The routing metrics of ELAG takes into account the residual energy level of the whole path which is different from pre-proposed data gathering protocols like GEAR. GEAR chooses a neighbor with the most remaining energy to forward data which can only balance the energy consumption locally. But ELAG can balance the energy consumption across the whole sensor network and finally prolong network lifetime.

In table 2, we present the difference in the system lifetime of the three protocols. ELAG achieves 89.25% of the upper bound of the lifetime, which is the highest one in the three protocols.

Table 2. System Lifetime

Protocol	System lifetime	Percent of the upper bound can be achieved
Directed diffusion	372.0077 s	82.58%
GEAR	392.0138 s	87.02%
ELAG	402.0073 s	89.25%

5.2 Control Overhead

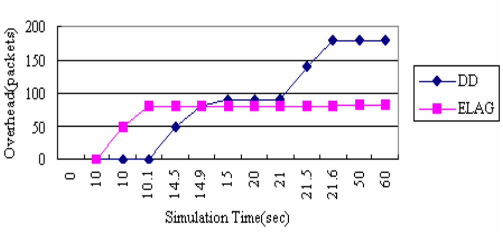


Fig. 2. Control overhead with directed diffusion and ELAG

In ELAG, control information is piggybacked into the command packets to set up path information in each intermediate node, which arouses some extra overhead. Global information obtained through this process ensures the data source node quickly find the energy-latency-optimal routing paths. Fig.2 shows the control overhead comparison of ELAG and directed diffusion which needs feedback information in the process of iterative routing optimization. In this sense, ELAG is a proactive routing paradigm while directed diffusion is a reactive routing paradigm. From Figure.2 we can see the overhead of directed diffusion is double of ELAG.

6 Conclusions

The key challenge in networks of energy constrained wireless integrated sensor nodes is maximizing network lifetime. In this paper, we use an idealized model to study the upper bound of the lifetime for large scale sensor networks, which explicitly considered the constraints of the limited energy and the limited end-to-end latency. In the model, we quantify the fundamental role played by the spheres in determining the energy consumption of routings in an ideal environment where all nodes have the same radio transmission range and transmit with the same energy. We then define a new composite metrics for energy-latency-optimal routing, and propose a distributed protocol that finds optimal routes. The ELAG protocol balances energy consumption across the network by selecting new optimal paths periodically. The simulation results indicate effectiveness of this protocol for enhancing network survivability.

Acknowledgment

This work is supported by National Natural Science Foundation of China (No. 60374072, 60434030)

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: a Survey. Computer Networks. 38 (2002) 393-422
2. Stankovic, J.A., Abdelzaher, F., Lu, C.Y., Sha, L., Hou, J.C.: Real-Time Communication and Coordination in Embedded Sensor Networks. In Proceedings of the IEEE, Vol. 91, No. 7 (2003) 1002- 1022
3. Yu, H.B., Zeng P., Wang, Z.F., Liang, Y., Shang, Z.J.: Research on Communication Protocols of Distributed Wireless Sensor Network. Journal of China Institute of Communications. Vol. 25, No 10 (2004) 102-110
4. Bhardwaj, M., Chandrakasan, A., Garnett, T.: Upper bounds on the lifetime of sensor networks. In: Proceedings of ICC2001, IEEE (2001) 785-790
5. Bhardwaj, M., Chandrakasan, A.: Bounding the lifetime of sensor networks via optimal role assignments. In Proceedings of INFOCOM2002, IEEE (2002) 1587-1596
6. Ritter, H., Schiller, J., Voigt, T., Dunkels, A., Alonso, J.: Experimental Evaluation of Lifetime Bounds for Wireless Sensor Networks. In Proceedings of EWSN2005, Istanbul, Turkey (2005)
7. Zhang, H.H., Hou, J.C.: Approaching the Upper Limit of alpha-lifetime for Wireless Sensor Networks. In: Proceeding of TAWN2004, Chicago, IL, USA (2004)
8. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks. In Proceedings of MobiCOM2000, Boston, Massachusetts (2000) 56-67
9. Yu,Y., Govindan, R., Estrin, D.: Geographical and Energy Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks. UCLA Computer Science Department Technical report UCLA/CSD-TR-01-0023 (2001)
10. Heinzelman, W., Chandrakasan, A.P., Balakrishnan, H.: Energy-Efficient Communication Protocols for Wireless Microsensor Networks. In Proceedings of Hawaiian International Conference on Systems Science (2000)

A Scalable Energy Efficient Medium Access Control Protocol for Wireless Sensor Networks

Ruizhong Lin, Zhi Wang, Yanjun Li, and Youxian Sun

National Laboratory of Industrial Control Technology,
Institute of Modern Control Engineering,
Zhejiang University, Hangzhou 310027, P.R.China
{rzlin, wangzhi, yjli, yxsun}@iipc.zju.edu.cn

Abstract. In this paper, we propose a scalable energy efficient medium access control protocol (SEMAC) based on time division multiple access (TDMA) technique for wireless sensor networks (WSNs), which uses the local information in scheduling, eliminates most collisions, is more energy efficient and is scalable to the number of sensor nodes in WSN. SEMAC uses the concept of periodic listen and sleep in order to avoid idle listening and overhearing. To balance the energy used in the whole network, SEMAC lets the node with lower energy be a winner in an election procedure based on their energy levels and the winner has more chances to sleep to save energy. We also use a clustering algorithm to form clusters so as to increase the scalability of SEMAC. The performance of SEMAC is evaluated by simulations, and the results show the gain in energy efficiency and scalability.

1 Introduction

Wireless sensor networks (WSNs), which are made by the convergence of sensor, micro-electro-mechanism system, wireless communication and distributed computing technologies, have been an active research area for a number of years now and received increasing interest due to their great numbers of potential applications [1] [2]. Most of the attention, however, has been given to the medium access control (MAC) protocols since they pay an important role in wireless communication for saving energy in a single node and one of the most important requirements in designing a WSN is low power. In most applications, the nodes are unattended and live only as long as their batteries can support. Thus the sensor node must use the limited resources effectively and manage the energy to extend the lifetime of the network as much as it can. Therefore, energy management is a challenging problem in designing a WSN [2]. Studies reveal that energy wastage occurs mainly from collision, overhearing, control packet overhead and idle listening [3]. These problems are present in all shared-medium networks and are generally taken care of by MAC techniques. The main goal of a MAC protocol in WSN is to allocate the shared wireless channels among sensor nodes as fairly as possible and to guarantee that no two interfering nodes transmit at the same time [3]. And the essential consideration of MAC protocols designing for

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 568–577, 2005.

WSN is energy efficiency as the power in a sensor node is limited and wireless communicating uses much more energy than sensing and computing.

Designing MAC protocols in WSN has become a broad research area and current MAC protocols designed for WSN can be classified into four categories: 1) *scheduling based*; 2) *collision free*; 3) *contention based*; and 4) *hybrid schemes* [4]. Scheduling (mostly time division multiple access, TDMA) based MAC protocols for WSN are seemed to be a promising technology since they are simple and can avoid collision in most cases and other classes of MAC protocols have this or that disadvantage which will be discussed in a little more detail in the next section. Kalidindi et al. proposed DE-MAC [3] which is based on TDMA and uses an electing mechanism to let the node with lower energy level has more chances to sleep to reduce energy usage [3].

However, in DE-MAC, the neighbor nodes have to keep active in listening the data from the node which owns the current slot(s) even if it has nothing to send. This unwanted listening wastes energy. And the scalability is not considered in DE-MAC, as it is important in large scale WSN. In this paper, we propose a scalable energy efficient medium access control (SEMAC) protocol for WSN to cope with these problems.

The rest of paper is structured as follows. Section 2 presents some relative work. In section 3, our proposed SEMAC is described. In section 4, performance metric of energy efficiency and scalability is evaluated through simulation in WSNSim. Finally, we make some concluding remarks.

2 Related Work

The advantages and disadvantages of the four classes of MAC mentioned above are reviewed and discussed as follows, with the requirements of *Real-time*, *Quality-of-Service*, *Decentralized*, *Power aware*, *Flexibility* and *Balance among multiple metrics* for MAC technology needed by WSN in considering [4]. Collision free protocols (e.g., the Implicit Prioritized Access Protocol [4] and TRAMA [4])are surely noteworthy because they save power by eliminating collisions. The problem in current collision free protocols is the use of multiple channels, which results in the nontrivial requirement on nodes and increasing of hardware cost. Another concern is the complexity of the protocol. Normally, a simple protocol is preferred because of the limited resource of nodes in the network [4].

Contention based MAC protocols often have difficulty in providing real-time guarantees and collisions also waste energy. However, there have been some advances in this area which can largely mitigate chances of collisions and reduce power consumption, e.g., T-MAC [4]. This could be useful in some applications where predictability is less critical and power consumption is main concern. On the other hand, for contention based protocols to be successfully used in WSN, a well-defined statistical delay bound is still needed [4]. And the hybrid protocols, which integrate more than one approach's advantages, may be useful in meeting the requirements in WSN. Yet the complexity of the protocols is still a problem and it must be considered to eliminate the disadvantages of these different kinds of methods.

Scheduling based (especially TDMA based) MAC protocols provide fair usage of the wireless channel and, if equipped with an adequate scheduling algorithm, could also avoid collisions [4]. But there are some shortcomings that make it difficult for many TDMA based MAC protocols to be broadly used in WSN. Firstly, some TDMA protocols use global information to do scheduling, which render those protocols impractical in general WSN, especially large scale WSN. Secondly, some protocols still have collisions, and it is quite difficult to control the collisions to the degree that does not hurt the guarantee of timeliness. Thirdly, the energy is wasted in collisions. Finally, the scalability of TDMA is still a problem to be used in large scale WSN.

The main idea of DE-MAC [3] is to let the nodes exchange information about their energy levels. The node with the lowest energy level will declare itself as a winner in an election and it owns two slots to transmit the data. When a node owns current slot(s) has nothing to send, it goes to sleep immediately, while its neighbors have to keep awake to listen to the channel according to their local scheduling table. In this way, the node with lower energy level will have more chances to sleep to save energy, thus balancing energy among the nodes and prolonging the lifetime of the whole network [3].

In DE-MAC, a node has to keep the radio awake in the slots assigned to its neighbors in order to receive packets from them even if the node owning current slot(s) has nothing to transmit. This needless listening wastes energy. Another problem of DE-MAC is that the scalability is not considered.

To deal with these problems in DE-MAC, we propose a MAC named SEMAC for WSN, since it is a Scalable Energy Efficient Medium Access Control Protocol. In SEMAC, the nodes the nodes listen to the channel for a while and if listen nothing, they also go to sleep. Thus, more energy can be saved to prolong the lifetime of WSN. SEMAC also adopts a clustering algorithm to form clusters to increase the scalability and uses a Listen-Listen-Send (LLS) mechanism to avoid collisions of the nodes in the borders of the adjacent clusters.

To evaluate the performance of the protocols and algorithms designed for WSN, a simulation environment is needed. Current simulation environments designed for WSN are as follows: SensorSim , TOSSIM , NS-2, EmStar, PowerTOSSIM [5] and et al. Few of these simulation environments have considered power consumption. SensorSim incorporates simple power usage and battery models, but does not appear to have been validated against actual hardware and real applications. TOSSIM provides a scalable simulation environment for WSN based on TinyOS. SensorSim, NS-2, and EmStar are all machine-level simulators, while TOSSIM is integrated in TinyOS and compiles a NesC application into a native executable that runs on the simulation host [5]. However, TOSSIM still does not provide any information on the power consumption, and in PowerTOSSIM, each NesC code of the nodes in simulating must be the same. In this paper, we develop a simulation framework named WSNSim, which is based on the energy model of Mote platform and the operation in each node can be defined to evaluate the performance of MAC protocols in condition monitoring applications.

3 SEMAC Protocol Design

SEMAC is based on TDMA and hence possesses the natural ability of avoiding extra energy wastage since the time at which a node can transmit data is determined by a TDMA algorithm and multiple nodes can transmit simultaneously without interference on the wireless channel. Time is divided into slots, and a node is assigned at least one slot to transmit. The length of a time slot can be determined according to different kinds of applications. Taking condition monitoring applications as an example, the time slot can be quite long since the data acquiring is not very frequent. The main advantages of a TDMA based protocol present in SEMAC are as follows [3].

- Packet loss due to collisions is absent because nodes do not transmit in the same slot.
- No contention mechanism is required for a node to start sending its packets since slots are preassigned to each node.
- No extra control overhead packets for contention are required.

The protocol contains two main phases, topology discovery and election progress. The former is to initialize the scheduling table of TDMA and to rediscover topology when some node fails or some node adds in. The latter is to elect the node with lowest energy level and assign more chance for it to sleep to save energy. The protocol is described in detail in the following subsections. Subsection 3.1 gives some definitions of the packets used in SEMAC. Subsection 3.2 describes the initialization of TDMA scheduling table and subsection 3.3 presents the algorithms of electing, sleeping and clustering.

3.1 Definitions of Packets in SEMAC

There are two kinds of packets in SEMAC in general, control packet (CP) and data packet (DP).Data Packet is the normal data packets received from higher layer protocols or generated by the node itself, which are to be routed to the base station. And in the condition monitoring application, the data are not very frequent. DP includes 2 bytes address, 1 byte type of the data, 1 byte cluster number, 1 byte length of the packet, 1 to 29 bytes of data, and 1 CRC byte.

Each CP contains three fields. The first byte - Type Field (TF) specifies the type of the control packet, and the following one byte - Cluster Field (CF) specifies the cluster number of the nodes, and the last field - Data Field (DF) is the data of the control packet. The control packets can also be classified into two types: Scheduling Packet (SP) and Voting Packet (VP). SP is used in forming a scheduling table while VP is used in voting for election of the lowest-energy-level winner. There are three kinds of SP, Request ID Packet (RIP), Sending ID Packet (SIP) and Sending Table Packet (STP) which can be defined as follows.

- RIP (TF = 1): a node sends RIP to its neighbors to start a topology discovery phase. DF includes 1 byte and specifies the ID number of the node who starts the discovery.

- SIP (TF = 2): a node sends SIP to the sponsor of the discovery. DF contains 1 byte, which is the node's ID number.
- STP (TF = 3): the node that starts the discovery sends STP to its neighbors the scheduling table it has built. DF contains several bytes including 1 byte to indicate the length of the data field and the scheduling table it has just built. The scheduling table includes the ID number and time slot(s) of each node.

Voting packets include Requiring Vote Packet (RVP), Sending Vote Packet (SVP), Winning Vote Packet (WVP) and Losing Vote Packet (LVP). Voting packets can be defined as follows.

- RVP (TF = 4): a node sends RVP to its neighbors to launch an election. DF includes 3 bytes and specifies the energy level of the node who wants to start the election.
- SVP (TF = 5): a node sends SVP to the sponsor of the election to vote. DF contains only 1 byte, which can be either positive vote (value = 55H)or negative vote (value = AAH).
- WVP (TF = 6): a node sends WVP to its neighbors to declare itself as a winner of the election. DF contains 3 bytes and specifies the energy level of the winner node.
- LVP (TF = 7): a node sends LVP to its neighbors to declare itself as a loser of the election. DF contains 3 bytes and specifies the energy level of the loser node.

3.2 Initialization of TDMA Scheduling Table

When sensor nodes are deployed randomly in the interest area by unmanned aircraft, the first task is to discover the neighbors to form the topology of the network. The main idea of the scheduling table initialization is to let the nodes exchange the identity (ID) numbers. Based on the ID numbers, the node with the smallest ID number will be in charge of assigning the time slots and setting up a scheduling table. Algorithm for the nodes to discover the topology is described as shown in Algorithm 1. And the topology rediscovering algorithm is similar.

3.3 Protocol Description

The nodes may be in any of the three phases after topology discovery and scheduling table forming: normal operation phase (NOP), voting phase (VOP) and topology rediscovery phase (TRP). In NOP, the nodes operate normally, routing data packets to the base station. And in VOP, critical nodes enter the voting phase to do a local election to readjust their slots. A node in the voting phase is integrated with the normal TDMA phase. So, control packets are sent along with normal data packets in the voting phase.

Voting Algorithms. The local voting phase is triggered by criticality of a node. A node is said to be critical if it falls below the previous election winners energy value. When a node enters this critical phase a local voting phase is triggered. A node in the voting phase is a winner if all its neighboring nodes energy levels are greater than its own energy level. Otherwise it is declared as a loser. The algorithm for a node i to operate in the voting phase is presented in Algorithm 2. And the algorithm for a receiver node j to operate in the voting phase is described in Algorithm 3.

Listening Algorithm. In normal operation phase, the activity of a node x in the time slot(s) is shown in Algorithm 4, where $\delta (0 < \delta < 1)$ is a parameter to adjust the wait period for node x to listen the channel.

Clustering Algorithms. To improve the scalability of SEMAC for using in large scale WSN, we adopt a simple listening based clustering algorithm to form clusters. In each cluster, the sensor nodes operate following the algorithms described above. The simple clustering algorithm is presented in Algorithm 5.

We use a cluster number in the packets to avoid collisions between clusters. And to avoid collisions of the nodes between the adjacent clusters, a Listen-Listen-Send (LLS) backoff mechanism is designed as shown in Algorithm 6, where $\alpha (0 < \alpha < 1)$ is a random factor for a node in the cluster border to backoff.

4 Simulation Environment and Results

To evaluate the performance of MAC protocols in condition monitoring applications, we develop a simulation framework named WSNSim, which is based on the energy model of Mote platform and the operation in each node can be defined. We perform the simulations in WSNSim and compare the average energy used by DE-MAC and SEMAC and a simple TDMA to evaluate the energy efficiency performance in different node numbers and different wait period (i.e., different δ). We also evaluate the scalability of SEMAC without Clustering and SEMAC with Clustering.

4.1 Simulation Framework: WSNSim

WSNSim is a component based, event driven runtime and Mote energy modeling simulation framework to simulate the energy used in each node for condition monitoring applications. The components in WSNSim are similar to Mote developed by University of California at Berkeley, which include CLOCK, SENSOR, ADC, LED, RADIO and APPLICATION.

The power model used in WSNSim is from MICA2 node with sensor board in a 3V power supply according to [5].

4.2 Evaluation of Performances

In our simulations, we set the nodes number N from 10 to 1000 for different factors of δ from 0.005 to 0.8, and simulate 50000 second for each case, in which the one time slot T_{slot} is set to 1s. The sensor data in each node is generated

Algorithm 1 BuildTAB

```
Wait for some time according to its own ID number
if not received RIP, then send RIP and wait for SIP
end if
if received RIP, then send SIP
end if
if received SIP and waited for enough time then build the scheduling table
    according to ID numbers
end if
send STP
```

Algorithm 2 RequestVOTE

```
if energy level of node $i$ < previous election winner's energy level then sends RVP
end if
if node $i$ received all SVP from its neighbors and each DF in SVP = 55H
then update the scheduling table and sends WVP
else if node $i$ received all SVP from its neighbors and
    some of DF in SVP = AAH
then send LVP
else if node $i$ not received all SVP from its neighbors or received extra SVP
then node $i$ send RIP to start topology rediscovery
end if
```

Algorithm 3 SendVOTE

```
if energy level of node $j$ > energy level sent by node $i$
then node $j$ send SVP with DF = 55H
else node $j$ send SVP with DF = AAH
end if
if node $j$ received WVP from node $i$
then update the scheduling table
else if node $j$ received LVP from node $i$
then node $j$ check its energy level to decide whether entering
    a voting phase or not
end if
```

Algorithm 4 ListenCHANNEL

```
if node $x$ owns the current slot then node $x$ sends its DP
    if node $x$ has nothing to send then node $x$ go to sleep
    end if
else if node $x$ not own the current slot then node $x$ check the scheduling table
    if one of node $x$'s neighbors is transmitting in current slot
      then node $x$ listen for $\delta * T_{slot}$ time
        if node $x$ received nothing in $\delta * T_{slot}$ time then node $x$ go to sleep
        end if
    end if
end if
```

Algorithm 5 Clustering

```
1: node $x$ becomes Remote to listen for a Head of cluster
2: if node $x$ listened a Head of Cluster then node $x$ join the cluster and go to step 9
3: else if node $x$ not listened a head after the listening period
4: then node $x$ declare itself as a Head and send for listeners
5:    if Head $x$ connected other listeners then go to step 9
6:    else if node $x$ connected no listeners then go to step 1
7:    end if
8: end if
9: Clustering finished
```

Algorithm 6 LLS

```
1: if node $x$ often listens other cluster packets then node $x$ go into LLS phase
2: end if
3: node $x$ listen to the channel for a period $\tau$
4: if node $x$ listened nothing then node $x$ backoff for a random period $\alpha * \tau$ and listen
   to the channel again
5:    if node $x$ listened nothing then node $x$ send its DP
6:    else if node $x$ listened something then go to step 3
7:    end if
8: else if node $x$ listened something then go to step 3
9: end if
```

in a uniform distribution, that is, averagely in 2s, the node will have 1 data to send.

Energy Efficiency: Fig.1 shows the effect of different δ in SEMAC, where nodes number $N = 100$, and different average energy used in DEMAC and SEMAC, and in TDMA and SEMAC is compared with different δ from 0.005 to 0.8 (i.e., t_{wait} is from 5ms to 800ms). The results in the left figure of Fig.1 show that different energy between DEMAC and SEMAC will decrease when δ increases, but the difference is still distinct even when the simulating time starts a little period and δ is 0.8. While in the right figure of Fig.1, the results shows that different energy between TDMA and SEMAC is little when the simulating time starts, and it will increase with the time's going on. This is due to the characteristic of condition monitoring application, since the sensor data generated by a uniform distribution and the simple TDMA protocol gives the same chance for each node to sleep.

Fig.2 shows the different energy used between three protocols with the same δ 0.005 and different nodes number N from 10 to 1000. The results show that when N increases, difference between DEMAC and SEMAC also increases. While the difference between TDMA and SEMAC appears change slowly. This is also because the simple TDMA is fair to each node while DEMAC and SEMAC give more chances to the nodes with lower energy levels.

Scalability: To verify the scalability of SEMAC, we compare average energy used in SEMAC with and without Clustering when the nodes number increases.

Fig.3 shows the different average energy usage in SEMAC without and with clustering in different nodes number N from 200 to 1000 where δ is 0.005. From the figure we can see that the difference increases with the increasing of N.

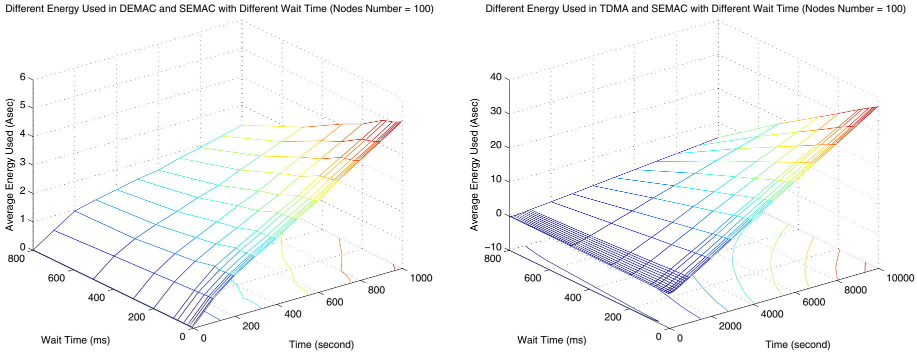


Fig. 1. Different average energy used comparison of DEMAC and SEMAC, TDMA and SEMAC with nodes number 100, different δ from 0.005 to 0.8

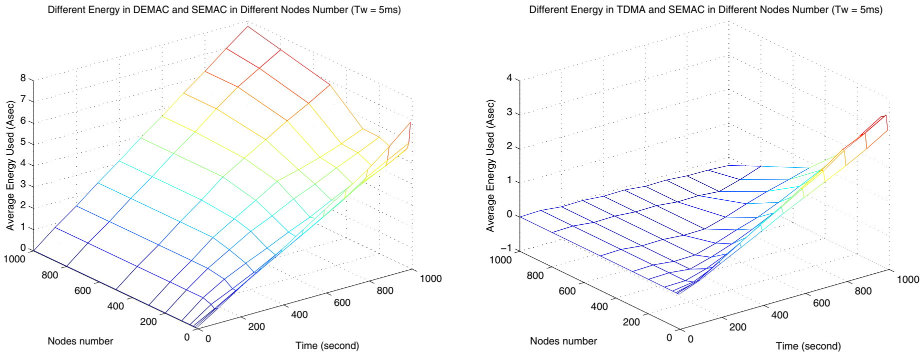


Fig. 2. Different average energy used comparison of DEMAC and SEMAC, TDMA and SEMAC with δ=0.005, different nodes number from 10 to 1000

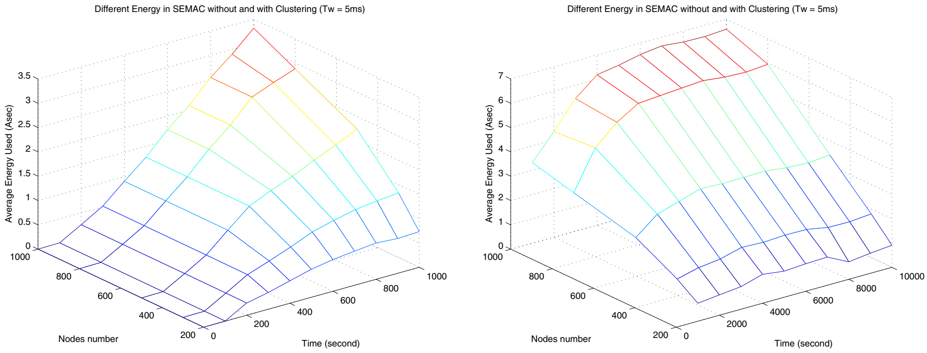


Fig. 3. Comparison of SEMAC without and with Clustering in different nodes number from 200 to 1000 where δ=0.005

5 Conclusions

The main original contributions of this paper are: 1 Propose SEMAC including TDMA table initialization, voting and channel listening algorithms to meet the requirements of condition monitoring applications. 2 Show that SEMAC with listening period factor δ can save more energy than DE-MAC and TDMA by simulating in WSNSim, a framework for simulating large scale WSN. 3 Show that SEMAC with clustering algorithm is more scalable.

The energy efficiency improvement is made with a very low computing cost or complexity, only to set a timer of $\delta * T_{slot}$, which is needed for a node to listen to its neighbors when it does not own the current time slot(s). Furthermore, the simple clustering algorithm increases the scalability of SEMAC with little communication throughput since clustering phase is integrated in the listening phase and collisions are avoid by using cluster number in packets and adopting LLS backoff mechanism in border nodes. In this sense, our protocol is interested in increasing the scalability with guaranteeing of energy efficiency. Furthermore, performance evaluations in other kinds of applications, e.g., break-in event processing and mobile target tracking, are to be worked on.

Acknowledgements

The authors thank the anonymous reviewers for their valuable comments that improve the quality of this paper and the financial support of Sino-France Advanced Research Program (No:PRA SI03-02), NSFC(No:60434030) and Zhejiang Academician Fund under project "Key Technologies and Development Platforms for Wireless Sensor Networks" (Academician Youxian Sun).

References

1. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Networks: a Survey. Computer Networks. **38** (2002) 393–422
2. Lin, R.Z., Wang, Z., Sun, Y.X.: Wireless Sensor Networks Solutions for Real Time Monitoring of Nuclear Power Plant. In: Proceedings of WCICA2004, IEEE (2004) 3663–3667
3. Kalidindi, R., Ray, L., Kannan, R., S., I.: Distributed Energy Aware MAC Layer Protocols For Wireless Sensor Networks. In: Proceedings of ICWN03. (2003) 1–5
4. Lin, R.Z., Wang, Z., Sun, Y.X.: Energy Efficient Medium Access Control Protocols for Wireless Sensor Networks and Its State-of-art. In: Proceedings of ISIE2004, IEEE (2004) 669–674
5. Shnayder, V., Hempstead, M., Chen, B., Allen, G., Welsh, M.: Simulating the Power Consumption of Large-scale Sensor Network Applications. In: Proceedings of SenSys2004, ACM (2004) 188–200

Connectivity and RSSI Based Localization Scheme for Wireless Sensor Networks

Xingfa Shen[1], Zhi Wang[1], Peng Jiang[2], Ruizhong Lin[1], and Youxian Sun[1]

[1] National Laboratory of Industrial Control Technology,
College of Info Science and Engineering, Zhejiang Univ., Hangzhou 310027, P.R.China
{xfshen, wangzhi, rzlin, yxsun}@iipc.zju.edu.cn

[2] College of Automation, Hangzhou Dianzi Univ., Hangzhou 310018, P.R.China
pjiang@hziee.edu.cn

Abstract. A multitude of applications of wireless sensor networks require that the sensor nodes be location-aware. Range-based localization schemes are sometimes not feasible due to hardware cost and resource restriction of the sensor nodes. As cost-efficient solutions, range-free localization schemes are more attractive for large-scale networks. This paper presents Weighted Centriod (W-Centriod), a novel range-free localization scheme extended on the basis of Centroid scheme, which takes received signal strength indicator (RSSI) metric into account besides connectivity metric used in Centroid scheme. It's shown that our W-Centriod method outperforms Centriod scheme significantly in terms of both the average localization error and the uniformity of error distribution across different positions, which decrease by 49.3% and by 37.7%, respectively, under the best circumstance. Moreover, a two-phase localization approach consisting of a field data collection phase and an off-line parameter optimization phase is proposed for localization in wireless sensor networks.

1 Introduction

Wireless sensor networks have greatly extended our ability to monitor and control the physical world. Such networks have been proposed for various applications including search and rescue, disaster relief, target tracking, and smart environments. The inherent characteristics of these sensor networks make a node's location an important part of their state. For such networks, location is being used to identify the location at which sensor readings originate or events happen or used in geographic routing protocols and some location based services (such as sensing coverage and network connectivity).

To incorporate location awareness into sensor networks, many localization algorithms have been proposed. Existing localization schemes can be divided into two categories by whether or not they require absolute range or angle information for estimating location [1]. Localization accuracy requirements vary with the applications. For high accuracy sensor localization, range-based localization schemes are preferred. Range accuracy on the order of centimeters has been reported [2]. However, due to some considerations such as size, cost and

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 578–587, 2005.

energy constraints of the nodes, range-based schemes are not practical for most applications in the large-scale sensor networks. Solutions in range-free schemes are being pursued as cost-effective alternatives to the range-based approaches.

This paper explores the possibility of integrating RSSI, a fine-grained metric, with range-free schemes to improve their localization accuracy. Based on this idea, we extend Centroid algorithm, a classic range-free localization scheme [3], which only uses connectivity metric.

The remainder of this paper is organized as follows: Section 2 discusses related work in localization for wireless sensor networks. Section 3 presents the radio propagation model used in this paper and then introduces the W-Centroid scheme. Section 4 describes our experiments and presents the experiment results with detailed performance comparison. In Section 5, some general problems that arise in deploying our localization method are further discussed. Finally, we conclude in Section 6.

2 Related Work

2.1 Range-Based Localization Schemes

Range-based localization schemes are commonly used as a method to obtain the fine-grained location information of nodes. Generally speaking, the range-based location discovery approaches consist of two basic phases: (1) distance (or angle) estimation and (2) distance (or angle) combining. By different distance (or angle) estimation methods they use, they can be further divided to four subsections: Time of Arrival (TOA), Time Difference of Arrival (TDOA), Angle of Arrival(AOA) and RSSI.

TOA and TDOA methods both compute signal propagation time to obtain range information. The difference between them is that the former records the time of arrival and transmission of the same signal while the latter uses two signals with different propagation speeds such as RF and acoustic signals. Cricket [4] is a famous infrastructure-based indoor localization systems using TDOA ranging method. In infrastructure-free ad-hoc sensor networks, a localization system called AHLos has been proposed [2]. TOA and TDOA technologies pose challenges for localization systems, which use them, due to energy constraints and hardware limitations.

AOA technique estimates the angle at which signals are received and uses geometric relationships to calculate node positions [5]. AOA has similar problems in hardware constraints and energy consumptions as TOA and TDOA. One additional problem is its poor scalability in large-scale sensor networks.

RSSI technique measures the power of the signal at the receiver. Based on the known transmit power, the effective propagation loss can be calculated. Either theoretical or empirical model is used to translate this loss into a distance estimate. This method has been used mainly for RF signals. RADAR system [6] is a typical in-door localization system based on RSSI technology. The main advantage of RSSI method is its low communication overhead and no additional hardware requirement. As shown in an empirical study of RF technology, however, problems such as multi-path fading, background interference and irregular signal propagation characteristics make range estimates inaccurate. Furthermore, RSSI method needs considerable efforts to obtain an empirical radio propagation model beforehand.

2.2 Range-Free Localization Schemes

As an alternate category of range-based localization solutions, range-free localization methods have been proposed to obtain coarse-grained location estimates in an acceptable limit. In this category, there are three major schemes: Centroid [3], DV-HOP [7] and Approximate Point-In-Triangulation (APIT) [1].

A very simple, range-free, connectivity-based and coarse-grained method, referred as Centroid algorithm, is proposed and evaluated for localization in outdoor environments that makes use of the inherent RF communications capabilities of sensor nodes. A fixed number of powerful reference points (or anchors) in the network with overlapping regions of coverage transmit periodic beacon signals containing their position information (X_i, Y_i). All neighboring nodes to be located keep an account of all received beacons. After receiving these beacons, a node estimates its location as the centroid of anchors for which the respective connectivity metrics exceed a certain threshold (a tunable parameter) using the following formula:

$$(X_{est}, Y_{est}) = \left(\frac{X_1 + ... + X_N}{N}, \frac{X_1 + ... + X_N}{N} \right) \tag{1}$$

In DV-HOP algorithm, for a node, the distances to anchors can be described by the minimum number of network hops between node and anchors and per-hop distance obtained through anchor communication. Then, given distances to and absolute positions of at least three different anchors, all nodes can calculate their own absolute positions via triangulation.Localization error will grow if node degree is decreasing when using the DV-hop technique.

APIT localization algorithm is an area-based, range-free localization method. This algorithm computes locations for nodes by determining the smallest possible area in which the node may be located. Only the most basic properties of radio wave propagation are assumed–the monotonicity of the reduction in radio signal strength along a single direction. The main disadvantage of APIT is that the algorithm's precision depends heavily on the node density and radio range. It's shown that the APIT scheme performs best when an irregular radio pattern and random node placement are considered.

3 W-Centroid Localization Scheme

3.1 Radio Propagation Models

There are three basic propagation models: the free space model, two-ray ground reflection model and the shadowing model. The free space model and the two-ray model predict the received power as a deterministic function of distance.

They both represent the communication range as an ideal circle. In reality, the received power at certain distance is a random variable due to multipath fading effects. The shadowing model is more general and widely-used . The shadowing model extends the ideal circle model to a richer statistic model: nodes can only probabilistically communicate when near the edge of the communication range.

The shadowing model consists of two parts. The first one is known as path loss model, which also predicts the mean received power at distance d, denoted by $\overline{P_r(d)}$. It uses a close-in distance d_0 as a reference. $\overline{P_r(d)}$ is computed relative to $P_r(d_0)$ as follows.

$$\frac{P_r(d_0)}{\overline{P_r(d)}} = \left(\frac{d}{d_0}\right)^{\alpha} \tag{2}$$

α is called the path loss exponent, and is usually empirically determined by field measurement. Table 1 gives some typical values of α. Larger values correspond to more obstructions and hence faster decrease in average received power as distance becomes larger.

Table 1. Some typical values of path loss exponent α

Environment		α
Outdoor	Free space	2
	Shadowed urban area	2.7 to 5
In building	Line-of-sight	1.6 to 1.8
	Obstructed	4 to 6

The second part of the shadowing model reflects the variation of the received power at certain distance. It is a log-normal random variable, that is, it is of zero-mean Gaussian distribution if measured in dB. In our proposed algorithm, a series of successive beacon signals are sampled by an receiver node in order to decrease the uncertainty of the RSSI measurements. For simplification, the second part of the shadowing model is ignored in deducing the W-Centroid algorithm.

3.2 W-Centroid Scheme

As shown in section 2.2, Centroid algorithm is a coarse-grained algorithm, whose localization accuracy is low and dependent on the distance between the adjacent anchors and the transmission range of these anchors. In order to increase the localization accuracy and decrease the influence of anchors distribution, RSSI metric can be used to reinforce the results of connectivity-based Centroid localization scheme.

In the network, nodes with overlapping regions of coverage serve as anchors R_i with known coordinates (X_i, Y_i). These anchors form a regular mesh and transmit periodic beacon signals with period T. It's assumed that the underlying MAC protocol ensures each anchor transmits one packet without any collision in any period T. Then, we define a few terms:

S : Sample size for connectivity metric for anchor R_i
t: Receiver sampling or data collection time
$N_{sent}(i,t)$: Number of beacons sent by R_i in time t
$N_{recv}(i,t)$: Number of beacons sent by R_i received in time t
CM_i : Connectivity metric for R_i
P_i: Average RSSI metric for R_i, in mW
$RSSI_j$: RSSI metric of the jth packet sent by R_i received, in $-dBm$
β : Threshold for CM
(X_{est}, Y_{est}): Estimated location of the receiver
(X_a, Y_a) : Actual location of the receiver

In order to improve the reliability of connectivity metric and RSSI metric in the presence of various radio propagation vagaries, each receiver listens for a fixed time period t and collects all the beacon signals it receives from the respective anchor. Then two metrics are computed for each anchor R_i respectively.

$$CM_i = \frac{N_{recv}(i,t)}{N_{sent}(i,t)} \times 100\% \tag{3}$$

$$P_i = 10^{\left(\frac{\sum_{j=1}^{N_{recv}(i,t)} RSSI_j}{N_{recv}(i,t)} \times \frac{-1}{10} \right)} \tag{4}$$

A straightforward idea to extend Centriod algorithm is to estimate a receiver's location as the weighted centroid of all anchors 'in-range', where the weight for R_i is a function of P_i. But what's the form of this weight function? A simple case is considered to answer this question, where the receiver A and two 'in range' anchors, B and C, are collinear, as shown in Figure 1.

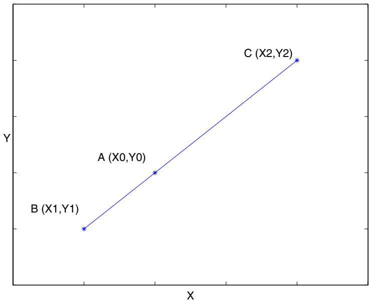


Fig. 1. A simple collinear case

Assume, at receiver A, the average strength of received signal from anchor B is P_1 and that from anchor C is P_2. Using Equation 2, the following two equations are got.

$$\frac{P_r(d_0)}{P_r(d_1)} = \left(\frac{d_1}{d_0}\right)^{\alpha} \qquad \frac{P_r(d_0)}{P_r(d_2)} = \left(\frac{d_2}{d_0}\right)^{\alpha} \tag{5}$$

From Figure 1, a geometric relationship can be easily obtained.

$$\frac{X_0 - X_1}{X_2 - X_0} = \frac{d_1}{d_2} \qquad \frac{Y_0 - Y_1}{Y_2 - Y_0} = \frac{d_1}{d_2} \tag{6}$$

Then, from Equation 5 and 6, the location of receiver A can be computed as follow.

$$X_0 = \frac{\sqrt[\alpha]{P_1}X_1 + \sqrt[\alpha]{P_2}X_2}{\sqrt[\alpha]{P_1} + \sqrt[\alpha]{P_2}} \qquad Y_0 = \frac{\sqrt[\alpha]{P_1}Y_1 + \sqrt[\alpha]{P_2}Y_2}{\sqrt[\alpha]{P_1} + \sqrt[\alpha]{P_2}} \tag{7}$$

Inspired by this simple case study mentioned above, the Centriod algorithm can be extended to W-Centroid scheme to estimate the location of receiver.

$$(X_{est}, Y_{est}) = \left(\frac{\sqrt[\alpha]{P_1}X_1 + \ldots + \sqrt[\alpha]{P_N}X_N}{\sqrt[\alpha]{P_1} + \ldots + \sqrt[\alpha]{P_N}}, \frac{\sqrt[\alpha]{P_1}Y_1 + \ldots + \sqrt[\alpha]{P_N}Y_N}{\sqrt[\alpha]{P_1} + \ldots + \sqrt[\alpha]{P_N}} \right)$$
$$s.t. \quad CM_i \geq \beta \quad (i = 1 \ldots N) \tag{8}$$

where P_i is the average RSSI metric for R_i (in mW), whose connectivity metric exceeds a certain threshold β. The accuracy of the estimate is characterized by the localization error LE, defined as

$$LE = \sqrt{(X_{est} - X_a)^2 + (Y_{est} - Y_a)^2} \tag{9}$$

4 Experimentation

4.1 Experiment Description

This paper provides a two-phase localization approach including realistic data collecting phase and off-line localization computing and parameter optimization phase.

In the first phase, field measures were performed about both connectivity metric and RSSI metric in outdoor environment. Our initial experiments are performed on the ground of the rooftop of our laboratory building. Our experimental testbed consisted of six mica2 motes. Four motes were placed at the four corners of a 6 m × 6 m square as anchor nodes. This square was further subdivided into 36 1 m × 1 m grids, and data were collected using a mote as receiver at each of the 49 grid intersection points. At each grid point, four anchors sent S=50 packets containing its own ID successively with an interval T= 0.2s, respectively. The receiver sampled for a time period of t=15s and then sent a result packet containing $N_{recv}(i, t)$ and P_i information to a base station mote connected with a laptop for off-line localization computing and parameter optimization.

In the second phase, based on the realistic data collected in the former phase, localization computing was performed using Centroid scheme and W-Centroid scheme with different parameter pairs. Then off-line parameter optimization was executed to choose a best parameter pair, α and β, to minimize the localization error, which can be used in future localization systems. In this phase, the software we used was Matlab.

4.2 Evaluation

This subsection provides a detailed quantitative analysis comparing the performance of the Centroid scheme and W-Centroid scheme with different parameter pairs. We use the *localizationerror* metric defined previously to characterize the performance.

Figure 2, 3 and 4 show the effect of varying path loss exponents on performance of Centroid and W-Centroid scheme. Similar conclusions can be drawn from these figures that W-Centroid scheme with different parameter pairs always outperforms than Centroid scheme, in terms of average localization error, maximum localization error and the standard variance of localization errors across 49 grid points.

From Figure 2, among different α values, α=2.5 performs best with least average localization error, which is about 0.8m, near 13% of the space between adjacent anchors (6 m in this paper). In contrast, the average localization errors of Centroid scheme are consistently higher than 1.5 m when varying the connectivity thresholds between 10% to 90%. This optimum value of α is slightly higher than the path loss exponent of free space shown in Table 3.1, which can be explained by that the motes are deployed on the ground in the experiments, where the radio power decreases much faster.

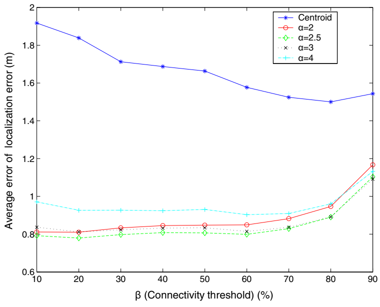


Fig. 2. Average localization error with different path loss exponents under varying connectivity thresholds

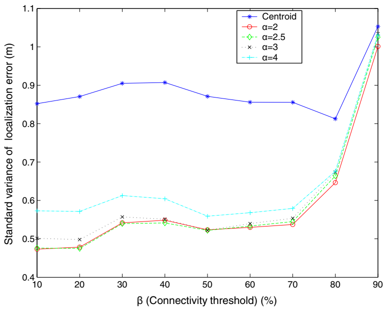


Fig. 3. Standard variance of localization errors with different path loss exponents under varying connectivity thresholds

Figure 3 presents the standard variance of localization errors across 49 grid points. From this figure, Centroid scheme has significantly higher standard variance than all the W-Centroid schemes with different parameter pairs. This is to say, the localization error distribution of Centroid scheme is suffering from greatest unevenness, which may lead to severe decreasing of application performance. In Fig. 4, we can draw a similar conclusion with Fig. 3. When β is varying from 10% to 80%, Centroid scheme keep a high max error of localization error about 3.6 m, which is higher than a half of the space between adjacent anchors.

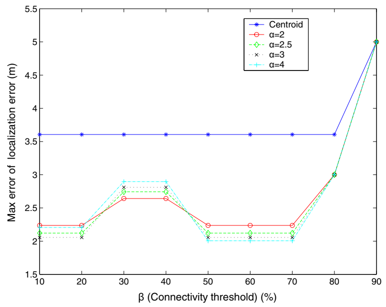


Fig. 4. Maximum localization error with different path loss exponents under varying connectivity thresholds

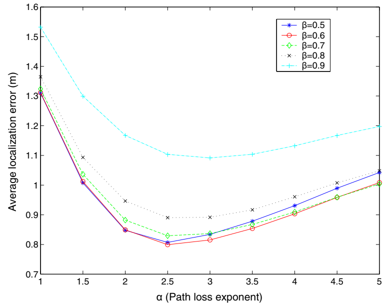


Fig. 5. Average localization error with different connectivity thresholds under varying path loss exponents

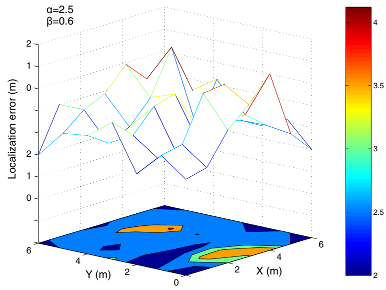


Fig. 6. The distribution of localization errors and its contour across all grid points when path loss exponent is 2.5 and connectivity threshold is 60%

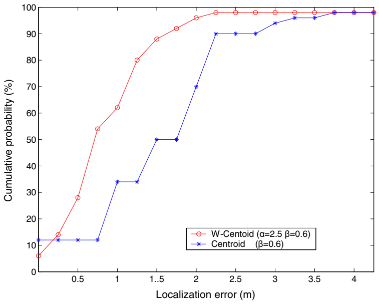


Fig. 7. The cumulative probability of localization error distribution when path loss exponent is 2.5 and connectivity threshold is 60%

Figure 5 shows the influence of different connectivity metric thresholds β on the average localization error. From this figure, β=60% performs best across all βs from 50% to 90%. βs less than 50% aren't taken into account because that low-quality links are much easily affected by radio vagaries and environmental noise and thus are not a reliable metric for using in localization scheme. From Figure 2 to 5, an optimum parameter pair for our experiments can be obtained: α=2.5 and β=60%.

Using this optimum parameter pair, the localization errors from the experiments are plotted as a function of the actual position with a contour of localization error, as shown in Figure 6. Under this optimum parameter setting, the average localization error and standard variance are 0.799 m and 0.533 m, respectively. These two figures decrease 49.3% and 37.7% compared to 1.577 m and 0.856 m for Centroid scheme (β=0.6), which is a significant performance improvement. The distribution of localization error across all grid nodes is irregular due to the uncertainty of radio propagation and the complexity of realistic environment.

Figure 7 shows the cumulative localization error distribution across all the grid points considering the Centroid scheme and W-Centroid scheme with α=2.5 and β=60%. In the result of the former one, for over 90% of the data points the localization error falls within 2.2 m while this figure decreases to about 1.7 m for the latter one.

5 Discussion

In this section, some general problems arising in deploying the proposed localization scheme and the practicable solutions to solve them are presented.

Anchors Deployment. In this paper, it's assumed that anchors are deployed in a regular mesh structure. In practice, however, the anchors are deployed randomly in an ad-hoc way across the network. In future work, the performance in realistic environment must be compared between different deployment styles of anchors: uniform deployment or random deployment.

Obviously, if anchors are deployed more densely, the quality of the proposed localization scheme will increase. But increasing the density of anchors requires additional cost of hardware and maintenance. As a result, it must be investigated that how to make a trade-off between performance improvement and cost increase resulting from anchor density increasing in future work.

Collision Avoidance. For our method to work well, neighboring anchors must be coordinated to avoid transmitting beacon signals at the same time. To achieve this, some collision avoidance schemes such as TDMA or randomized transmission can be adopted in our localization scheme.

Transmission Power Adjustment. Adaptive localization accuracy can be achieved by adjusting the transmission power of anchors. Given a constant anchor density, increasing the transmission power can increase the granularity of the localization regions and hence improve the accuracy of location estimates. Increasing the transmission power, however, will consume more energy. Therefore, an interesting idea is adjusting the transmission power to achieve different localization accuracies adaptive to the application requirements.

Robustness. The robustness to anchor failures of our scheme must be explored in our future work. A certain degree of anchor redundancy will be provided to tolerate anchor failures. If one anchor fails, another backup anchor will join in to maintain the quality of localization.

6 Conclusion

This article proposes a RF-based, receiver-based and rang-free localization scheme called W-Centroid, which is a simple, practical and cost-efficient localization solution for large-scale but resource-constrained sensor networks. As shown in the results of our outdoor experiments based on motes, taking RSSI metric into account besides connectivity metric used in Centroid scheme improves the accuracy of localization significantly without any additional hardware cost and neglectable energy consumption.

Moreover, a two-phase localization approach is proposed for localization in wireless sensor networks. First phase is field data collecting phase, in which the value of connectivity and RSSI metrics are collected. Second phase is offline parameter optimization phase based on analysis tools in order to identify a preferable parameter pair for future applications. In the future, we will extend our solution to handle more sophisticated network configuration and fulfill diverse application requirements.

Acknowledgements

This work is supported in part by NSFC under contract number 60434030 and Sino-France Advanced Research Program under contract number PRA SI03-02) and Zhejiang Academician Fund under project " Key technologies and development platforms for wireless sensor networks"(Academician Youxian Sun).

References

1. He, T., Huang, C., Blum, B.M., Stankovic, J.A., Abdelzaher., T.: Range-free Localization Schemes for Large Scale Sensor Networks. Technical Report CS-2003-06, University of Virginia, Computer Science Department (2003)
2. Savvides, A., Han, C.C., Srivastava, M.B.: Dynamic Fine-Grained Localization in Ad-Hoc Networks of Sensors. In: Proceedings of MOBICOM '01. (2001)
3. Bulusu, N., Heidemann, J., Estrin, D.: Gps-less Low cost Outdoor Localization for Very Small Devices. IEEE Personal Communications Magazine. **7** (2000) 28–34
4. Priyantha, N.B., Chakraborty, A., Balakrishnan, H.: The Cricket Location-Support System. In: Proceedings of MOBICOM '00. (2000)
5. Niculescu, D., Nath, B.: Ad Hoc Positioning System (APS) using AoA. In: Proceedings of INFOCOM' 03. (2003)
6. Bahl, P., Padmanabhan, V.: RADAR: An In-Building RF-based User Location and Tracking System. In: Proceedings of INFOCOM 2000. (2000)
7. Niculescu, D., Nath, B.: DV Based Positioning in Ad hoc Networks. Journal of Telecommunication Systems. (2003)

A Self-adaptive Energy-Aware Data Gathering Mechanism for Wireless Sensor Networks[*]

Li-Min Sun, Ting-Xin Yan, Yan-Zhong Bi, and Hong-Song Zhu

Institute of Software, Chinese Academy of Science,
P.O.Box 8718, Beijing, P.R.China 100080
{sunlimin, tingxin03, yanzhong02, hongsong}@ios.cn

Abstract. Sensor networks are composed of a large number of densely deployed sensors. The sensor nodes are self-organized and form an ad hoc network. As the energy supply to sensor nodes is limited and cannot be replenished, energy efficiency is an important design consideration for sensor networks. We design and evaluate an energy efficient routing algorithm for data querying sensor networks that propagates routing instructions and build data paths by considering both the hop count to the sink node and the minimum residual energy of that path. The proposed Dynamic Energy Aware Routing (DEAR) algorithm can effectively choose a data path with low energy consumption and high residual energy. The simulation results show that DEAR can prolong the lifetime of networks compared with Directed Diffusion, Minimum Transmission Energy routing and Energy Aware Routing.

1 Introduction

A large number of sensor nodes are densely deployed nearly or in the sensing area to collect and transmit objective information in sensor networks. Sensor network is especially adapted to the inhospitable physical environments such as remote geographic regions or toxic urban locations. It will also enable low maintenance sensing in more benign, but less accessible environments: large industrial plants, aircraft interiors etc. Each node in the sensor network may consist of one or more sensors, a microprocessor, a low power radio and portable power supply. A few nodes may also contain localization hardware, such as a GPS (Global Positioning System) unit or a ranging device. Sensor nodes usually have nonreplenishable energy resource, which causes energy efficiency an important design consideration for sensor networks. Besides, the network lifetime is related to the nodes who first drain there energy, so how to make the energy consumption evenly in the network is another important issue in energy aware routing research.

As wireless communication is the major factor in energy consumption, the performance of communication protocol can strongly influence the energy consumption efficiency. The coding and modulation mechanism of wireless channel and the MAC protocol can determine the energy consumption efficiency of a single

[*] This work is supported by National Natural Science foundation of China under grants No.60434030 and No.60373049.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 588 – 597, 2005.

node, and the routing protocol is a critical factor in determining the energy consumption efficiency of the whole network. The sensor nodes usually knows limited local information of sensor network due to limited resources, so the routing protocol in sensor networks have to make routing decision only by the information from neighbor nodes. The routing protocols should consider not only the optimization of total energy consumption but also the equilibration of energy consumption among sensor nodes. If only considering the former factor, the network may be separated into several parts because the nodes in the optimized path may drain their energy much more quickly than other nodes far from the path.

Sensor networks are data centric networks and always have tight correlation to application scenarios. We can regard a sensor network as a distributed real-time database. The querying data are separated and stored in the sensor nodes and acquired through sending instructions to the areas where interested data are stored.

In this paper, we propose Dynamic Energy Aware Routing (DEAR) for the data querying sensor networks. The aim of DEAR is to both reduce the total energy consumption of data delivery and make the energy consumption evenly among the nodes in the sensor networks so that DEAR can greatly increase the lifetime of sensor networks. In DEAR, the next hop is decided by both the hop count to sink node and residual energy of the data path. As the data being transmitted from destination to sink node, the data path will be adjusted dynamically according to the residual energy of each node along it.

The rest of the paper is organized as follows. We briefly discuss related work in section 2. We discuss DEAR in detail in section 3, and present our simulation results in section 4. In section 5 we draw the conclusions.

2 Related Works

Our work for the data querying sensor networks is closely related to the energy aware routing, data querying routing and geographic routing.

Rather than using traditional metrics such as hop-count or delay for finding routers, energy aware routing use the amount of energy consumption or residual energy as the routing criteria. This kind of routing protocol is suitable for the sensor networks because sensor nodes are restricted in energy supply. In MTE (Minimum Transmission Energy) [1], the energy consumption is in proportion to the square of communication distance, the routing mechanism is designed to find out the path with least value of total energy consumption. MTE only accounts the energy consumption for transmission while ignore the receiving consumption. Strictly speaking, MTE isn't the least energy consumption routing. In Maximum Minimum Power available routing [2], the available energy in a path is the node's residual energy which is the least among all the sensor nodes. This routing mechanism aims to choose the path from sink node to the destination with the maximum available energy. Both MTE and Maximum Minimum Power available routing needs global information of sensor network, so they are ideal routing mechanisms. Energy Aware Routing (EAR) proposed by Rahul C. Shah al [5] broadcasts query messages in the network to builds multi-path from the destination to sink node. Each path has a probability to be chosen and the probability is determined by both the energy consumption and residual energy. This mechanism makes the energy consumption of each node evenly thus prolong the lifetime of sensor network. EAR needs periodically flooding of query messages to maintain data paths, thus increase the protocol cost.

Directed Diffusion (DD) [4] is a query-based routing mechanism. Sink node broadcasts the query instructions all over the network. The query instruction contains several parameters such as query type, data rate, time stamp, etc. Every sensor node only needs to know its neighbors and they build gradients point to them. As the instructions spreading, a data path from destination to the sink node is built. Directed Diffusion will choose the path with highest data rate. Directed Diffusion is a data-centric protocol for sensor network applications. It achieves some level of energy savings by selecting empirically good paths, and by caching and processing data in-network. However, low rate data flooding throughout the network wastes a considerable amount of energy. Rumor routing [6] is another kind of query-based routing. In rumor routing, agents from sink node and destination will be delivered randomly and form a path once the two kind of agents encounter. The data path in rumor routing is not optimized and it cannot avoid routing loop.

It is supposed that each node knows its geographic information in Geographic and Energy Aware Routing (GEAR) [8]. GEAR routing uses a greedy mechanism which uses the distance to estimate the energy cost for forwarding instructions from sink node to the destination area, so it is a minimum energy consumption routing. GEAR avoids broadcasting instruction so as to reduce the protocol cost. GEAR only uses local information, so it may fall in the routing void. Greedy Perimeter Stateless Routing (GPSR) [7] is another geographic routing used for Data Central Storage in sensor networks. In GPSR, it also uses greedy mechanism to forward instructions which essentially tries to find the minimum hops path to the destination region by location information. It also defines a set of perimeter nodes and proposes a Right-Hand rule to bypass routing void.

The DEAR routing we proposed in this paper is used for data querying sensor networks. The query message is sent to the destination region according to nodes' position information. At the same time, the data path is decided by the hop counts to sink node and nodes' available energy, and it is updated dynamically with nodes' changing during the lifetime of sensor networks.

3 Dynamic Energy Aware Routing (DEAR)

3.1 Network Assumptions

A data-querying sensor network consists of a single sink node and a lots of sensor nodes. Sink node receives the exterior queries, generates query instructions and sends them to the destination. After receiving the query instructions, the destination transmits the matching results. Data flows are from the data source to the sink node while instructions are sent in the opposite direction. The task parameters such as destination position and task type are carried in query instructions. We discuss data-querying sensor network on the assumptions that

1. Each node knows its own location and remaining energy level, and its neighbors' locations and remaining energy levels through a simple neighbor hello protocol.

Note that a node can obtain its location information at low cost from GPS or some localization system.

2. The link is bi-directional, i.e., if a node hears from a neighbor, then its transmission range can reach the neighbor.
3. Sink node continuously have sufficient energy supply.

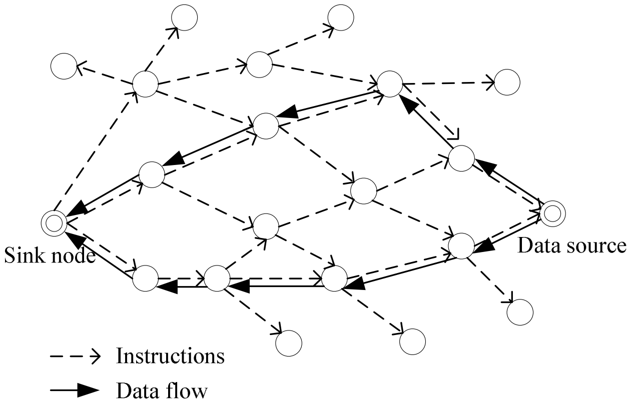


Fig. 1. A typical data-querying sensor network

A typical data-querying network is shown in Fig.1. The instructions and data are transmitted over the network in an opposite direction. During the process the query instructions being broadcasted, each node selects one or several nodes as its next hop, thus several data paths from destination to the sink node are built. In Directed Diffusion, query instructions are flooded over the whole network, and the data paths are built according to the data rate among sensor nodes [4]. In our mechanism, we consider both the length of path and the minimum energy to construct the data path. The instructions process is similar to restricted flooding.

3.2 Routing Decision Function

We define a function of path selection in our DEAR routing mechanism. There is a pair of parameters <hop count, residual energy> called as Routing Decision Parameters Pair (RDPP) in the function. We note RDPP as (H, R). Each node between sink node and destination is on a path to the sink for forwarding data. The hop count of node i is the number of nodes from it to sink node, and the residual energy of the data path is represented by the least residual energy of each node along the path. We use the quotient of residual energy and energy consumption of sending a fixed length message to evaluate the residual energy of each node.

There is a filed recording the RDPP value in query message. The initialization value of RDPP in the query message sent by sink node is RDPP is $(0, \infty)$. We call the following function about node i as the routing decision function.

$$F_i(H_i, R_i) = H_i / R_i \,. \tag{1}$$

When the node j near sink node receives the initialization query massage, it updates the RDPP as (1, R_j) in the massage and forwards the massage. The R_j value is the available energy of node j. When a node receives query massages from several neighbors, it uses the routing decision function (1) to compute the corresponding value for each neighbor, and select the neighbor with minimum value as next hops to sink node.

3.3 Path Building Process

The DEAR consists of two main processes, one is path building, and the other is path revising. While the query massages are delivering to the destination, the data path to sink node is built hop by hop in the opposite direction. The entire path will be setup when the query massages arrive to the destination. Along with data transmitting, the data path will be adjusted dynamically in response to changing of nodes' energy and the network topology.

The path building process is initialized by sink node sending a query massage. The query massage sets RDPP as (0, ∞), and contains the position of destination and task type. When node i first receives an instruction message from one of its neighbors, it will record the RDPP value and other parameters from that neighbor node into a task list. Then node i will wait for a certain time to receive instructions from other neighbor nodes and it will record all of the RDPP values from neighbor nodes into the list.

Next, Node i checks all the neighbors with following conditions.

$$D(Ni, Nd) \le D(Nj, Nd)\ . \tag{2}$$

Here $D\ (N_i, N_d)$ denote the distance from node i to the destination. This will ensure that the next hop is not nearer to the destination. If there is no node satisfying the condition, node i will simply drop the instruction message, otherwise node i will choose its next hop by the following steps.

From all the neighbors satisfying the conditions above, Node i chooses node j which has the least value of routing decision function as its next hop node, i.e. the downstreaming node in the aspect of data path. If there are multiple nodes with the same least value of routing decision function, we choose the one whose instruction message came earlier to break the tie. If node i has only one neighbor, it has to choose the only neighbor as the next hop.

After choosing the next hop, say node j, node i will set its hop count H_i=H_j+1 and set the new remaining energy of the data path as $R_i = \min \{R_j, ER_i\}$. Here ER_i denotes the residual energy of node i itself, so R_i is the minimum residual energy in the path up to node i. Then node i uses (H_i, R_i) to substitute the former RDPP in the message and broadcasts it to all its neighbors.

In the routing decision function, we consider both the factor of hop count and residual energy. At the early period after deploying sensor networks, the routing decision function is mainly determined by the number of hops because sensor nodes have sufficient energy. As time goes by, the remaining energy of sensor nodes reduce and the residual energy take greater part in the routing decision function. If H is a constant, then DEAR is equivalent to the Maximum Minimum Power available routing. If R is a constant, then DEAR is equivalent to the shortest path routing.

DEAR colligates both hop count and residual energy. If considering the correlation between the number of hops and node's location, the residual energy takes greater proportion than the hop count in the routing decision function.

3.4 Path Revising Process

The destination node collects data and sends matching data out to its next hop according to the routing decision function mentioned. The node which receives data will forward it to its next hop too. So data from destination will be forwarded to the sink node along the data path.

The remaining energy will reduce gradually. The nodes in the data path may be no longer the best node according to the routing decision function. In order to utilize the energy of each node evenly and adapt the topology changes, we need to adjust the data path dynamically.

When the residual energy changes over a certain degree, the node will send out an "announcing message" to all of its neighbors. In order to avoid adjusting too frequently, we define a global threshold value. If the energy variety exceeds the threshold, or if the residual energy has been less than the residual energy value in the RDPP, the node will send out the announcing message.

When one node receives an announcing message, it will process it just like receiving an instruction message. The process in the view of node i is as following:

1. Verify whether the neighbor satisfies the location condition according to equation (2), if satisfying, go to the next step, otherwise drop the announcing message;
2. Revise the value of H and R in the task list according to announcing message. If the announcing message is not from the next hop, there will be no further process, otherwise go to next step;
3. Node i will recompute the routing decision function. If node i has the only one neighbor, it will send the new routing decision function value to the upstreaming nodes; if node i has more than one neighbors, node i will choose the one who offer the least routing decision function value as next hop. Then node i sends an announcing message to its upstreaming node.

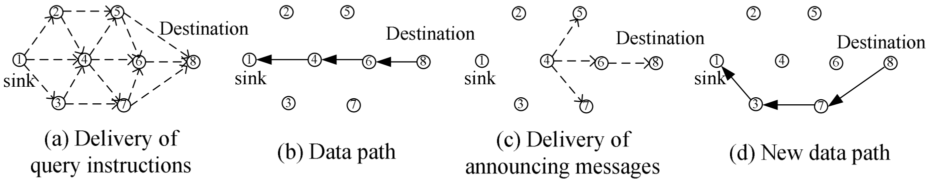

Destination
sink
①
②
③
④
⑤
⑥
⑦
⑧
(a) Delivery of query instructions
(b) Data path
(c) Delivery of announcing messages
(d) New data path

Fig. 2. illustration of path building and path revising process

The path building and revising processes are shown in Fig.2. Fig.2-(c) supposes the residual energy of node 4 decreases a lot, it sends out the announcing message to node 5, 6, and 7. This message may cause nodes 6 to generate an announcing message too because node 4 is the next hop of node 6. So this revising may cause node 8 to choose node 7 as its new next hop node to substitute node 6. The new data path is shown in Fig.2-(d).

If node i failed to communicate with its downstreaming node for a period of time, it will assert a link failure. In this situation, node i will delete the failed node from its list and recompute the routing decision function from the left nodes in the list. The hop count and residual energy would be changed, and node i will broadcast out an announcing message.

3.5 Further Discussion

1. Multi-path routing

In DEAR, there is only one data path from destination to the sink node. In fact, several potential paths are built in the path building process. If we sort the neighbors according to the value of routing decision function and assign the forwarding probability to each of these neighbors, then it will form multiple data path from destination to the sink node. This mechanism can make the energy consumption more evenly among sensor nodes in the network

2. Routing decision function

Considering the simplicity for computation, we use a division of number of hops and residual energy as our routing decision. In the sensor networks, it is a critical problem of how to design and evaluate the performance of a routing decision function. We may do further research on the routing decision function and introduce multiple routing decision functions in a system to achieve sufficient flexibility.

3. Node's location

In DEAR, we use nodes' locations as a precondition of routing mechanism. It can reduce the total amount of instructions and announcing messages delivered and avoid routing loop.

4 Simulation and Evaluation

We carry out the simulation in such a scene: network area is a rectangle and sensor nodes are randomly generated in the network area. In each simulation, the destination nodes were randomly selected from a small scope in the left part of the network field, and the sink node was picked up randomly near the right edge of the network. We compare our mechanism with Directed Diffusion (DD), Energy Aware Routing (EAR) and the conventional Minimum Transmitted Energy (MTE) algorithm.

In our simulation, the sink nodes generates messages to refresh the routes every 50 simulation ticks and each destination node send one data message every 10 simulation ticks. Every node has the same initial energy in our simulations. The initial power of the destination nodes and sink nodes is set to be more sufficient in order to dry other nodes up. The power cost for sending a unit data was three times higher than that for receiving. We assume there are no computing delay and no network losses in the network.

The algorithms are compared in three aspects, they are: the time of the first node runs out of energy, the network survivability and the average delay of sink gathering data. The results are given in following sections.

4.1 Time of the First Node Dies

The time of the first node dies reflects the lifetime of the network and the fairness of power consuming in many conditions. If the routing protocol transmits data through fixed paths, the nodes on the paths would be dried up much faster than other nodes. As a result, the network may be divided into several separated parts and they are hard to communicate with each other.

The scope of the network was fixed at 100m x 100m in the simulation. The number of the nodes in the network was increased from 50 to 200 in step of 50. Fig.3 gives out the simulation results of the four algorithms for the comparison of the time of the first node dies. The Y-axis stands for the simulation ticks. We can see DEAR performance best in the four test scenes. As the number of neighbors in communication range goes up along with the nodes density increases, more messages should be handled, which shorten the time of the first node dies.

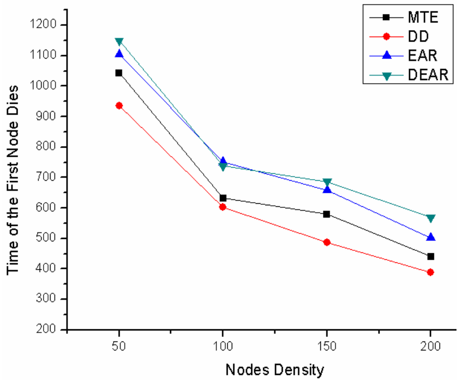


Fig. 3. The comparison of the time of the first nodes dies

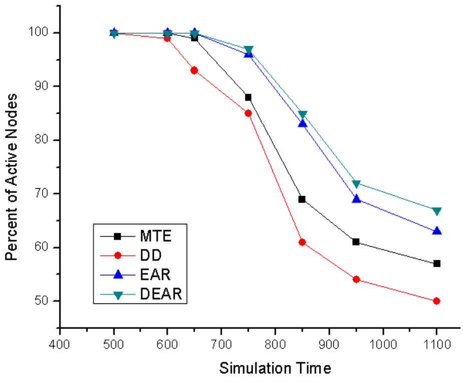


Fig. 4. Percents of active nodes vs. Simulation time

4.2 Network Survivability

Network survivability is another metric for comparing the lifetime of the networks. In this paper, network survivability is defined as the percent of the active nodes maintaining connectivity in the network. The simulation is executed in a network containing 100 nodes in a 100m x 100m scope. We set a large enough radio radius to ensure the network connectivity even more than half of the nodes died. The result of the simulation is shown in Fig.4. DD considers little about the energy, so it lost nodes earlier than other protocols. EAR can work better than both DD and MTE because it concentrates on consuming energy equitably. Since the contents of its route table can be refreshed during the paths building phases with a very low frequency, it cannot update the routes according to the power conditions in time. DEAR can adjust the routes dynamically according to the change of the residual power of the nodes, therefore it consumes energy more evenly than EAR does and can maintain longer lifetime of the network.

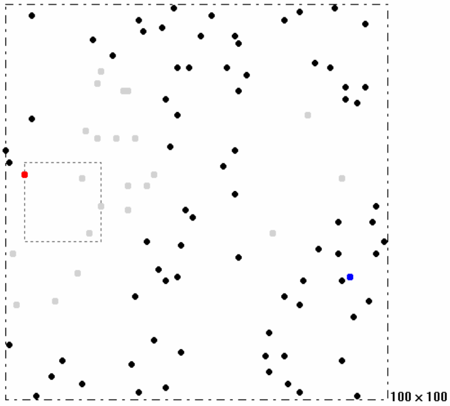


Fig. 5. A snapshot of the network after a simulation ended

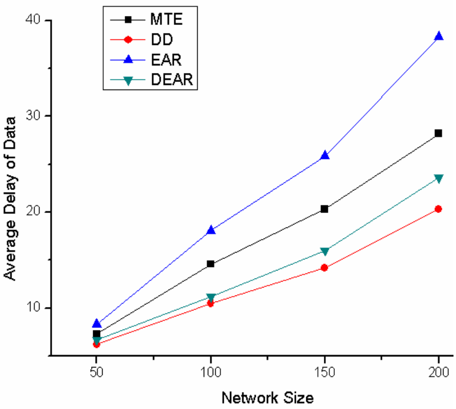


Fig. 6. Average delay of data gathering vs. network size

Fig.5 is the snapshot of a network after a DEAR simulation ended. The radio radius of each node is 35m in this simulation. The red node in the left dashed rectangle is the destination node and the blue node near the right edge of the network is the sink node. The grey nodes are dead nodes, and the black nodes are the survivals. Note that DEAR equitably dried up the energy of the nodes those around the destination node, instead of drying up the nodes only in the direction towards to the sink node. Due to having to handle more instructions and announcing messages, the nodes near the destination node commonly die earlier than those near the sink.

4.3 Average Delay of Gathering Data

It is well known that having short average delay of gathering data is an important feature of DD. Many other energy-efficient routing algorithms trade off the temporal performance to gain more efficiency of energy. DEAR gives attention to both aspects. Fig.6 shows the comparison of their temporal performance in different network size. In order to consume power evenly, DEAR may not always choose the paths that deliver data quickly. So it is hard for DEAR to gain less average delay than DD when the network runs a long time. But DEAR works much better than the other two energy-efficient algorithms obviously.

5 Conclusion

In this paper, we presented a new energy aware routing protocol named DEAR. DEAR is suitable for the data querying sensor networks and it considers both the aspects of energy consumption and delivering delay. In DEAR, the sink node spread routing instruction in a restricted flooding manner, the instruction contains hop count and minimum residual energy information. We define routing decision function as the quotient of hop count and residual energy and use it as the routing criteria. Simulation results show that DEAR can acquire better performance than DD, EAR and MTE in the standard of network lifetime.

References

1. W.R.Heinzelman, A.Chandrakasan, and H.Balakrishnan, Energy-Efficient Communication Protocol for Wireless Microsensor Networks. IEEE International Conference on System Sciences, January 4-7, 2000, Maui, Hawaii
2. J.-H. Chang and L. Tassiulas, "Maximum Lifetime Routing in Wireless Sensor Networks," Proc. Advanced Telecommunications and Information Distribution Research Program (ATIRP'2000), College Park, MD, Mar. 2000
3. J.-H. Chang and L. Tassiulas, "Energy Conserving Routing in Wireless Ad-hoc Networks ," Proc. IEEE INFOCOM 2000, pp.22-31, Tel Aviv, Israel, Mar. 2000
4. Chalermek Intanagonwiwat, Ramesh Govindan, Deborah Estrin, John S. Heidemann, Fabio Silva: Directed diffusion for wireless sensor networking. IEEE/ACM Transactions on Networkin. 11(1): 2-16 (2003)
5. Rahul C. Shah, Jan Rabaey, Energy Aware Routing for Low Energy Ad Hoc Sensor Networks IEEE Wireless Communications and Networking Conference (WCNC), March 17-21, 2002, Orlando, FL
6. D. Braginsky and D. Estrin. Rumor Routing Algorithm For Sensor Networks. The First Workshop on Sensor Networks and Applications (WSNA'02), October 2002, Atlanta, GA.
7. B. Karp and H.T. Kung. GPSR: Greedy Perimeter stateless Routing for Wireless Networks. MobiCom2000, Boston, Massachusetts. Aug 6-11, 2000
8. Yan Yu, Ramesh Govindan and Deborah Estrin. Geographical and Energy Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks. UCLA Computer Science Department Technical Report UCLA/CSD-TR-01-0023, May 2001
9. Ya Xu, John S. Heidemann, Deborah Estrin: Geography-informed energy conservation for Ad Hoc routing. MOBICOM 2001: 70-84
10. J. Newsome and D. Song. GEM: Graph EMbedding for Routing and Data-Centric Storage in Sensor Networks Without Geographic Information. The First ACM Conference on Embedded Networked Sensor Systems (Sensys03), Los Angeles, CA, USA. November, 2003
11. A. Rao, S. Ratnasamy, C. Papadimitriou, S. Shenker, and I. Stoica. Geographic Routing without Location Information. MobiCom'03. San Diego, California, USA. September 14-19, 2003
12. Scott Shenker, Sylvia Ratnasamy, Brad Karp, Ramesh Govindan, Deborah Estrin: Data-centric storage in sensornets. Computer Communication Review 33(1): 137-142 (2003)

An Adaptive Energy-Efficient and Low-Delay MAC Protocol for Wireless Sensor Networks

Seongcheol Kim

Software School, Sangmyung University,
Seoul, Korea
sckim@smu.ac.kr

Abstract. To increase the life of the sensor networks, each sensor node has to conserve energy. Due to the fact that each sensor node has one battery energy to remain alive for long times, energy management is a one of critical issues in wireless sensor networks. In this paper we propose a power efficient and low-delay MAC protocol (MT-MAC, Modified T-MAC) for wireless sensor networks. We first address the protocol design problem and suggest a novel solution based on media access control protocol. The MT-MAC uses network traffics to determine active and sleep periods to save energy and enhances the packet transmission latency. The MT-MAC shows better energy saving than the previous proposed protocols, S-MAC and T-MAC.

1 Introduction

Wireless sensor networks (WSNs) are an emerging research area. Many applications like environmental monitoring, military, home security, context-aware personal assistants, and medical monitoring are related with the WSNs.

WSNs usually contain thousands or millions of sensors, which are randomly and densely deployed. A sensor node is a small electronic or electromechanical device with processor and transceiver. Sensor nodes may have attached sensing devices like a light detector or magnetometer. A sensor node can communicate only within a limited range. Wireless sensor networks are characterized by the limited energy supply because wireless sensor nodes are typically powered by batteries. So the important design issue for medium access control protocols in wireless sensor networks is to save energy. It is important to realize that the failure of individual nodes may not harm the overall functioning of a sensor network, because neighboring nodes can take over provided that the node density is high enough. Therefore the main parameters to optimize for are network lifetime, the time until the network gets partitioned.

Energy efficient protocols have been proposed for Medium Access Control (MAC), topology control, and data aggregation. The main issue of these works is in the design of novel sleep scheduling schemes wherein nodes turn off their communication radios during the sleep. A MAC protocol decides when competing nodes may access the shared medium and tries to ensure that no two nodes are interfering with each other's transmission. The MAC layer operates on a local scale and lacks the global information to optimize for network lifetime [1], [2], [3], [4]. In contrast to typical WLAN protocols, MAC protocols designed for sensor networks usually trade

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 598 – 606, 2005.

off performance such as delay, throughput, and fairness for energy efficiency. For example, a low duty cycle MAC is energy efficient. But it increases the packet delivery delay. An intermediate node may have to wait until the receiver node wakes up before it can forward a packet.

Sensor nodes have special low power hardware and batteries may be difficult or impossible to recharge. So protocols for wireless sensor networks must focus on energy efficiency. The primary source of energy consumption in WSN is the radio. So minimizing the radio's activity is the major target to solve. Much energy is wasted due to the following sources of overhead [5], [6], [7], [8], [9], [10].

Collisions. When two nodes send packets at the same time, packets are corrupted. Since this packet is discarded, the energy consumption per successful transmission will increase.

Idle Listening. A sensor node does not know when it will receive packet from its neighbors. So the node should keep its radio in receive mode at all time. Typical radios consume two order of magnitude more energy in receive mode than in standby mode.

Overhearing. A node may receive a packet not destined to it. In this case, the node consumes energy to receive packet.

Control Packet Overhead. The MAC headers and control packets used for signaling do not contain application data and are therefore considered overhead.

Nodes are in one of the following states: sleeping, listening, sending, and receiving. Among these four states, sending and receiving states consume the highest amount of energy. But the listening state consumes nearly as much power as the sending and receiving states. Furthermore nodes typically spend most of their time in listening state. The example of power consumption of the Mica2 Mote sensors is shown in table 1. So the listening power consumption becomes the most important factor of the nodes lifetime in WSNs.

To extend the lifetime, it is required to transit from the listening state to the sleeping state, in which power consumption is much less than that of sending/receiving state. In many sensor network applications, nodes are in idle listening state for a long time if no sensing events happen. If the data rate during this period is very low, it is not necessary to keep nodes listening all the time. So each node goes to sleep, wakes up, and listens to see if any other node wants to talk to it.

An important goal of a MAC protocol is to increase the percentage of sleeping period of the radio. But in contrast to typical WLAN protocols, MAC protocols designed for WSNs usually trades off throughput or/and latency for energy efficiency. The goal of this paper is to define a MAC protocol for the WSNs with energy efficiency and low latency.

Table 1. Characteristics of a sensor radio [11]

Ratio State	Power Consumption (mW)
Transit	81
Receive	30
Idle Listening	30
Sleep	0.003

The rest of this paper is organized as follows: Section 2 introduces the related works and the details of the proposed protocol are presented in Section 3. Section 4 presents and discusses the simulation results, and the paper concludes with Section 5.

2 Related Works

The main goal in designing a MAC protocol for WSNs is to minimize energy consumption, while limiting latency and loss of data throughput. There are so many works done in past few years in the design of energy-efficient, low latency communication protocols. Current MAC design for wireless sensor networks can be classified into two categories: contention-based protocols and frame-based TDMA protocol. IEEE 802.11, PAMAS [2], S-MAC [12], and T-MAC [13] are examples of contention-based protocols.

In contention-based protocols nodes can start a transmission at any random time and must contend for the channel. So the main challenge with contention-based protocols is to reduce the energy consumption caused by collisions, overhearing, and idle listening. These works design scheduling schemes to control duty cycles of communication radios. The advantages of the contention-based protocols are the low implementation complexity, the ad hoc nature, and the flexibility to accommodate mobile nodes and traffic fluctuations.

In a frame-based TDMA protocol, time is divided into time slots, which nodes can use to transfer data without having to content for the medium or having to deal with energy wasting collisions of transmissions. After the frame length, which consists of several time slots, the node again has a period time reserved for it. The major advantage of frame-based TDMA protocols is the inherent energy-efficiency due to the lack of collisions, overhearing, and idle-listening overheads. The main challenge with frame-based TDMA protocols is to operate efficiently in ad-hoc sensor networks without any infrastructure.

Most MAC protocols for wireless sensor networks have been based on conventional wireless protocol, IEEE 802.11. The medium access control in the IEEE 802.11 MAC protocol is based on carrier sensing (CSMA) and collision detection. If a node wants to transmit a packet, it must first sense the radio channel to check whether it is free for a specified time called the Distributed Inter Frame Space (DIFS).

The IEEE 802.11 MAC provides low-level support for power management such as buffering data for sleeping nodes and synchronizing nodes to wake up for data delivery. In the IEEE 802.11 specification, a node can be in one of two power management modes, active mode or power-save mode. In active mode, a node is awake and may receive packet at any time. But in power-save mode, a node wakes up periodically to check for incoming packet.

The IEEE 802.11 power-save mode attempts to conserve energy on idle nodes by powering their wireless interfaces off for a specified period time. So each node in the IEEE 802.11 power-save mode may be in one of two states: sleeping and awake. In sleeping state the node is unable to communicate but is consuming very little power. But in a wake state the node is fully powered and capable of communication.

All nodes in the IEEE 802.11 power-save mode are assumed to be synchronized and awake at the beginning of each time interval. The main issue of these works is in the design of novel sleep scheduling schemes wherein nodes turn off their communication radios during the sleep. Next, we consider some previous contention-based protocols.

PAMAS. The PAMAS [2] power-saving MAC protocol turns off a node's radio when it overhears a packet not destined to it. The PAMAS protocol uses out-of-band signaling to reduce the overhearing, while preserving throughput and latency. Whenever a node overhears signaling destined for another node it calculates the time until the associated data transfer finishes. The radio is turned off, and will be turned on when the medium becomes available again for other transfers. The effectiveness of PAMAS is limited to reducing the power consumption of processing unnecessary packets.

STEM. STEM (Sparse Topology and Energy Management) [10] uses asynchronous beacon packets in a second control channel to wakeup destination nodes. STEM achieves low power consumption of wakeup radio by using a large duty cycle ratio, instead of assuming a low power wakeup radio. After transmissions have finished, the node turns its radio off in the data channel. STEM does not provide mechanisms for indicating the power management state of a node.

DMAC. DMAC [8] attempts to reduce the amount of latency experienced in a sensor network that is employing a power save mechanism. DMAC assumes that there is just one destination in the network to form a routing tree. DMAC adjusts the duty cycles adaptively according to the traffic load in the network. Data prediction is employed when each single source has low traffic rate but the aggregated rate at an intermediate node is larger than what the basic duty cycle can handle. DMAC also proposes using a staggered wake-up schedule to create a pipeline for data propagation.

S-MAC. S-MAC [12] is designed to save energy on single radio architecture. S-MAC protocol prevents overhearing by in-channel signaling, using RTS (Request To Send) and CTS (Close To Send) packets. And S-MAC proposes a virtual clustering to allow nodes to synchronize on a common frames structure. An S-MAC slot starts off with a small synchronization phase, followed by a fixed length active period, and ends with a sleep period. In a sleep period, nodes turn off their radio. But S-MAC uses a fixed sleep interval regardless of networks traffics. When a node is in active period, it can communicate with its neighbors. During the sleeping period, transmitted packets will be queued. Since all packets are transmitted during active period, the energy consumption on idle listening can be reduced.

T-MAC. The main idea of the T-MAC [13] is based on the S-MAC. In other words, T-MAC is an extended version of S-MAC. In T-MAC, a node has active and sleeping periods. But the two parts can be adjusted with network traffic. During an active period, if a node cannot receive any packet from its neighbor for a time TA, it will enter sleeping period immediately. The time called TA is a threshold value. The value TA determines the minimal amount of idle listening per frame. Since a node can enter into sleeping period quickly, the sleeping period will increase. So T-MAC can save energy mode than S-MAC. But in T-MAC, it doesn't consider when a node wake-up from sleeping period. It just considers when a node can enter into sleeping period. T-MAC protocol is more energy effective than S-MAC protocol by decreasing active period and increasing sleeping period. In other words, T-MAC focuses on when the transition occurs from active state to sleeping state. But they didn't consider when the sleeping period ends. So the main idea of the proposed protocol in this paper is to design wake-up mechanism according to the network traffic in addition to the T-MAC.

Figure 1 shows the basic scheme of S-MAC and T-MAC protocols. As shown in the figure, every node periodically wakes up to communicate with its neighbors. Since all messages are packed into the active part in S-MAC, the time between messages, and therefore the energy wasted on idle listening is reduced. But in T-MAC, the active time is not fixed, instead adaptive to the network traffic.

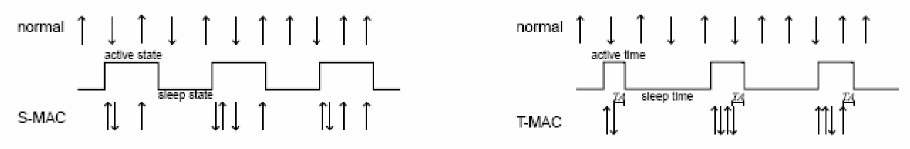


Fig. 1. The basic S-MAC and T-MAC protocol scheme. S-MAC has fixed length active periods, but T-MAC has adaptive active periods [13].

The new protocol proposed in [5] increases energy efficiency by allowing packet buffering. And the protocol uses two channels: primary and wakeup channel is used to wakeup neighbors. It is assumed the second radio is only wakeup. Furthermore they use only packet inter-arrival time to adjust the value of T. But the protocol has the following problem: each busy tone must wake up a node's entire neighborhood since the intended receiver's identifier is not encoded on the wakeup channel.

The primary channel is used for sending data and control packets, and the capable of transmitting a busy tone, rather than actual data. The goal of the proposed protocol is to find the optimal T value, for a given data rate minimizing the energy consumption. After selecting the optimal value of T, the sender will send T with its data packet by piggyback. Then receiver will schedule a triggered wakeup T time in the future, taking account transmission delay. So the protocol tries to sleep as soon as possible after data communication and predict when it should next wakeup based on previous traffic patterns.

We propose a modified T-MAC protocol (MT-MAC) in this paper. The proposed MT-MAC considers not only the packet inter-arrival time but also the active periods. The design goal of MT-MAC is to provide a traffic-adaptive for energy-efficient and low-delay MAC protocol in wireless sensor networks. The MT-MAC is closely related to the protocol in [5].

3 The Modified T-MAC (MT-MAC)

As mentioned previously, in S-MAC nodes periodically sleep to reduce energy consumption in listening to an idle channel. S-MAC also sets the radio to sleep during transmissions of other nodes. T-MAC adjusts the length of time sensors are awake between sleep periods based on communication of neighbors. T-MAC also transmits all messages in bursts of variable length to reduce idle listening and sleep between bursts. Doing this less energy is wasted due to idle listening when data traffic is light.

There are some previous researches considering saving energy by adjusting the network traffics [5], [14]. The main goal in [5] is to find the optimal value of T, for a given data rate, which minimizes the energy consumption. The sender estimates its sending rate through a weighted average of the inter-arrival time of packets. But the control scheme of [5] uses two radios and considers only one parameter to calculate the value of T. The primary channel is used for sending data and control packets, whereas the wakeup channel is used to wakeup neighbors. A queue threshold is used to control delay or limit the storage usage on a sensor. To avoid costly full wakeups, a sensor estimates the rate at which it is sending data and tries to schedule a triggered wakeup with a receiver T seconds after its previous data transmission.

In the proposed MT-MAC protocol an entire frame time T_{frame} will be changed according to the network traffic. Let T_{frame} be an entire frame time and $T_{threshold}$ be the time that a source node should be active. So we have following relationship.

$$T_{frame} - T_{threshold} = T_{sleep} \tag{1}$$

Sleeping period T_{sleep} can be obtained from above equation. The Modified T-MAC (MT-MAC) uses two parameters: $T_{threshold}$ and an average of the inter-arrival time of packets. Both parameters reflect traffic conditions as discussed before. Each node calculates the length of $T_{threshold}$ and the average of the inter-arrival time of packets. These two parameters are used in calculating wake-up time and the entire frame time T_{frame}. And in the MT-MAC protocol, nodes enter sleeping period from active period if they doesn't receive any packet from neighbor nodes for TA. This action is exactly the same as T-MAC. But wake-up time will be depended on the network traffic.

Large $T_{threshold}$ value means that there are many transmissions between neighbor nodes. So if nodes get large $T_{threshold}$, the node in sleeping period must wakeup early to communicate each other since there may be lots of packets to be transmitted. And small $T_{threshold}$ means that there are little transmissions between neighbor nodes. So nodes with small $T_{threshold}$ values in sleeping period can stay in sleeping mode long enough to save energy consumption. But when there are little transmissions but the triggered events will be happen just before TA value. In this case $T_{threshold}$ value will be large but network traffic is small.

Even though the average inter-arrival time of packets is short, it doesn't mean that total length of $T_{threshold}$ is small. As many packets arrive in sleeping period, then the length of $T_{threshold}$ will increase. So in the proposed protocol, the average inter-arrival time of packets and the length of $T_{threshold}$ will be considered to obtain adequate T_{frame}.

The sequence of steps followed by a node in MT-MAC is the following:

1) Estimates the packet inter-arrival time and $T_{threshold}$
2) Calculates appropriate T_{frame}
2) Sends data packets to neighbors with its chosen T_{frame} and $T_{threshold}$ values
3) Returns to sleep if no more packet is sent or received for TA
4) Returns to wakeup mode (T_{frame} - T_{sleep}) time later

The entire frame time T_{frame} and sleep time T_{sleep} are a function of two parameters as follows:

$$T_{frame} \sim f_n(\text{packet inter-arrival time, } T_{threshold}) \tag{2}$$

Internet traffic is usually considered as burst. So once a node gets congested, this congestion keeps going on some time. If a node gets larger $T_{threshold}$, the node may receive more packets from its neighbor nodes for some time.

The advantages of the proposed protocol are as follow. First, it is adaptive to network conditions. If there are many packet transmissions in networks, the protocol can quickly respond by adjusting entire frame time and sleeping time. This can deduce packet transmission latency. Second, MT-MAC uses single radio. Third, nodes can transit from active period to sleeping period and from sleeping period to active period quickly, which can reduce energy consumptions and packet transmission latency.

4 Simulation Results

We implemented our protocol in *ns-2* [15] by modifying the IEEE 802.11 MAC and physical layers. In our experiments, we compared three protocols, S-MAC, T-MAC, and MT-MAC. When a node is not transmitting, its radio is set to receive. Energy consumption in the model is based on the amount of energy the real nodes use.

We assume that the node consumes 20 μA while sleeping, 4 mA while receiving, and 10 mA while transmitting a DC-balanced signal [5]. Most parameters used in our simulations are the same as in [5]. There are 100 nodes in 10 by 10 grids in our simulation. We also have selected a radio range so that non-edge nodes all have 8 neighbors. And we used a randomized shortest path routing method for the nodes-to-sink communication pattern. Next hops are eligible if they have a shorter path to the final destination than the sending node.

We tested S-MAC protocol with a frame length of one second, and with several lengths of the active time, varying from 75 ms to 900 ms in our simulation. And for the T-MAC protocol, we used a frame length of 600 ms and interval TA with a length of 15 ms. For packet inter-arrival time we assume Poisson packet arrivals. Results shown in this paper are obtained through 8 times simulations.

Figure 2 shows the energy consumption of three protocols. As we can see MT-MAC consumes less energy. We can also see that the energy consumption does not heavily depend on the packet inter-arrival time.

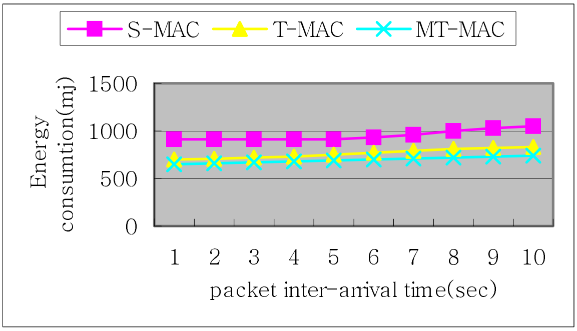


Fig. 2. Energy consumption of the protocols

This is because in MT-MAC, a source node can enter sleep mode early and awake late. In other words, when packet transmissions do not happen frequently but packets usually arrive just before time TA, then the sleep period will be increased in MT-MAC. So MT-MAC shows good performance when packets are transmitted almost constant time periods.

End-to-end delay comparison for the protocols is shown in the Figure 3. As described previous section, MT-MAC uses variable frame time according to the network traffic. So when there are many packet transmissions in the network, sleeping period will be reduced and nodes can send packets more quickly.

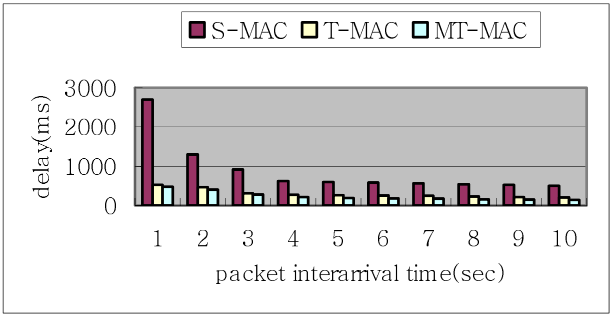


Fig. 3. End-to-end delay comparisons

5 Conclusions

The important design issue for medium access control protocols in wireless sensor networks is to save energy. The main sources of energy wastage are collisions, idle listening, overhearing, and control packet overhead. One approach to prevent energy wastage due to above sources is to control the node receiver by setting it to sleep mode when no data is expected. Wake-up based power management protocol is one of the schemes to save energy in WSNs. The main issue of these works is in the design of novel sleep scheduling schemes wherein nodes turn off their communication radios during the sleep.

This paper presents a new MAC protocol for wireless sensor networks. The proposed protocol, MT-MAC is an extension version of T-MAC. T-MAC automatically adjusts its sleep and active periods based on the network traffic patterns of a node’s neighbors. T-MAC focuses on when the transition occurs from active state to sleeping state. But they didn't consider when the sleeping period ends. MT-MAC also considers network traffic for determining active and sleep periods. But in MT-MAC packet inter-arrival time and the active period are used to determine variable frame time and sleep period. MT-MAC has very good energy conserving proprieties comparing with previous results such as S-MAC and T-MAC. Future work will focus on the effects of the parameters used.

Acknowledgments

This work was supported by Ministry of Education and Human Resources Development through the Embedded Software Open Education Resource Center (ESC) at Sangmyung University.

References

1. E. Shih, P. Bahl, and M. J. Sinclair, “Wake on Wireless: An Event Driven Energy Saving Strategy for Battery Operated Devices,” in Proceedings of the Eighth Annual International Conference on Mobile Computing and Networking, 2002
2. C. S. Raghavendra and S. Singh, “PAMAS-power aware multi-access protocol with signaing for ad hoc networks,” *SIGCOMM Computer Communication Rev.*, vol. 28, no. 3, pp. 5–26, 1998
3. Yuan Li, Wei Ye, John Heidemann, “Schedule and Latency Control in S-MAC,” Poster, in *UCLA CENS research review, 2003*
4. Rong Zheng, Jennifer C. Hou and Lui Sha, “Asynchronous Wakeup For Ad Hoc Networks, *“ in Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2003, pp. 35–45
5. Matthew J. Miller and Nitin Vaidya, "Minimization Energy Consumption in Sensor Networks Using a Wakeup Radio,” in *Proc. Wireless Communications and Networking Conf.*, 2004, pp. 120–125
6. Xue Yang and Nitin H. Vaidya, "A Wakeup Scheme for Sensor Networks: Achieving Balance between Energy saving and End-to-end Delay,” *in Proceedings of the IEEE Real-Time and Embedded Technology and Applications Symposium, 2004*, pp. 19–26
7. Ramaraju Kalidindi, Lyndia Ray, Rajgopal Kannan, Sitharama Lyengar, "Distributed Eergy Aware MAC Layer Protocol For Wireless Sensor Networks," in *International Conference on Wireless Networks*, Las Vegas, Nevada, June 2003
8. Gang Lu, Bhaskar Krishnamachari, Cauligi S. Raghavendra, "An Adaptive Energy-Efficient and Low-Latency MAC for Data Gathering in Wireless Sensor Networks," *Proc. of the 18th International Parallel and Distributed Processing Symposium (IPDPS'04)*
9. R. Zheng and R. Kravets, “On-demand Power Management for Ad Hoc Networks,” *in Proceedings of the 22nd International Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2003),* 2003
10. C. Schurgers, V. Tsiatsis, S. Ganeriwal, and M. Srivastava, ”Topology Management for Sensor Networks: Exploiting Latency and Density,” *in Proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2002),* 2002,pp. 135–145
11. MICA2 Mote Datasheet. http://www.xbow.com/products/Product_pdf_files/Wireless_pdf/6020-0042-01_A_MICA2.pdf
12. Wei Ye, John Heidemann, Deborah Estin, "An energy-Efficient MAC Protocol for Wireless Sensor Networks,” *In 21st Conference of the IEEE Computer and Communications Societies (INFOCOM)*, vol. 3, pages 1567-1576, June 2002
13. Tijs van Dam and Koen Langendoen, "An Adaptive Energy-Efficient MAC Protocol for Wireless Sensor Networks." *in Proceedings of the 1st international conference on Embedded networked sensor systems (SenSys ’03*), 2003, pp. 171–180
14. Matthew J. Miller and Nitin H. Vaidya, “Power save mechanisms for multi-hop wireless networks,” *in Proceedings of the 1st International Conference on Broadband Networks, 2004,* pp. 518–526
15. *ns-2* Network Simulator, http://www.isi.edu/nsnam/ns

Sensor Management of Multi-sensor Information Fusion Applied in Automatic Control System

Yue-Song Lin and An-ke Xue

Institute of Intelligent Information and Control Technology, Hangzhou Dianzi University,
Hangzhou, 310018, China
lysjxb@mail.hz.zj.cn

Abstract. Based on the classical control system, the framework of information fusion in stochastic optimal control system and the architecture of multisensor information fusion system are developed. The fusion estimation algorithm for control system is applied to state estimation. A structure of sensor management(SM) method and the design method are developed. The practical example shows the performance of state estimation is improved by applying information fusion. But if the model error is very large, the measurement model of each sensor must be selected rightly to ensure the performance of state estimation.

1 Introduction

Multisensor fusion can combine the information from different sensors to obtain the more accurate information about the system. Twenty years ago, information fusion was first applied in the field of military. Since then, this technique has greatly developed in other application fields [1-4].

In the industrial processing control system, some sensors are very precise and expensive and the number of sensors also increases greatly. But sensors can’t measure some state values directly in process control system due to the limitation process technology. Moreover, the information provided by each single sensor is generally uncertain, incomplete, inconsistent, or imprecise. Multi-sensor system can improve accuracy of measurement, increase system robust, and enhance reliability comparing to single sensor [5,6]. The performance can greatly be improved by using the information fusion technique in the automatic control system because of the improved state estimation.

So many state fusion estimation algorithms are developed [7-11], but the design methods of multisensor system is seemed to catch less interest of researchers. The performance of state fusion estimation and controlling can be guaranteed by reasonable structure of multisensor. This is a sensor management(SM) problem.

SM methods of information fusion can help designers choose adaptive sensors, sensing point and operating modes of sensors based on the performance of the whole system and develop a more economic system [12]. So how to utilize different information sources and how to design a more economic system is the key issues in the future automatic system.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 607 – 615, 2005.

This paper develops the framework of information fusion applied in stochastic optimal control system in Section II. Section III discusses the application of SM in control system. At last an example shows the improvement of state estimation by fusion measurements from multisensor system.

2 Information Fusion For Control System

2.1 System Framework

Based on the theory of automatic control, the framework of information fusion in stochastic optimal control system is presented as in Fig.1. In this framework (Fig. 1), the Kalman filter is replaced by information fusion. The architecture of multi-sensor information fusion system is detailed as Fig. 2.

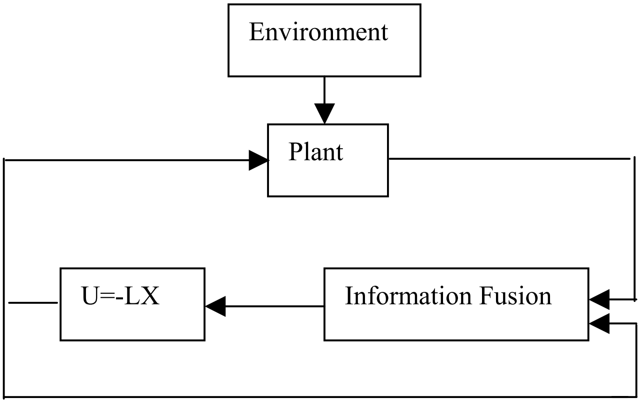


Fig. 1. Information fusion in stochastic optimal control system

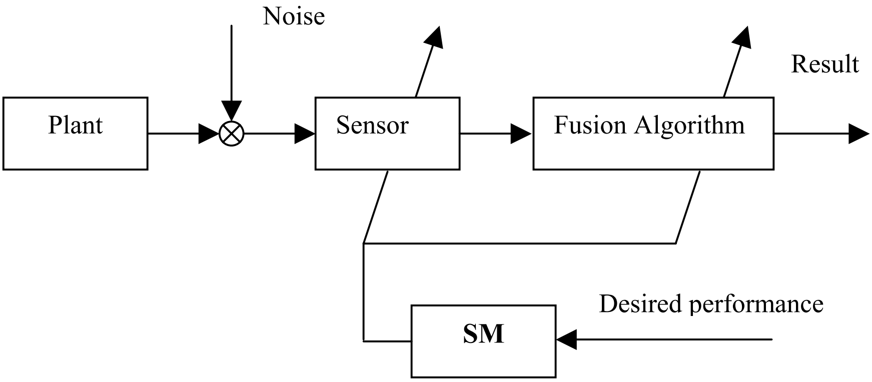


Fig. 2. The architecture of multisensor information fusion system

2.2 Dynamic Models

Consider a dynamic control system that is sensed by N sensors. The plant and sensors are modeled by the following discrete-time state models:

$$x(k) = A(k)x(k-1) + B(k)u(k) + v(k) \tag{2.1}$$

$$y_n(k) = c_n(k)x(k) + w_n(k) \qquad n = 1,2,\cdots N \tag{2.2}$$

where *k* represents the discrete-time index, *x*(*k*) is the state-vector, *u*(*k*) the input vector, $y_n(k)$ measurement vectors, *v*(*k*) and $w_n(k)$ Gaussian noise with zero mean, covariance matrices *Q*(*k*) and $R_n(k)$, respectively.

2.3 State Fusion Estimation

Let the observation vector $y_n(k)$ be fused into *N* blocks:

$$Y(k) = [y_1^T(k), y_2^T(k), \cdots, y_{N_j}^T(k)]^T \tag{2.3}$$

Then the measurement model is

$$Y(k) = C(k)x(k) + W(k) \tag{2.4}$$

where

$$C^T(k) = [c_1^{\ T}(k) \quad c_2^{\ T}(k) \quad \cdots \quad c_N^{\ T}(k)]^T \tag{2.5}$$

$$W^T(k) = [w_1^{\ T}(k) \quad w_2^{\ T}(k) \quad \cdots \quad w_N^{\ T}(k)]^T \tag{2.6}$$

$$COV[W(k) \quad W(j)^T] = R(k)\delta_{kj} \tag{2.7}$$

where

$$R^{-1}(k) = diag[R_1^{\ -1}(k) \quad R_2^{\ -1}(k) \quad \cdots \quad R_N^{\ -1}(k)] \tag{2.8}$$

Then based on Kalman filter, we have the state estimation of the system (2.1) and (2.4) as the following

$$\hat{x}(k \mid k) = \hat{x}(k \mid k-1) + K(k)[Y(k) - C(k)\hat{x}(k \mid k-1)] \tag{2.9}$$

where

$$\begin{aligned} K(k) &= P(k \mid k)C^T(k)R^{-1}(k) \\ &= P(k \mid k)[c_1^{\ T}(k)R_1^{\ -1}(k), c_2^{\ T}(k)R_2^{\ -1}(k), \cdots, c_N^{\ T}(k)R_N^{\ -1}(k)] \\ &= [K_1(k), \quad K_2(k), \quad \cdots, \quad K_N(k)] \end{aligned} \tag{2.10}$$

$$\hat{x}(k \mid k-1) = A(k)\hat{x}(k-1 \mid k-1) + B(k)u(k) \tag{2.11}$$

$$P^{-1}(k \mid k) = P^{-1}(k \mid k-1) + C^T(k)R^{-1}(k)C(k) \tag{2.12}$$

$$P(k \mid k-1) = A(k)P(k-1 \mid k-1)A^T(k) + Q(k) \tag{2.13}$$

Then by (2.4), (2.5) and (2.6), we can obtain

$$\begin{aligned} \hat{x}(k \mid k) &= \hat{x}(k \mid k-1) + \sum_{i=1}^{N} K_i(k)[y_i(k) - c_i(k)\hat{x}(k \mid k-1)] \\ &= \hat{x}(k \mid k-1) + \sum_{i=1}^{N} P(k \mid k)c_i^{\ T}(k)R_i^{\ -1}(k)[y_i(k) - c_i(k)\hat{x}(k \mid k-1)] \end{aligned} \tag{2.14}$$

and by (2.12), (2.5) and (2.6), the covariance of estimation is

$$P^{-1}(k \mid k) = P^{-1}(k \mid k-1) + \sum_{i=1}^{N} c_i^{\ T}(k)R_i^{\ -1}(k)c_i(k) \tag{2.15}$$

Thus, (2.14) and (2.15), together with (2.11) and (2.13), are the state fusion estimation algorithms.

3 SM in Control System

The input of SM is desired performance, and SM can be designed off-line on basis of desired performance. Based on the input, SM decides what sensor or what group of sensors the system employs, what work mode or function they have, where to point them, how to control and supervise them, how efficient current fusion algorithm is, and how to develop performance of overall system.

In practical control system, SM is dramatically limited. The reasons are as the following: 1) Some sensors are very expensive. It is impossible to equip them because of the limitation of the whole production cost. 2) Some states value can't be measured by sensors directly, partly due to the imperfection of sensor technology itself, and partly due to the limitation of process technology.

Therefore, SM has to consider the practical situation of industrial process. The practical work is often to select the appropriate measurable point, more inexpensive sensors and decide its work mode, and choose the suitable information fusion algorithm to obtain better performance of state estimate by less cost.

The next example shows how to design a simple sensor manager off-line. To simplify the problem, we assume the control input is zero and only design the work mode and measurement point of sensors.

The plant model is as following:

$$x(k) = \begin{bmatrix} 0 & -0.16 \\ 1 & -0.1 \end{bmatrix} x(k-1) + v(k) \tag{3.1}$$

Two sensors measure the system, the measurement models:

$$y_1(k) = \begin{bmatrix} 0 & 0.5 \end{bmatrix} x(k) + w_1(k) \tag{3.2}$$

$$y_2(k) = \begin{bmatrix} c_{21} & c_{22} \end{bmatrix} x(k) + w_2(k) \tag{3.3}$$

where $v(k)$, $w_1(k)$, $w_2(k)$ is Gaussian noise with zero mean, covariance matrices 1, 0.5, 0.5, respectively. The measurement model shows that due to the limitation of dynamic system, sensor 1 is fixed to measure only one state: x_2, and its working mode is unchanged. The measuring point and sensing mode of sensor 2 can be selected, but there is still a limit:

$$c_{21},\ c_{22} \in \begin{bmatrix} 0 & 10 \end{bmatrix} \tag{3.4}$$

We choose the steady covariance of Kalman filter P to be desired performance,

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \tag{3.5}$$

The two sensors are integrated into the following measurement model:

$$y(t) = \begin{bmatrix} 0 & 0.5 \\ c_{21} & c_{22} \end{bmatrix} x(t) + W(t) \tag{3.6}$$

where $W(t) \sim N(0, R(t))$, $R(t) = diag(0.5, 0.5)$. The quantity of p_{11}, p_{12}, p_{22} against (c_{21}, c_{22}) is reported in Fig.3, Fig.4 and Fig.5.

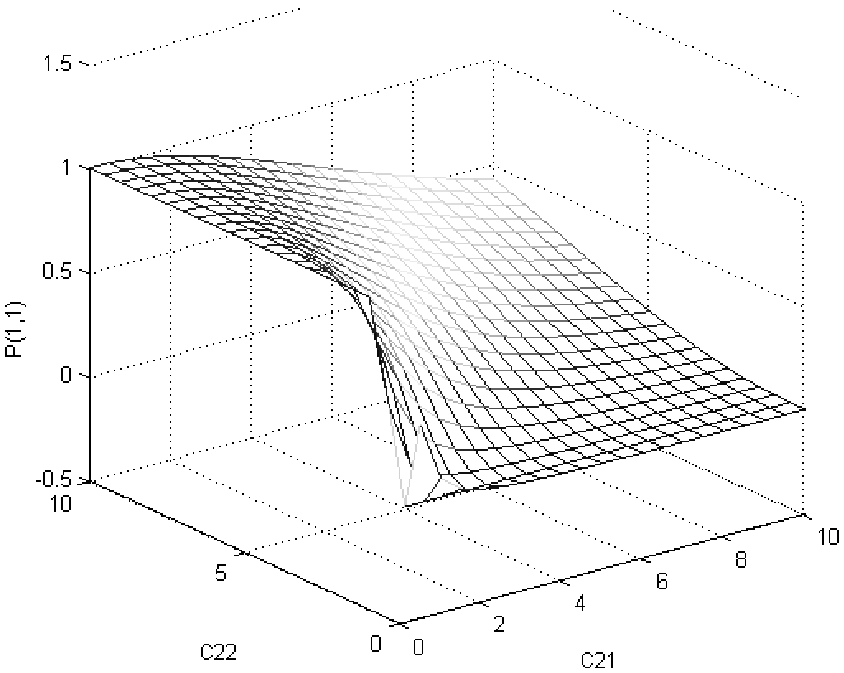


Fig. 3. p_{11} against (c_{21}, c_{22})

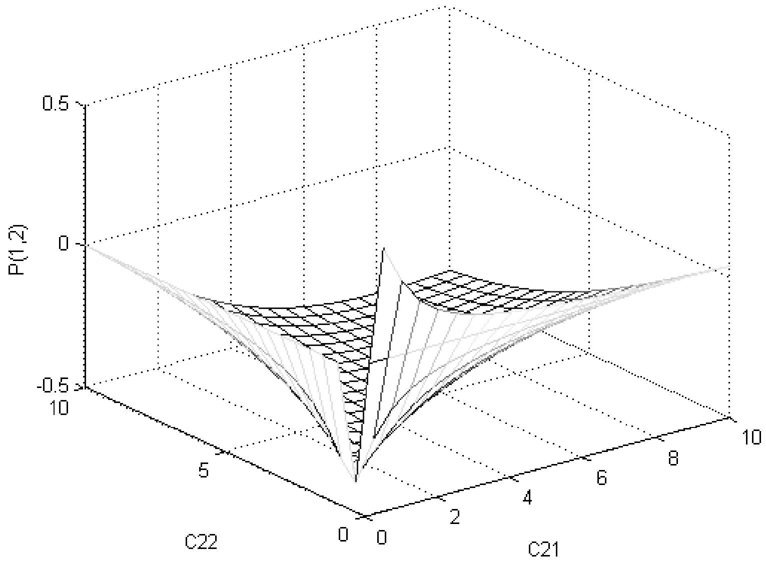


Fig. 4. p_{12} against (c_{21}, c_{22})

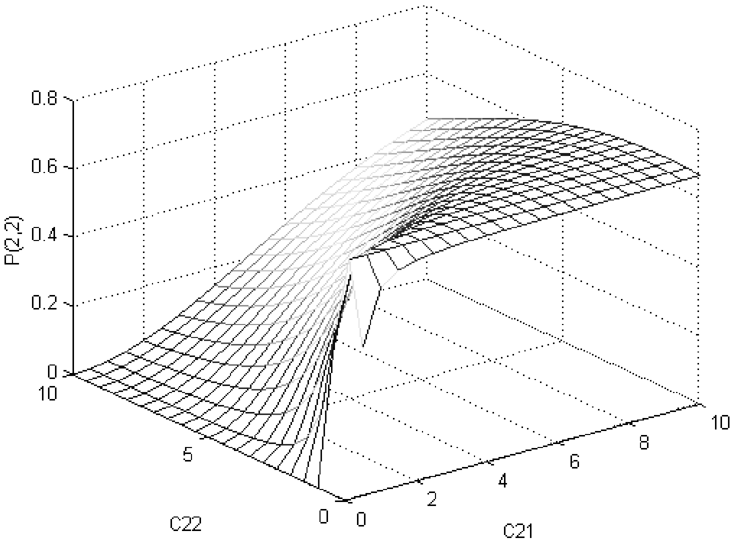


Fig. 5. p_{22} against (c_{21}, c_{22})

From the figures, we can see the changing tendency of p_{12}, p_{11}, p_{22} against (c_{21}, c_{22}) is different. Therefore if all the states are critical in the control system, all the covariance must be considered. Values from 5 to 10 of (c_{21}, c_{22}) are advisable.

4 An Example

Six scenarios are tested by (2.14) to illustrate the improvement achieved using information fusion. In Scenario I, plant model and measurement models are the same as (3.1), (3.2) and (3.3). In Scenario II and III, we consider the error of plant model. We presume the real plant dynamic in Scenario II is

$$x(k) = \begin{bmatrix} 0.1 & -0.14 \\ 1.1 & -0.1 \end{bmatrix} x(k-1) + v(k) \tag{4.1}$$

and in Scenario III is

$$x(k) = \begin{bmatrix} 0.9 & -0.14 \\ 1.7 & -0.1 \end{bmatrix} x(k-1) + v(k) \tag{4.2}$$

The plant model for state estimation is (3.1). Then the covariance error of process noise is considered in Scenario IV. The real covariance of process noise is 1.5 while the given covariance is 1. At last, the errors between the real covariance of measurement noise and the given covariance model are considered. The real measurement covariance of two sensors is (1, 0.8) and (0.7, 0.6) in Scenario V and Scenario VI, respectively, and the given measurement covariance model for state estimation is (0.5,0.5).

Table 1. Covariance of state estimation by Kalman filter of one sensor and by information fusion of two sensors

			Sensor 2	Sensor 2	Sensor 2	Informa-	Informa-	Informa-
Scenario		Sensor 1	C_{21}=7	C_{21}=1	C_{21}=10	tion	tion	tion
			C_{22}=5	C_{22}=10	C_{22}=1	fusion I	fusion II	fusion III
I	P_{11}	1.0258	0.4093	0.9937	0.0151	0.2867	0.9889	0.0116
	P_{12}	0.0074	-0.5650	-0.0991	-0.1021	-0.3914	-0.0984	-0.0670
	P_{22}	1.0085	0.7997	0.0149	1.0278	0.5540	0.0148	0.6714
II	P_{11}	1.0294	0.4122	0.9965	0.0152	0.2952	0.9917	0.0150
	P_{12}	0.0589	-0.5693	-0.0991	-0.1033	-0.3945	-0.0984	-0.0672
	P_{22}	1.0598	0.8060	0.0148	1.0405	0.5622	0.0147	0.6815
III	P_{11}	2.5871	0.4081	0.0204	0.0204	0.2934	0.0140	0.0159
	P_{12}	1.5945	-0.5615	-0.1596	-0.1596	-0.3916	-0.0937	-0.0854
	P_{22}	2.3843	0.7928	1.6557	1.6557	0.5578	0.9780	1.0457
IV	P_{11}	1.5321	0.6120	1.4905	0.0203	0.3754	1.4808	0.0156
	P_{12}	0.0075	-0.8488	-0.1488	-0.1533	-0.5191	-0.1474	-0.0781
	P_{22}	1.2564	1.1970	0.0198	1.5393	0.7560	0.0197	0.8990
V	P_{11}	1.0389	0.4116	0.9938	0.0180	0.3729	0.9938	0.0168
	P_{12}	0.0146	-0.5635	-0.0989	-0.1020	-0.5075	-0.0988	-0.0898
	P_{22}	1.5212	0.8028	0.0178	1.0307	0.7218	0.0178	0.9058
VI	P_{11}	1.0310	0.4101	0.9938	0.0161	0.3210	0.9908	0.0135
	P_{12}	0.0103	-0.5645	-0.0990	-0.1021	-0.4379	-0.0985	-0.0076
	P_{22}	1.2136	0.8008	0.0159	1.0288	0.6210	0.0158	0.7667

Table 2. Covariance reduction ratio of each scenario

Scenario		COR_{11}	COR_{21}	COR_{12}	COR_{22}	COR_{13}	COR_{21}	COR_{mean}
I	P_{11}	0.7205	0.2995	0.0360	0.0048	0.9887	0.2318	0.3802
	P_{22}	0.4507	0.3072	0.9853	0.0067	0.3343	0.3468	0.4052
II	P_{11}	0.7132	0.2838	0.0366	0.0048	0.9854	0.0132	0.3395
	P_{22}	0.4695	0.3025	0.9861	0.0068	0.3570	0.3450	0.4111
III	P_{11}	0.8866	0.2811	0.9946	0.3137	0.9939	0.2206	0.6151
	P_{22}	0.7661	0.2964	0.5928	0.4137	0.5614	0.3684	0.4998
IV	P_{11}	0.7550	0.3866	0.0335	0.0065	0.9898	0.2315	0.4005
	P_{22}	0.3983	0.3684	0.9843	0.0051	0.2845	0.4160	0.4094
V	P_{11}	0.6411	0.0940	0.0434	0	0.9838	0.0667	0.3048
	P_{22}	0.5255	0.1009	0.9883	0	0.4045	0.1212	0.3567
VI	P_{11}	0.6887	0.2173	0.0390	0.0030	0.9869	0.1615	0.3494
	P_{22}	0.4883	0.2245	0.9870	0.0063	0.3682	0.2548	0.3882

The result is shown in Table.1. In the table p_{11}, p_{12} and p_{22} is defined as (3.5). The results shown in the first column are the state estimation covariance of sensor 1. The measurement model is (3.2). Data in the second to the firth column are the covariance of state estimation only applying the measurement from sensor 2. The measurement model is (3.2), where (c_{21}, c_{22}) is (7,5), (1,10) and (10,1), respectively. The fusion measurement model is (3.6), where (c_{21}, c_{22}) is (7,5), (1,10) and (10,1) in information fusion I, II and III, respectively. The results of three fusions are shown in the fifth to the seventh column in Table.1.

From the data of Table.1, we can see the covariance p_{11} and p_{22} estimated by information fusion are smaller than by only one sensor in all scenarios except in scenario V. The covariance reduction ratio COR is defined as

$$COR_{ij} = \frac{PS(i) - PF(j)}{PS(i)} \tag{4.3}$$

where PS and PF are p_{11} or p_{22} estimated by one sensor and by information fusion, respectively, i=1 ,2 , j=1 ,2 ,3. The mean of covariance reduction ratio $\mathrm{COR}_{\mathrm{mean}}$ is defined as:

$$COR_{mean} = \sum_{i=1}^{2} \sum_{j=1}^{3} COR_{ij} \Big/ 6 \tag{4.4}$$

Notice that the value of $\mathrm{COR}_{\mathrm{mean}}$ is between 0.3048 and 0.6151. The largest is in Scenario III and the smallest in Scenario V. From the table, we can conclude that state estimation by information fusion have greatly developed the robust of estimation. For example, the model error in Scenario III is larger than in Scenario II, but the mean of covariance reduction ratio $\mathrm{COR}_{\mathrm{mean}}$ in Scenario III is more than that in Scenario II. Besides, COR and $\mathrm{COR}_{\mathrm{mean}}$ in Scenario II and Scenario III almost have no difference. This means the state estimation is more insensitive to the model errors through fusing measurement data from different sensor.

Comparing the different errors of model, we can note that the information fusion system is more sensitive to the covariance error of measurement noise compared with other system error. The mean of covariance reduction ratio $\mathrm{COR}_{\mathrm{mean}}$ is the smallest in Scenario V. We can also note that COR_{22} is 0 in Scenario V, where the measurement model of sensor 2 is (1,10) and the fusion model of two sensors (0, 0.5) and (1,10), but CORs with other measurement model and fusion model are nonzero. Therefore, the performance of estimation can't be developed when the measurement model is selected incorrectly in this condition. To improve the performance of the system, the measuring point and the sensing mode of sensors must be adjusted.

5 Conclusion

In this paper we discuss the application of information fusion technique in control system and develop state fusion estimation algorithm and SM for the control system. The example shows the performance of state estimation is greatly developed by information fusion compared with by single sensor. We can also conclude from the example that the estimated state is more robust against the model error, especially the error of process model, through combining data from different sensors. However, if the model error is very large, the measurement model of each sensor must be selected rightly to ensure the performance of state estimation.

The work in this paper is a first step. We expect the result presented can be useful for further work on the design of a practical system. This subject is under investigation.

Acknowledgements

This work is supported by the Key National Natural Science Foundation of China (No.60434020).

References

1. I. Wadi, R. Balendra: An Intelligent Approach to Monitor and Control the Blanking Process. Advances in Engineering Software. 30(1999) 85-92
2. P. G. Mathews, M. S. Shummugam: Neural-network Approach for Predicting hole quality in Reaming. International Journal of Machine Tools & Manufacture. 39(1999) 723-730
3. Hrianmayee Vedam: Signed Digraph Based Multiple Fault diagnosis. Computer Chem. Engng. 21(1997) 655-660
4. Shang-liang Chen, Y. W. Jen: Data Fusion Neural Network for Tool Condition monitoring in CNC Milling Machining. International Journal of Machine Tools & Manufacture. 40(2000) 381-400
5. JIN Xue-bo, SUN You-xian: Optimal Fusion Estimation Covariance of Multisensor Data Fusion on Tracking Problem. Proceedings of the 2002 IEEE International Conference on Control Applications. (2002) 1288-1289
6. JIN Xue-bo: State fusion estimation covariance of measurement fusion. Advances in Modelling Series B: Signal Processing and Pattern Recognition. 6(2004) 43-52
7. K. C., Chang, Saha, R. K., Bar-Shalom, Y.: On optimal track-to-track fusion. IEEE Transaction on Aerospace and Electronic Systems. (1997) 1271-1276
8. Qiang Gan, Chis J. Harris: Comparison of Two Measurement Fusion Methods for Kalman-Filter-Based Multisensor Data Fusion. IEEE Transaction on Aerospace and Electronic Systems. 1(2001) 273-280
9. JIN Xue-bo, SUN You-xian: Optimal State Estimation for Data Fusion with Correlated Measurement Noise. Journal of Zhejiang University. 1(2003) 60-64
10. JIN Xue-bo, SUN You-xian: Distributed uncorrelated optimal fusion algorithm for multisensor system. Fifth World Congress on Intelligent Control and Automation Conference Proceedings (2004) 1575-1579
11. Chris J. Harris, Qiang Gan: State estimation and multi-sensor information fusion using data-based neurofuzzy local linearisation process models. Information fusion 2(2001) 17-29
12. G. W. Ng, K. H. Ng: Sensor Management – What, Why and How, Information Fusion. 1(2000) 67-75

A Survey of the Theory of Min-Max Systems*

Yiping Cheng

School of Electronic and Information Engineering,
Beijing Jiaotong University, China
ypcheng@ustc.edu

Abstract. Min-max systems are discrete event systems whose timing involves the maximum, minimum, and addition operations. In recent years, progresses have been made in such topics as cycle time computation, ultimate periodicity of trajectories, structural properties, cycle time assignment, etc. They are surveyed in this paper. Furthermore, as an attempt to open new directions for further research we propose two new models referred to as *and-or net* and *and-or event graph*. It is found that timed and-or event graphs can be algebraically described by min-max systems. We hope that these new models would accommodate more new results.

1 Introduction

Min-max systems are discrete event systems whose behavior can be modeled by a state transition equation $x(k+1) = F(x(k))$ where F is a min-max function. A min-max function $F : \mathbb{R}^n \to \mathbb{R}^n$ is composed of n scalar-valued min-max functions which are built from terms of the form $x_i + a$, where $1 \leq i \leq n$ and $a \in \mathbb{R}$, by application of finitely many max and min operations.

Min-max functions/systems were introduced by Olsder and Gunawardena as a natural generalization of max-plus matrices [1] and as a mathematical tool to describe the timing of asynchronous systems [2,3]. Min-max systems arise in those discrete event systems where the maximum and addition operations are not enough to describe the timing behavior and consequently the minimum operation has to be introduced as well. When the minimum operation is introduced, the system becomes nonlinear in the max-plus algebraic sense and therefore the study of min-max systems is much harder than that of max-plus algebra. The theory of min-max systems was reported to have applications in such areas as digital circuits [4,5,6], manufacturing plants [7], traffic networks, etc. For those readers who are not quite familiar with the basic concepts of min-max functions, the reference [3] is a good starting point.

In recent years, considerable progresses have been made in a few topics of the theory of min-max systems such as cycle time computation, ultimate periodicity of trajectories, structural properties, cycle time assignment, etc. It is a purpose

* This work was supported by the National Natural Science Foundation of China under grant 60404010.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 616–625, 2005.

of this paper to survey these new developments. However, as limited by our scope and paper length, this survey is by no means complete, for example, it does not cover the max-plus algebra and the theory of topical functions which are closely related to the theory of min-max systems.

2 Basic Properties

Min-max functions form a subclass of the class of topical functions which are homogeneous

$$\forall x \in \mathbb{R}^n, \forall h \in \mathbb{R}, F(x_1 + h, \cdots, x_n + h) = (F_1(x) + h, \cdots, F_n(x) + h) \ , \qquad (1)$$

monotone with respect to the usual product ordering on $\mathbb{R}^n$,

$$\forall x, y \in \mathbb{R}^n, x \leq y \Rightarrow F(x) \leq F(y) \ , \qquad (2)$$

and nonexpansive in the sup norm (i.e. $\|x\| = \max_{1 \leq i \leq n} |x_i|$),

$$\forall x, y \in \mathbb{R}^n, \|F(x) - F(y)\| \leq \|x - y\| \ . \qquad (3)$$

Besides the three foregoing fundamental properties, there are also some properties that are of interest to researchers however some of them are only exhibited by some min-max functions.

Property C(F): The cycle time $\chi(F) = \lim_{k \to \infty} F^k(\xi)/k$ exists in $\mathbb{R}^n$, where F^k is the function obtained by k times applying F. Note that if for some ξ the limit $\chi(F)$ exists, then for all ξ this limit exists and is independent of ξ, because F is nonexpansive in the sup norm.

Property GE(F): F admits a cycle time and generalized eigenvector $(\eta, v) \in \mathbb{R}^n \times \mathbb{R}^n$ such that for all $k \geq 0$, $F(v + k\eta) = v + (k+1)\eta$.

Property I(F): F has a cycle time with identical coordinates, i.e. there is a $\lambda \in \mathbb{R}$ such that $\chi(F) = (\lambda, \cdots, \lambda)$.

Property E(F): F admits an eigenvalue $\lambda \in \mathbb{R}$ and corresponding eigenvector $x \in \mathbb{R}^n$ such that $F(x) = \lambda + x$.

Let C denote the assertion that C(F) holds for all min-max function F. Let GE denote the assertion that GE(F) holds for all min-max function F. Being a fundamental theorem of min-max functions, GE was first proved in [8] which was published even before DEDS emerged as a discipline. It was later independently proved in [9] and [10], and its first constructive proof was provided in [11]. The theorem C follows immediately from GE.

The equivalence between I(F) and E(F) follows immediately from GE. A constructive proof of the equivalence without using GE was given in [12].

3 Computation of Cycle Time

The cycle time is the most important performance metric of a min-max system and therefore the computation of cycle time is a central problem of min-max functions/systems. As a starting point one needs to decide how the min-max functions are represented in the computer. Basically there are two ways to represent a min-max function. One is called "tree representation" where each component function is represented as a min-max expression like a tree according to the definition; the other is called "max-plus representation" where each component function is represented as a conjunctive form [3], i.e. the min of some max-only functions.

It is the more usual practice to use the max-plus representation. Let

$$f(x) = \min_{a \in \mathcal{R}} a \otimes x \; . \tag{4}$$

Let $\mathcal{S} = \mathcal{R}_1 \times \cdots \times \mathcal{R}_n$ where the operation $\times$ is defined to be rectangular product defined in [9], then $\mathcal{S}$ is a set of $n \times n$ matrices and

$$F(x) = \min_{A \in \mathcal{S}} A \otimes x \; . \tag{5}$$

The *duality theorem*, first introduced as a conjecture in [2], states that

$$\chi(F) = \min_{A \in \mathcal{S}} \mu(A) \tag{6}$$

where $\mu(A)$ is the cycle time of A as defined in the max-plus algebra [7]. Note that the duality theorem was originally given in dual form including both max-plus and min-plus expressions but for our purposes here only the max-plus expression is needed, moreover, the equivalence between (6) and the duality theorem in its original form can be easily established. The duality theorem is closely related to GE and the proofs of GE in [9,10,11] are also proofs of the duality theorem.

The duality theorem suggests an algorithm to compute the cycle time but it is exponential. Cochet-Terrasson, Gaubert and their co-workers proposed in [13,9] an algorithm to compute the spectral radius of max-plus matrices and an algorithm to compute the cycle time of min-max functions. The key idea inside their cycle time algorithm is *policy iteration* which has been used successfully in stochastic control. Cheng and Zheng proposed in [11] another policy-iteration-based algorithm and along with it a computer implementation was provided. There are two major differences between the two cycle time algorithms: one is that different policy improvement schemes are used in the two algorithms; and the other is that the former algorithm calculates the cycle time components in parallel, whereas the latter computes the cycle time by first calling a policy-iteration-based algorithm to find the spectral radius (the largest component of the cycle time), and then recursively calling itself to compute the cycle time of the reduced system.

Subiono and van der Woude [14] proposed power algorithms to compute the cycle time. Their algorithms rely on the ultimate periodicity of trajectories of min-max systems. However, in [14] no details are supplied on how to test whether the trajectory has entered a periodic behavior after a transient phase, which is seemingly trivial but in fact often very time and space consuming to which the contributing factors include the length of the transient phase, the period, the number of different periodicity patterns.

It is now still open to determine theoretically the complexity of the cycle time computing problem, and for any of the above-mentioned cycle time algorithms there are still no definite complexity results.

4 Ultimate Periodicity of Trajectories

It had long been observed that any trajectory $x(0), x(1), x(2), \cdots$ of any min-max system always ends up with a periodic behavior, i.e. there is some $K \in \mathbb{N}$, $\eta \in \mathbb{R}^n$, such that $x(k+p) = p\eta + x(k)$ for $k \geq K$. Its proofs appeared in [10,15] which were both based on a functional analysis result on nonexpansive maps. A purely algebraic proof of this fact is not yet available.

There exists a maximum period $M(n)$ for n-dimensional min-max systems however $M(n)$ is yet unknown. Moreover, no major progress has been reported in this direction since the writing of [16] and therefore §2.1.3 of [16] remains an up-to-date survey of this direction of research.

5 Structural Properties

For min-max functions a property is said to be structural if it does not depend on the numerical values (as long as they remain finite) characterizing the function.

The most studied structural property is perhaps the *SEE* (structural existence of eigenvalue) property which means that the min-max function has an eigenvalue for all values of its parameters. It should be noted that this property is in fact not a single property, but a family of properties since different grammars or structural conventions may be used in forming the min-max function. If the forming structure is the most free one as allowed by the original grammar of min-max functions given in [3], then the *SEE* property is the *balanced property* introduced by Gunawardena in [3]. If one imposes a limitation on the above structure by stipulating that any parameter a can only appear in a term like $x_i + a$ (this limitation does not result in any loss of expressive power in that all min-max functions can still be formed by using this structure), then the *SEE* property becomes the *inseparability* introduced by Zhao in [17]. For bipartite min-max functions [1], the *SEE* property was studied by Olsder in [18] and later van der Woude and Subiono gave in [19] a sufficient and necessary condition called *irreducibility* in the language of max-plus algebra and min-plus algebra.

Zhao [17] showed that the test of inseparability is equivalent to the test of whether a monotone boolean function has a nontrivial fixed point, and Yang and Zhao [20] showed that the complement of the latter problem is NP-complete, therefore the test of inseparability is co-NP-complete. It was hence derived in [20] that the test of the balanced property is co-NP-hard.

For min-max functions that have an eigenvalue, it is also of interest to analyze the structure of the eigenspace (the set of eigenvectors). Zhao [21] showed that the decision problem of whether there are at least two eigenvectors that are not mutually “equivalent” (two eigenvectors are said to be equivalent if one can be obtained from the other by simply adding a finite number to all its components) is NP-complete. Cheng and Zheng [11] obtained a boolean necessary condition for the existence of a finite upper bound on $x_i - x_j$ for all the eigenvectors x of F.

6 Control or Cycle Time Assignment

Zhao [22], Chen and Tao [23,24,25] studied the control of min-max systems. A very general framework for the control of min-max systems can be formulated as follows. Consider the nonautonomous min-max system

$$x(k+1) = F(x(k), u(k)) \tag{7}$$

where $x(k)$ is $n \times 1$ state sequence, $u(k)$ is $p \times 1$ control sequence, $F : \mathbb{R}^{n+p} \mapsto \mathbb{R}^n$ is a "non-square" min-max function. A general form of control law is

$$u(k) = G(x(k)) \tag{8}$$

where $G : \mathbb{R}^n \mapsto \mathbb{R}^p$ is a (possibly non-square) min-max function implementing the control law. Through proper selection of control law one aims to achieve some desired properties for the closed-loop system, usually assigning the cycle time to a desired location, which closely parallels the pole assignment problem in linear systems theory. The control (or cycle time assignment) problem is extremely difficult for the general case, and to obtain results usually some simplifying assumptions on the forms of (7) and (8) are made.

In [22] the model of the controlled system takes the form:

$$\begin{aligned} x(k+1) &= F(x(k)) \vee G(u(k)) \\ y(k) &= H(x(k)) \\ u(k) &= K(y(k), y(k-1), \cdots, y(k-d)) \end{aligned}$$

where $\vee$ is the infix notation for max, d is a nonnegative integer and F, G, H, K are min-max functions with G, H, K possibly non-square. It is assumed that the feedback map K has all-nonnegative parameters since only nonnegative delay can be implemented in real systems. The control objective is to stabilize the system without degrading its performance, i.e. to make the closed-loop system have an eigenvalue equal to the largest cycle time component of the original system. The control law is obtained using the concept of inseparability thus the stabilization is rather a robust one.

In [23,24,25] the model of the controlled system takes the form:

$$\begin{aligned} x(k+1) &= F(x(k)) \vee B \otimes u(k) \\ y(k) &= C \otimes x(k) \\ u(k) &= K \otimes y(k) \end{aligned}$$

where F is a min-max function and B, C, K are max-plus matrices with appropriate sizes. The control objective is to assign the cycle time of the closed-loop system to a desired location. Several assignability conditions have been obtained in [23,24,25] mainly using graph-theoretic methods.

7 Models for the Future: And-Or Net and And-Or Event Graph

It is felt by the author of this paper that through the past few years of fast development the min-max systems theory is now becoming mature. To keep the momentum of research in this area it may be crucial to find new application backgrounds in order that new ideas, problems, and methods could be discovered. However for a broadened application area we believe that the current model is too restrictive and hence extensions are necessary to allow more real-world systems to be modeled.

Note that the max-plus algebra is an algebraic description of timed event graphs. Similarly, we would also expect min-max functions to be an algebraic description of a certain graph-based model which includes timed event graphs as a subclass. The model would not constitute a subclass of Petri nets [26], since in Petri nets a transition is fired only when it receives tokens from all its predecessor places, which is an "and-causality", but to introduce the min operation in timing it is necessary for the model to include its dual "or-causality". We therefore introduce here a model named *and-or net*.

Definition 1 (and-or net). *An and-or net is a 4-tuple* (P, T, E, M_0)*, where* P *is the set of places;* T *is the set of transitions,* $T = T_a \cup T_o$ *and* $T_a \cap T_o = \emptyset$*,* T_a *is the set of and-transitions,* T_o *is the set of or-transitions;* $E \subseteq (P \times T) \cup (T \times P)$ *is the set of edges;* $M_0 : P \mapsto \mathbb{Z}$ *specifies for each place the initial number of tokens which must be nonnegative.*

If $(p, t) \in E$*, then* p *is said to be a predecessor place of* t*, and* t *is said to be a successor transition of* p*. If* $(t, p) \in E$*, then* t *is said to be a predecessor transition of* p*, and* p *is said to be a successor place of* t*.*

An and-transition is said to be enabled if each of its predecessor places contains at least one token. An or-transition is said to be enabled if at least one of its predecessor place contains at least one token.

A firing of an enabled transition decreases the number of tokens by one for each of its predecessor places and increases the number of tokens by one for each of its successor places.

Notice that according to this definition a place may have negative number of tokens during the dynamic process. Intuitively, we may think of a negative number of tokens as a positive number of *anti-tokens*.

Definition 2 (and-or event graph). *An and-or net is called an and-or event graph if each place has exactly one predecessor and one successor transition.*

The above definitions deal only with the ordering of events, and questions pertaining to when events take place are not addressed. For questions pertaining to performance evaluation it is necessary to introduce time. This can be done in several ways. Here we choose one of them:

Definition 3 (timed and-or net). *A timed and-or net is a 5-tuple* (P, T, E, M_0, D) *where* P, T, E, M_0 *are defined as in Definition 1, and* $D : E \mapsto \mathbb{R}$ *specifies the delays of token propagation either from a transition to a place, or from a place to a transition.*

Furthermore, an and-transition is fired as soon as it receives tokens from all its predecessor places, an or-transition is fired as soon as it receives token from any of its predecessor places.

A definition of timed and-or event graph then follows naturally:

Definition 4 (timed and-or event graph). *A timed and-or net is called a timed and-or event graph if each place has exactly one predecessor and one successor transition.*

Note that for a timed and-or event graph, if only the timing of events (an event is a firing of a transition) is concerned, then it does not matter how the delay is distributed along the path from a predecessor transition to a successor transition of the same place, therefore we may assume without loss of generality that all delays are from transitions to places, i.e. there is no delay from places to transitions; in this case a transition will be fired as soon as it is enabled. This assumption shall be used in the following discussion.

To illustrate the concept of (timed) and-or event graph let us see an example depicted in Fig. 1. Note that circles indicate places, dots within a circle indicate tokens in the place, bars indicate and-transitions, boxes indicate or-transitions. The place whose predecessor transition is t_i and whose successor transition is t_j is denoted by p_{ij}. The delay from transition t_i to place p_{ij} is denoted by τ_{ij}. The number of tokens of p_{ij} is denoted by $M(p_{ij})$.

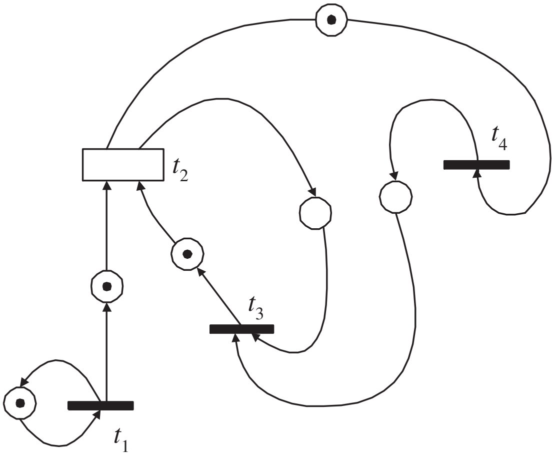


Fig. 1. An and-or event graph

Denote the time of the k-th firing of transition t_i by $x_i(k)$, then we have:

$$x_1(k) = x_1(k-1) + \tau_{11} \tag{9}$$

$$x_2(k) = \min(x_1(k-1) + \tau_{12}, x_3(k-1) + \tau_{32}) \tag{10}$$

$$x_3(k) = \max(x_2(k) + \tau_{23}, x_4(k) + \tau_{43}) \tag{11}$$

$$x_4(k) = x_2(k-1) + \tau_{24} \ . \tag{12}$$

We need an explanation why (10) holds. According to the firing rule, at time 0 the transition t_2 is fired, hence $x_2(1) = 0$; after this firing $M(p_{12})$ and $M(p_{32})$ both become 0, since t_2 is an or-transition we have $x_2(2) = \min(x_1(1) + \tau_{12}, x_3(1) + \tau_{32})$, assume without loss of generality that the first token from t_1 arrives at t_2 before the first token from t_3 does, i.e. $x_1(1) + \tau_{12} < x_3(1) + \tau_{32}$, then t_2 fires for the second time at $x_2(2) = x_1(1) + \tau_{12}$; after this firing $M(p_{12})$ becomes 0 and $M(p_{32})$ becomes -1, now we must stress that it is the arrival of the second, not the first, token from t_3 that forms an enabling condition for t_2's third firing, because when the first token from t_3 arrives at p_{32}, it is "annihilated" with the anti-token already in the place, and $M(p_{32})$ becomes 0, but not 1. Therefore $x_2(3) = \min(x_1(2) + \tau_{12}, x_3(2) + \tau_{32})$. By repeating this reasoning infinitely many times we obtain (10).

From (9–12) we obtain the following min-max system:

$$\begin{aligned} x_1(k) &= x_1(k-1) + \tau_{11} \\ x_2(k) &= \min(x_1(k-1) + \tau_{12}, \max(x_2(k-1) + \tau_{23}, x_4(k-1) + \tau_{43}) + \tau_{32}) \\ x_4(k) &= x_2(k-1) + \tau_{24} \ . \end{aligned}$$

However, not all timed and-or event graphs can be converted to min-max systems as easily as the above example is, because sometimes self-references will occur and they must be removed to obtain a min-max system. Self-references are caused by dead cycles within a (timed) and-or event graph. A (timed) and-or event graph can be viewed as a graph whose nodes are transitions and whose edges are places. A *dead cycle* within a (timed) and-or event graph is a cycle whose forming edges (places) have 0 as sum of their initial token numbers. For example, let us have a slight modification on Fig. 1 to change the initial token number of p_{32} to be 0, then p_{32} and p_{23} form a dead cycle. Hence (9,11,12) still hold but (10) should be changed as follows:

$$x_2(k) = \min(x_1(k-1) + \tau_{12}, x_3(k) + \tau_{32}) \ . \tag{13}$$

And combining (13,11) we obtain:

$$x_2(k) = \min(x_1(k-1) + \tau_{12}, \max(x_2(k) + \tau_{23}, x_4(k) + \tau_{43}) + \tau_{32}) \ . \tag{14}$$

There is a self-reference of $x_2(k)$ on the right-hand side of (14), but it can be removed. Since $\tau_{23} + \tau_{32} \geq 0$, the second operand of the min operation on the right-hand side of (14) is not less than $x_2(k)$, therefore (14) is equivalent to the following equation (15) and (9,12,15) constitute a min-max system:

$$x_2(k) = x_1(k-1) + \tau_{12} \ . \tag{15}$$

The foregoing argument can be generalized. We may conclude, though without a formal proof, that if a timed and-or event graph contains no dead cycle, or although it contains dead cycle(s) but the dead cycle(s) can be safely eliminated, then it can be algebraically described by a min-max system. It should also be noted that for this to hold it is crucial to allow negative numbers of tokens for places during the dynamic process.

The values of the models proposed here are threefold:

1. The logical model: and-or event graph, which was previously hidden under the timed algebraic model: min-max system, is now discovered.
2. By relaxing the restrictions imposed on timed and-or event graphs one can obtain various different extensions of min-max systems, which is a very systematic method of extension.
3. The and-or event graphs model and the and-or nets model are more friendly to the engineering system designers than the min-max systems model is. Moreover, since they give internal (with details of event occurrence mechanism) rather than only phenomenal description of the system's dynamic behavior, we may establish a supervisory control theory for them as we did for Petri nets.

8 Conclusion

In this paper the recent developments in the theory of min-max systems are surveyed and the models of and-or net and and-or event graph are introduced. The connections between min-max systems and the proposed models are revealed. It is hoped that the proposed models would open new directions for future research and accommodate more new results in this area.

References

1. Olsder, G.J.: Eigenvalues of dynamic max-min systems. Discrete Event Dynamic Systems **1** (1991) 177–207
2. Gunawardena, J.: Cycle times and fixed points of min-max functions. In Cohen, G., Quadrat, J.P., eds.: 11th International Conference on Analysis and Optimization of Systems. Volume 199 of LNCIS., Springer (1994) 266–272
3. Gunawardena, J.: Min-max functions. Discrete Event Dynamic Systems **4** (1994) 377–406
4. Burns, S.M.: Performance analysis and optimization of asynchronous circuits. PhD thesis, California Institute of Technology (1991)
5. Lee, T.K.: A general approach to performance analysis and optimization of asynchronous circuits. PhD thesis, California Institute of Technology (1995)
6. Sakallah, K.A., Mudge, T.N., Olukotun, O.A.: Analysis and design of latch-controlled synchronous digital circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits **11** (1992) 322–333
7. Baccelli, F., Cohen, G., Olsder, G., Quadrat, J.: Synchronization and Linearity. Wiley, New York (1992)

8. Kohlberg, E.: Invariant half-lines of nonexpansive piecewise-linear transformations. Math. Oper. Res. **5** (1980) 366–372
9. Gaubert, S., Gunawardena, J.: A non-linear hierarchy for discrete event dynamical systems. In: Proc. 4th Workshop on Discrete Event Systems, Cagliari, Italy (1998)
10. Olsder, G.J., Perennes, S.: Iteration of (min,max,+) functions. Draft [online], available: `http://citeseer.nj.nec.com/310205.html` (1997)
11. Cheng, Y., Zheng, D.Z.: A cycle time computing algorithm and its application in the structural analysis of min-max systems. Discrete Event Dynamic Systems: Theory and Applications **14** (2004) 5–30
12. Cochet-Terrasson, J., Gaubert, S., Gunawardena, J.: A constructive fixed point theorem for min-max functions. Dynamics and Stability of Systems **14** (1999) 407–433
13. Cochet-Terrasson, J., Cohen, G., Gaubert, S., Gettrick, M.M., Quadrat, J.: Numerical computation of spectral elements in max-plus algebra. In: Proceedings 5th IFAC conference on System Structure and Control. Volume 2., Elsevier (1998) 667–674
14. Subiono, van der Woude, J.: Power algorithms for (min, max, +)- and bipartite (min, max, +)-systems. Discrete Event Dynamic Systems **10** (2000) 369–389
15. Cheng, Y., Zheng, D.Z.: Ultimate periodicity of orbits for min-max systems. IEEE Trans. AC **47** (2002) 1937–1940
16. Gunawardena, J.: From max-plus algebra to non-expansive mappings: a nonlinear theory for discrete event systems. Theoretical Computer Science **293** (2003) 141–167
17. Zhao, Q., Zheng, D.Z., Zhu, X.: Structure properties of min-max systems and existence of global cycle time. IEEE Trans. on Automatic Control **46** (2001) 148–151
18. Olsder, G.J.: On structural properties of min-max systems. TWI report 93-95 (1993)
19. van der Woude, J., Subiono: Conditions for the structural existence of an eigenvalue of a bipartite (min, max, +)-system. Theoretical Computer Science **293** (2003) 13–24
20. Yang, K., Zhao, Q.: Balance problem of min-max systems is co-NP-hard. Systems & Control Letters **53** (2004) 303–310
21. Zhao, Q., Zheng, D.Z.: Structural properties of min-max functions and eigenspace of min-max functions. In: Lecture Notes on Control and Information Sciences (LNCIS). Volume 294. (2003) 393–400
22. Zhao, Q., Zheng, D.Z.: On stabilization of min-max systems. Automatica **39** (2003) 751–756
23. Chen, W.: Cycle time assignment of nonlinear discrete event dynamic systems. Systems Science and Mathematical Sciences **13** (2000) 213–218
24. Chen, W., Tao, Y.: On cycle time assignment of min-max systems. In: Proceedings of the IASTED International Conference on Circuits, Signals and Systems. (2003) 29–34
25. Tao, Y., Chen, W.: Cycle time assignment of min-max systems. International Journal of Control **76** (2003) 1790–1799
26. Peterson, J.L.: Petri Net Theory and the Modeling of Systems. Prentice Hall (1981)

Performance Bounds for a Class of Workflow Diagrams*

Qianchuan Zhao

Center for Intelligent and Networked Systems,
Department of Automation, Tsinghua University,
Beijing 100084, China
zhaoqc@tsinghua.edu.cn

Abstract. Recently the study of workflow diagrams has received considerable attention in business process modelling. Formal methods such as Petri nets have been used to analyze and verify of logical properties. However, to our best knowledge, due to the complexity caused by the extreme flexible nature of workflow processes, little work has been done on the performance analysis for workflow diagrams except intensive simulations or approximation analysis based on Stochastic Petri net (SPN) or queueing theory. In this paper, timed workflow diagrams with both AND and OR logic will be modelled and analyzed as stochastic min-max systems. We will provide provable bounds on average tournaround time. The OR logic (known also as the Discriminator [1]) requires that a downstream event happens whenever one of the upstream events happens. This is different from the AND logic modelling synchronization which requires that the output event happens when all input events happen.

1 Introduction

Recently the study of workflow diagrams has received considerable attention in business process modelling. Formal methods such as Petri nets have been used to analyze and verify logical properties. Due to the complexity caused by the extreme flexible nature of workflow processes, little work has been done on the performance analysis for workflow diagrams. Typical performance indices include resource availability and utilization, and average turnaround time. To our best knowledge, existing results are either based on stochastic Petri nets [2] [3] or queueing network theory [4]. Most of the results assume exponential distributions for task processing times. This assumption has been shown being able to give good approximation results. In this paper, we take a different approach which enable provable performance bounds. We will focus on deriving bounds for the first two order moments for the turnaround time for a class of workflow diagrams contructed recursively using elementary building blocks (AND, XOR and OR

* This work was supported in part by NSFC Grant No. 60274011, Fundamental Research Funds from Tsinghua University, Chinese Scholarship Council and NCET program of China.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 626–635, 2005.

blocks) under the assumption of infinite servers. Following the basic idea of max-plus algebra method to Discrete Event Systems (DES) (see e.g. [5]), we establish our main results by analyzing the algebraic relation of timing of the events. From this aspect, this work can be seen as an extension of the results in [6] to min-max systems with conditional behavior. Our results can be applied to general task processing time distributions. In this sense, our work is a complementary to the results of [2] and [3] which assume exponential distribution in firing time.

2 Model

Let T_i, $i = 1, \ldots, n$ be a finite set of task blocks. (F_j, J_j), $j = 1, \ldots, m$ are a finite set of pairs of fork and join points. Fork points can be AND-fork, XOR-fork or OR-fork. Join points can be AND-join, XOR-join and OR-join. We assume that for each j, F_j and J_j must be the same type.

Definition 1. *A workflow diagram is directed graph constructed recursively as follows based on the tree hierarchy of blocks.* `START` *and* `END` *are starting and ending blocks respectively.* `START` *and* `END` *are blocks. Other blocks are obtained only in one of the following ways:*

1. [serial block] a set of blocks connected in serial. Task blocks are serial blocks. Graphically, serially connected blocks are shown in Figure 1.

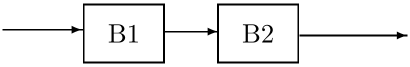


Fig. 1. Block diagram for serially connected blocks

2. [AND block] a block starts with an AND-fork point and ends with an AND-join point with equal number of out and in branches. Each branch is a block connecting the fork and join points. Graphically, an AND block is shown in Figure 2. The circles are fork and join points. A symbol '$\wedge$' will be used in the circles to indicate the AND-fork and AND-join.

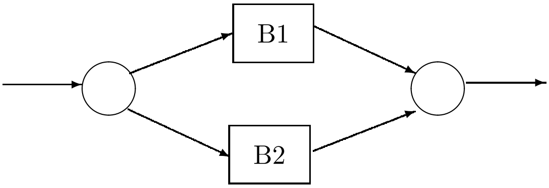


Fig. 2. Block diagram for AND, OR or XOR blocks

3. [XOR block] a block starts with an XOR-fork point and ends with an XOR-join point with equal number of out and in branches. Each branch is a block *connecting the fork and join points. Graphically, an XOR block is also shown in Figure 2. A symbol '$\oplus$' will be used in the circles to indicate the XOR-fork and XOR-join.*

4. [OR block] a block starts with an OR-fork point and ends with an OR-join point with equal number of out and in branches. Each branch is a block connecting the fork and join points. Graphically, an OR block is again shown in Figure 2. A symbol '$\vee$' will be used in the circles to indicate the OR-fork and OR-join.

5. [LOOP block] a block starts with an XOR-join point and ends with an XOR-fork point with two branches: the forward branch and the backward branch. The forward branch is a block connecting the join point and the fork point. The backward branch connecting the join point and the fork point. Graphically, LOOP block is shown in Figure 3.

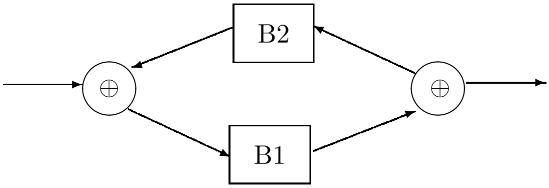


Fig. 3. Block diagram for loops

A workflow diagram W generated by T_i, $i = 1, \ldots, n$ and (F_j, J_j), $j = 1 \ldots, m$, is any block (will be denoted by TOP) such that

1. starts with the block `START` *and ends with the block* `END`*;*

2. conforms to the grammar of blocks;

3. all (fork,join) pairs in all blocks are distinct.

4. all task blocks appear only once in the diagram.

To evaluate the performance of workflow diagrams, we need to introduce time and routing related parameters. We attach with each task block T_i a random variable t_i representing the execution of the task. We assume that the distribution of t_i are D_i with mean value a_i and $\Delta(T_i) = E(t_i^2) < +\infty$. We also attach with each XOR-fork point with k branches a routing probability as p_{ju}, $u = 1, \ldots, k$ such that $p_{ju} \geq 0$ and $\sum_{u=1}^{k} p_{ju} = 1$. For a XOR-fork F_j of a LOOP block, we always assume this probability is given as p_{j1} (for exiting the loop) and p_{j2}(for the backward branch). The semantics of the execution of a workflow diagram is as follows: START, END, fork and join points cost no time; at each AND-fork, all branches are activated simultaneously and the corresponding AND-join is enabled at the time the tasks in all branches finish; at each OR-fork point, all branches are activated and the corresponding OR-join point is activated at the time when the tasks in one of the branches (such OR-join point is known also as the Discriminator [1]) finish; at the XOR-fork point in each XOR block, only one branch is activated according to the probability distribution and the corresponding XOR-join point is activated once the tasks on the activated branch finish; the XOR-join point in each LOOP block is activated for the first time once the LOOP block is activated, then the forward branch is activated, and once finished, the XOR-fork point F_j will be activated and the backward branch is then activated with probability p_{i2}. All random variables are assumed to be independent. For simplicity, we assume that there are an infinite number of servers and thus omit potential waiting of tasks.

3 Main Results

It is well known that as a discrete event system, the fork-join structure makes it very difficult to exactly evaluate the performance of workflow systems except very simple situations. As we mentioned in the Introduction, in this section, we will instead establish upper and lower bounds on average turnaround time. Our method will be based on the min-max formulation of the timing of each event as samples of the time and routing parameters in workflow diagrams.

We do this for four types of blocks and will show that the entire diagram's performance bounds can be obtained recursively. In a sample path of the execution of the workflow diagram, let $s(B)$ and $x(B)$ be the start and end time of activation for all activated block B. A block is activated if one of the task is activated. Denote $D(B) = x(B) - s(B)$ as the turnaround time, i.e., the time length of activation for block B. If two blocks B_1 and B_2 are connected in serial, it is assumed that $x(B_1) = s(B_2)$. We assume that $s(START) = 0$. Introduce a pair of real numbers $(\underline{D}(B), \overline{D}(B))$ for each block B as follows. We will prove that they are a lower and an upper bounds for the mean value $E(D(B))$ for the duration of the activation of the block. We also define $\Delta(B)$ as an upper bound for the second order absolute moment $E(D(B)^2)$ of $D(B)$.

- For all task blocks,

$$\underline{D}(T_i) = E(t_i) = a_i,\ \overline{D}(T_i) = E(t_i) = a_i,\ \Delta(T_i) = Var(t_i) + a_i^2. \tag{1}$$

- Let B be a serial block consisting of the block sequence $B_1 \to \ldots, \to B_l$.

$$\underline{D}(B) = \sum_{u=1}^{l} \underline{D}(B_u),\ \overline{D}(B) = \sum_{u=1}^{l} \overline{D}(B_u),$$
$$\Delta(B) = \sum_{u=1}^{l} \Delta(B_u) + \sum_{u \neq v} \overline{D}(B_u)\overline{D}(B_v). \tag{2}$$

- Let B be an AND block consisting of blocks in k branches $B_1, \ldots, B_k$.

$$\underline{D}(B) = \max_{u=1,\ldots k} \underline{D}(B_u),\ \overline{D}(B) = \max_{u=1,\ldots k} \overline{D}(B_u) + \sum_{u=1\ldots,k} \Delta(B_u),$$
$$\Delta(B) = \max_{u=1,\ldots,k} (\Delta(B_u)) + \sum_{u=1}^{k} \phi(D(B_u)), \tag{3}$$

where
$\phi(D(B_u)) = \sqrt{(\Delta(B_u) - (\underline{D}(B))^2) \times (\Delta(B_u) + 3(\overline{D}(B_u))^2)} + \Delta(B_u) - (\underline{D}(B))^2.$

- Let B be an XOR block consisting of blocks in k branches $B_1, \dots, B_k$ with F_j as its fork point.

$$\underline{D}(B) = \sum_{u=1,\dots k} \underline{D}(B_u) * p_{ju},\ \overline{D}(B) = \sum_{u=1,\dots k} \overline{D}(B_u) * p_{ju},$$
$$\Delta(B) = \sum_{u=1}^{k} \Delta B_u * p_{ju}. \tag{4}$$

- Let B be an OR block consisting of blocks in k branches $B_1, \dots, B_k$ with F_j as its fork point.

$$\underline{D}(B) = [\min_{u=1,\dots k} \underline{D}(B_u) - \sum_{u=1,\dots,k} \Delta(B_u)]^+,\ \overline{D}(B) = \min_{u=1,\dots k} \overline{D}(B_u),$$
$$\Delta(B) = \min_{u=1,\dots,k} \Delta(B_u) + \sum_{u=1}^{k} \phi(D(B_u)), \tag{5}$$

where $[x]^+ = \max(x, 0)$.

- Let B be a LOOP block consist of forward branch B_1 and backward branch B_2 with F_j as its fork point.

$$\begin{aligned}
\underline{D}(B) &= \underline{D}(B_1) * p_{j1} + \sum_{k=1}^{\infty} [\underline{D}(B_1) + k(\underline{D}(B_2) + \underline{D}(B_1))] * p_{j2}^k p_{j1}, \\
&= \underline{D}(B_1) * [\frac{1}{(1 - p_{j2})^2}] p_{j1} + \underline{D}(B_2) \frac{p_{j2}}{(1 - p_{j2})^2} p_{j1}, \\
\overline{D}(B) &= \overline{D}(B_1) * p_{j1} + \sum_{k=1}^{\infty} [\overline{D}(B_1) + k(\overline{D}(B_2) + \overline{D}(B_1))] * p_{j2}^k p_{j1}, \\
&= \overline{D}(B_1) * [\frac{1}{(1 - p_{j2})^2}] p_{j1} + \overline{D}(B_2) \frac{p_{j2}}{(1 - p_{j2})^2} p_{j1} \\
\Delta(B) &= \xi(\pi_{D(B_1)}, \pi_{D(B_2)}, \pi_{N_j}),
\end{aligned} \tag{6}$$

where
$P(N_j = k) = p_{j2}^k p_{j1}$,

$$\begin{aligned}
\xi(\pi_{D(B_1)}, \pi_{D(B_2)}, \pi_{N_j}) = {} & \Delta(B_1) S p_{j1} + \Delta(B_2) p_{j2} S p_{j1} + \overline{D}(B_1)^2 p_{j2} T p_{j1} + \\
& + \overline{D}(B_2)^2 p_{j2}^2 T p_{j1} + 2 * \overline{D}(B_1) \overline{D}(B_2) p_{j2} T p_{j1},
\end{aligned}$$

$S = (1 - p_{j2})^{-2}$, $T = 2 * (1 - p_{j2})^{-3}$,
π_x is the distribution of random variable x.

Proposition 1. *For a given workflow diagram W, for all blocks B in W, it must be true that*

$$\overline{D}(B) \geq ED(B) = E(x(B) - s(B)) \geq \underline{D}(B). \tag{7}$$

Especially it is true for TOP

$$\overline{D}(TOP) \geq ED(TOP) = E(x(END) - s(START)) = E(x(END)) \geq \underline{D}(TOP). \tag{8}$$

Proof. See appendix.

One application of the bounds is to give theoretical guarantee for the probability of the duration of the execution of the workflow diagram being larger than a given up limit z as the following corollary.

Corollary 1. *For all given positive real number z, it holds that*

$$P(D(B) > \overline{D}(B) + z) \leq \frac{\Delta(B) - (\underline{D}(B))^2}{z^2}. \tag{9}$$

Proof. Since $\overline{D}(B) \geq E(D(B))$, we have $P(D(B) > \overline{D}(B) + z) \leq P(|D(B) - E(D(B))| > z)$. It then follows from the Chebyshev Inequality (see e.g.[7])

$$P(|D(B) - E(D(B))| > z) < \frac{Var(D(B))}{z^2} \tag{10}$$

that

$$P(D(B) > \overline{D}(B) + z) \leq \frac{Var(D(B))}{z^2}. \tag{11}$$

The desired result follows from the fact that $Var(D(B)) \leq \Delta(B) - (\underline{D}(B))^2$.

4 An Example

In this section, we illustrate the usage of our main results by an example shown in Figure 4. Assume that all the four task blocks B_1, B_2, B_3 and B_4 have the same Gamma distribution with parameter $a = 0.1$ and $\gamma = 0.1$. The pdf of the Gamma distribution is $\frac{a^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-ax}$. Let the fork point of the loop F_1 have parameters $p_{11} = 0.8$ and $p_{12} = 0.2$. $\underline{D}(B_i) = \overline{D}(B_i) = 1$. $\Delta(B_i) = 1.1$. As a building block, the OR block B_{12} has $\underline{D}(B_{12}) = 0.37, \overline{D}(B_{12}) = 1, \Delta(B_{12}) = 2.58$. With this, we can calculate for the AND block B_{123} as $\underline{D}(B_{123}) = 1, \overline{D}(B_{123}) = 2.88, \Delta(B_{123}) = 9.46$. Furthermore, we can establish for the entire workflow diagram as a LOOP block B_{1234}: $\underline{D}(B_{1234}) = 1.5, \overline{D}(B_{1234}) = 3.85, \Delta(B_{1234}) = 21$. Simulation shows that the mean value of $D(B_{1234}) \approx 1.6$ with variance about 1.46. Clearly, the mean value falls in the interval $[\underline{D}(B_{1234}), \overline{D}(B_{1234})] = [1.5, 3.85]$. Using Corollary 1, we can further establish probability guarantee on $D(B_{1234})$. For example, we have $P(D(B_{1234}) > 20) \leq \frac{21-1.5^2}{(20-3.85)^2} = 7\%$.

It is interesting to note that our results apply also for the workflow systems studied in [2] and [3]. For the example in [2], we have $\underline{D}(B) = 5 + 6 + 0.9 * (\max(7, 7 + 7) + 2) + 0.1 * 4 = 25.8$ in contrast to the simulation 28.678 and analysis result 29.2 reported in [2]. For the example in Fig. 9(m) in [3], we have $\underline{D}(B) = \max(((2+6+4+1)*0.2+(6+8))/0.8, 5)+1+1+4 = 26.75$ in contrast to the simulation 28.8628 and analysis result 29.72 reported in [3].

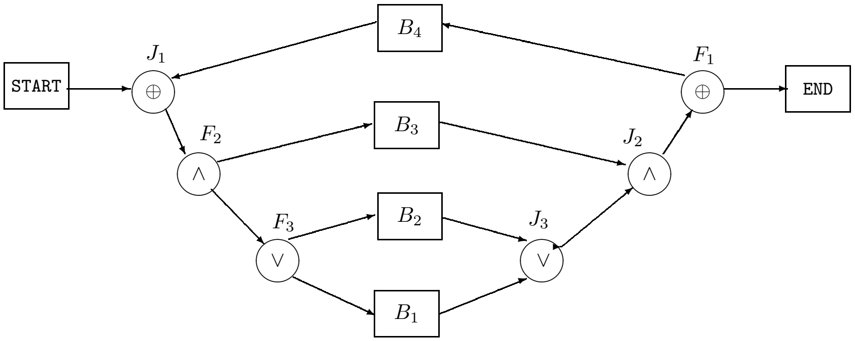


Fig. 4. An example of workflow diagram

5 Conclusions

In this paper, we establish upper and lower bounds for the average turnaround time and upper bound for the second order moment for the turnaround time for a class of workflow diagrams under the assumption of infinite servers. They hold for general task processing time distribution with finite first two order moments. Future work is to consider waiting when there are only limited number of servers for each tasks.

References

1. van der Aalst, W., ter Hofstede, A., Kiepuszewski, B., Barros, A.: Workflow patterns. Distributed and Parallel Databases **14** (2003) 5-51
2. Lin, C., Qu, Y., Ren, F., Marinescu, D.C.: Performance equivalent analysis of workflow systems based on stochastic petri net models. In: Engineering and Deployment of Cooperative Information Systems, First International Conference (EDCIS 2002) Beijing, China, September 17-20, 2002 / Y. Han, S. Tai, D.Wikarski (Eds.), Springer Verlag, LNCS 2480 (2002) 1-64
3. Li, J., Fan, Y., Zhou, M.: Performance modeling and analysis of workflow. Ieee Transactions On Systems, Man, And Cyberneticspart A: Systems And Humans **34** (2004) 229-242
4. Zerguini, L.: On the estimation of the response time of the business process. In: 17th UK Performance Engineering Workshop, University of Leeds. (2001)
5. Baccelli, F., Cohen, G., Olsder, G.J., Quadrat, J.P.: Synchronization and Linearity. John Wiley and Sons (1992)
6. Zhao, Q., Zheng, D.Z., Zhu, X.: Structure properties of min-max systems and existence of global cycle time. IEEE Trans. Automat. Contr. **46** (2001) 148-151
7. Fristedt, B., Gray, L.: A Mordern Approach to Probability Theory. Birkhäuser, Boston (1997)

Appendix

Before presenting the proof of Proposition 1, we need several lemmas. Among them, Lemma 2, 1 and 4 are for the first order moment results; Lemma 3, 5, 6 and 7 are for the second order moment results.

Lemma 1. *For independent non-negative random variables x_u, $u = 1, \ldots, k$, it is true that*

$$E(\max_{u=1\ldots,k} x_u) \leq \max_{u=1\ldots,k} Ex_u + \sum_{u=1}^{k} \sqrt{Var(x_u)} \tag{12}$$

Proof. It follows from $z \leq Ez + |z - Ez|$ for any random variable z and

$$E|x_u - Ex_u| \leq \sqrt{E(|x_u - Ex_u|^2)} = \sqrt{Var(x_u)}. \tag{13}$$

For a random variable x, denote $\phi'(x) = \sqrt{(Ex^2 - (Ex)^2) \times (Ex^2 + 3(Ex)^2)} + Ex^2 - (Ex)^2$. It is easy to show the following lemmas.

Lemma 2.

$$E|x^2 - Ex^2| \leq \phi'(x) \tag{14}$$

Lemma 3. *For independent non-negative random variables x_u, $u = 1, \ldots, k$, it is true that*

$$E(\max_{u=1\ldots,k} x_u)^2 \leq \max_{u=1\ldots,k} E(x_u)^2 + \sum_{u=1}^{k} \phi(x_u) \tag{15}$$

Lemma 4. *For random variables x_u, $u = 1, \ldots, k$, it is true that*

$$E \min_{u=1,\ldots,k} x_u \leq \min_{u=1,\ldots,k} Ex_u \tag{16}$$

Lemma 5. *For independent random variables x_u, $u = 1, \ldots, k$, it is true that*

$$E(\sum_{u=1}^{k} x_u)^2 = \sum_{u=1}^{k} Ex_u^2 + \sum_{u \neq v} Ex_u Ex_v \tag{17}$$

Based on the concept of conditional expectation, we have

Lemma 6. *For random variables x_u, $u = 1, \ldots, k$, and a random variable N taking value from $\{1, \ldots, k\}$ which is independent of x_u, it is true that*

$$E(x_N)^2 = \sum_{u=1,\ldots,k} E(x_u)^2 * p(N = u) \tag{18}$$

For independent random variables x, y, N, let

$$\xi'(\pi_x, \pi_y, \pi_N) = \sum_{k=0}^{\infty} [(k+1)Ex^2 + kEy^2 + (k+1)k * (Ex)^2 + \\ + k(k-1) * (Ey)^2 + 2 * (k+1)k * ExEy] * p(N = k),$$

where π_x,π_y and π_N are distributions of x, y and N.

Lemma 7. *For i.i.d. random variables x_u, y_u, $u = 1, 2, \ldots,$, and a random variable N taking value from $1, 2, \ldots$, which is independent of x_u and y_u, it is true that*

$$E(x_1 + \sum_{u=1}^{N}(x_{u+1} + y_u))^2 = E(E((x_1 + \sum_{u=1}^{k}(x_{u+1} + yu))^2 | N = k)) \quad (19)$$

$$= \xi'(\pi_x, \pi_y, \pi_N) \quad (20)$$

[**Proof of Proposition 1**]
First of all, it is true from

$$x(B) = s(B) + t_i$$

that for a single task block $B = T_i$,

$$E(x(B) - s(B)) = E(t_i) = a_i. \quad (21)$$

Thus, (7) is true for the trivial case of task blocks. Now let us use induction on recursive constructions of blocks to prove the results. Assume that (7) is true for all immediate building blocks of a given block B. Without loss of generality, we assume all blocks contain only two sub-blocks or branches. The proof will be divided into five cases.

1) If B consists of connected blocks B_1 and B_2, we have $D(B) = D(B_1) + D(B_2)$. We have $E(D(B)) = E(D(B_1)) + E(D(B_2)) \geq \underline{D}(B_1) + \underline{D}(B_2)$ $= \underline{D}(B)$, and $ED(B) \leq \overline{D}(B_1) + \overline{D}(B_2) = \overline{D}(B)$, since (7) is assumed to be true for B_1 and B_2.

2) If B is an AND block consisting of two branches B_1 and B_2, we have $D(B) = \max(D(B_1), D(B_2))$. Thus, by noting that $max(x, y) = -min(-x, -y)$, it follows from Lemma 4 that

$$E(D(B)) \geq \max(E(D(B_1)), E(D(B_2))) \geq \max(\underline{D}(B_1), \underline{D}(B_2)). \quad (22)$$

Then we have

$$ED(B) \leq \max(ED(B_1), ED(B_2)) + E\max(|D(B_1) - ED(B_1)|, |D(B_2) - ED(B_2)|)$$

$$\leq \max(\overline{D}(B_1), \overline{D}(B_2)) + E(|D(B_1) - ED(B_1)| + |D(B_2) - ED(B_2)|) \leq \overline{D}(B).$$

The last inequality follows from Lemma 1 and Lemma 2 by observing $\phi(D(B_u)) \geq \phi'(D(B_u))$.

3) If B is an XOR block consisting two branches B_1 and B_2 with F_j as fork point, we have $D(B) = D(B_1) * p_{j1} + D(B_2) * p_{j2}$. It follows that $ED(B) = ED(B_1) * p_{j1} + ED(B_2) * p_{j2} \geq \underline{D}(B_1) * p_{j1} + \underline{D}(B_2) * p_{j2} = \underline{D}(B)$. and $ED(B) = ED(B_1) * p_{j1} + ED(B_2) * p_{j2} \leq \overline{D}(B_1) * p_{j1} + \overline{D}(B_2) * p_{j2} = \overline{D}(B)$. The equality has been proved before as Theorem 3 in [2].
4) If B is an OR block consisting of two branches B_1 and B_2 with F_j as fork point, we have $D(B) = \min(D(B_1), D(B_2))$. Similar to the case of AND block, we have
$E(D(B)) = E \min(D(B_1), D(B_2)) \leq \min(E(D(B_1)), E(D(B_2)))$
$\leq \min(\overline{D}(B_1)), \overline{D}(B_2))) = \overline{D}(B)$ and
$E(D(B)) \geq \min(E(D(B_1)), E(D(B_2))) - E \max(|D(B_1) - ED(B_1)|, |D(B_2) - ED(B_2)|) \geq \min(\underline{D}(B_1)), \underline{D}(B_1)) - E \max(|D(B_1) - ED(B_1)|, |D(B_2) - ED(B_2)|)$
$= \underline{D}(B)$.
5) If B is a LOOP block consisting forward branch B_1 and backward branch B_2 with F_j as its fork point. We have
$E(D(B)) = ED(B_1) * p_{j1} + \sum_{k=1}^{\infty} [ED(B_1) + k * (ED(B_2) + ED(B_2))] * p_{j2}^k$
$\geq \underline{D}(B_1) * p_{j1} + \sum_{k=1}^{\infty} [\underline{D}(B_1) + k * (\underline{D}(B_2) + \underline{D}(B_2))] * p_{j2}^k = \underline{D}(B)$ and
$E(D(B)) = ED(B_1) * p_{j1} + \sum_{k=1}^{\infty} [ED(B_1) + k * (ED(B_2) + ED(B_2))] * p_{j2}^k$
$\leq \overline{D}(B_1) * p_{j1} + \sum_{k=1}^{\infty} [\overline{D}(B_1) + k * (\overline{D}(B_2) + \overline{D}(B_2))] * p_{j2}^k = \overline{D}(B)$. The first equality has been proved before as Theorem 4 in [2].
Using the induction method, we complete the proof by combining cases 1)-5) and Lemma 3, 5, 6 and 7 for the second order estimation.

A Hybrid Quantum-Inspired Genetic Algorithm for Flow Shop Scheduling

Ling Wang[1], Hao Wu[1], Fang Tang[2], and Da-Zhong Zheng[1]

[1] Department of Automation, Tsinghua University, Beijing 100084, China
wangling@mail.tsinghua.edu.cn
[2] Dept. of Physics, Beijing Univ. of Aeronautics and Astronautics, Beijing, 100083, China
tangfang@buaa.edu.cn

Abstract. This paper is the first to propose a hybrid quantum-inspired genetic algorithm (HQGA) for flow shop scheduling problems. In the HQGA, Q-bit based representation is employed for exploration in discrete 0-1 hyperspace by using updating operator of quantum gate as well as genetic operators of Q-bit. Then, the Q-bit representation is converted to random key representation. Furthermore, job permutation is formed according to the random key to construct scheduling solution. Moreover, as a supplementary search, a permutation-based genetic algorithm is applied after the solutions are constructed. The HQGA can be viewed as a fusion of micro-space based search (Q-bit based search) and macro-space based search (permutation based search). Simulation results and comparisons based on benchmarks demonstrate the effectiveness of the HQGA. The search quality of HQGA is much better than that of the pure classic GA, pure QGA and famous NEH heuristic.

1 Introduction

Flow shop scheduling is a typical combinatorial optimization problem that has been proved to be strongly NP-complete [1]. Due to its strong engineering background, flow shop scheduling problem has gained much attention and wide research in both Computer Science and Operation Research fields. The permutation flow shop scheduling with J jobs and M machines is commonly defined as follows. Each of J jobs is to be sequentially processed on machine 1, …, M . The processing time $p_{i,j}$ of job i on machine j is given. At any time, each machine can process at most one job and each job can be processed on at most one machine. The sequence in which the jobs are to be processed is the same for each machine. The objective is to find a permutation of jobs to minimize the maximum completion time, i.e. makespan $C_{\max}$ [1-8]. Due to its significance in both theory and applications, it is always an important and valuable study to develop effective scheduling approaches.

Denote $c_{i,j}$ as the complete time of job i on machine j , and let $\pi = (\sigma_1, \sigma_2, \ldots, \sigma_J)$ be any a processing sequence of all jobs. Then the mathematical formulation of the permutation flow shop problem to minimize makespan can be described as follows:

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 636 – 644, 2005.

$$\begin{cases} c_{\sigma_1,1} = p_{\sigma_1,1}, \\ c_{\sigma_j,1} = c_{\sigma_{j-1},1} + p_{\sigma_1,1}, \quad j = 2,..., J \\ c_{\sigma_1,i} = c_{\sigma_1,i-1} + p_{\sigma_1,i}, \quad i = 2,..., M \\ c_{\sigma_j,i} = \max\{c_{\sigma_{j-1},i}, c_{\sigma_j,i-1}\} + p_{\sigma_j,i}, i = 2,..., M; j = 2,..., J \\ C_{\max} = c_{\sigma_J,M} \end{cases} \tag{1}$$

The optimal solution π^* should satisfies the following criterion:

$$\pi^* = \arg\{C_{\max}(\pi) \to \min\} \tag{2}$$

So far, many approaches have been proposed for flow shop scheduling. However, exact techniques are applicable only to small-sized problems in practice, and the qualities of constructive heuristics [2] are often not satisfactory. So, intelligent methods have gained wide research, such as simulated annealing [3], genetic algorithm [4], evolutionary programming [5], tabu search [6] and hybrid heuristics [7], etc.

Recently, Han and Kim [8-10] proposed some quantum-inspired genetic algorithms (QGAs) for knapsack problem and achieved good results. However, solution of flow shop scheduling should be a permutation of all jobs, while in knapsack problem solution is a 0-1 matrix. That is to say, the QGA cannot directly apply to scheduling problems. Moreover, the QGA only performs exploration in 0-1 hyperspace, i.e. micro-space, while classic permutation genetic algorithm applies permutation-based exploration directly in solution space. Besides, the Q-bit based quantum search and evolutionary genetic search need to be well coordinated, and the exploration and exploitation behaviors need to be well balanced. Thus, in this paper, we propose a GA inspired by quantum computing for flow shop scheduling problem to minimize makespan, especially to develop a hybrid strategy by combining Q-bit based search and permutation based search to achieve better performance.

The organization of the remaining content is as follows. In Section 2, the quantum-inspired GA is presented. Then, the hybrid quantum-inspired GA for permutation flow shop scheduling problem as well as it implementation are proposed in Section 3. The simulation results and comparisons are provided in Section 4. Finally, we end the paper with some conclusions in Section 5.

2 Quantum-Inspired Genetic Algorithm

2.1 Representation

In QGA for a minimization problem, a Q-bit chromosome representation is adopted based on the concept and principles of quantum computing [8-10]. The characteristic of the representation is that any linear superposition of solutions can be represented. The smallest unit of information stored in two-state quantum computer is called a Q-bit, which may be in the "1" state, or in the "0" state, or in any superposition of the two. The state of a Q-bit can be represented as follows:

$$|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{3}$$

where α and β are complex numbers that specify the probability amplitudes of the corresponding states. $|\alpha|^2$ and $|\beta|^2$ denote the probabilities that the Q-bit will be found in the "0" state and "1" state respectively. Normalization of the state to the unity guarantees $|\alpha|^2 + |\beta|^2 = 1$.

A Q-bit individual as a string of m Q-bits is defined as follows:

$$\left[\begin{array}{c|c|c|c} \alpha_1 & \alpha_2 & \cdots & \alpha_m \\ \beta_1 & \beta_2 & \cdots & \beta_m \end{array}\right] \tag{4}$$

where $|\alpha_i|^2 + |\beta_i|^2 = 1$, $i = 1,2,\ldots,m$.

For example, for a following three-Q-bit with three pairs of amplitudes,

$$\left[\begin{array}{c|c|c} 1/\sqrt{2} & 1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} & \sqrt{3}/2 \end{array}\right] \tag{5}$$

the states can be represented as follows:

$$\frac{1}{4}|000\rangle + \frac{\sqrt{3}}{4}|001\rangle - \frac{1}{4}|010\rangle - \frac{\sqrt{3}}{4}|011\rangle + \frac{1}{4}|100\rangle + \frac{\sqrt{3}}{4}|101\rangle - \frac{1}{4}|110\rangle - \frac{\sqrt{3}}{4}|111\rangle \tag{6}$$

This means that the probabilities to represent the states $|000\rangle$, $|001\rangle$, $|010\rangle$, $|011\rangle$, $|100\rangle$, $|101\rangle$, $|110\rangle$ and $|111\rangle$ are 1/16, 3/16, 1/16, 3/16, 1/16, 3/16, 1/16 and 3/16, respectively. By consequence, the above three-Q-bit system contains the information of eight states. Evolutionary computing with Q-bit representation has a better characteristic of population diversity than other representation, since it can represent linear superposition of state's probabilities.

2.2 Operations

In this paper, selection, crossover and mutation operators in QGA are designed as follows.

Selection. All individuals of the population are firstly ordered from the best to the worst, then the top $N/5$ individuals are copied and the bottom $N/5$ individuals are discarded to maintain the size of population, N. Thus, good individuals have more chance to be reserved or to perform evolution.

One Point Crossover. One position is randomly determined (e.g. position i), and then the Q-bits of the parents before position i are reserved while the Q-bits after position i are exchanged.

Mutation. One position is randomly determined (e.g. position i), and then the corresponding α_i and β_i are exchanged.

Catastrophe Operation. To avoid premature convergence, a *catastrophe operation* is used in QGA: if the best solution does not change in certain consecutive generations, we regard it is trapped in local optima, then the best solution is reserved and the others will be replaced by solutions randomly generated.

Rotation operation. In addition, a rotation gate $U(\theta)$ is employed in QGA to update a Q-bit individual as a variation operator [8-10]. (α_i, β_i) of the i-th Q-bit is updated as follows:

$$\begin{bmatrix} \alpha_i^{'} \\ \beta_i^{'} \end{bmatrix} = U(\theta_i) \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{bmatrix} \cdot \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \tag{7}$$

where θ_i is rotation angle. $\theta_i = s(\alpha_i, \beta_i)\Delta\theta_i$, $s(\alpha_i, \beta_i)$ is the sign of θ_i that determines the direction, $\Delta\theta_i$ is the magnitude of rotation angle whose lookup table is shown in Table 1. In the Table, b_i and r_i are the i-th bits of the best solution $\boldsymbol{b}$ and the binary solution $\boldsymbol{r}$ respectively.

Table 1. Lookup table of rotation angle

r_i	b_i	$f(\boldsymbol{r}) < f(\boldsymbol{b})$	$\Delta\theta_i$	$s(\alpha_i, \beta_i)$			
				$\alpha_i\beta_i > 0$	$\alpha_i\beta_i < 0$	$\alpha_i = 0$	$\beta_i = 0$
0	0	false	0	0	0	0	0
0	0	true	0	0	0	0	0
0	1	false	0	0	0	0	0
0	1	true	0.05π	-1	+1	±1	0
1	0	false	0.01π	-1	+1	±1	0
1	0	true	0.025π	+1	-1	0	±1
1	1	false	0.005π	+1	-1	0	±1
1	1	true	0.025π	+1	-1	0	±1

Evaluation. When evaluating the solution, a binary string $\boldsymbol{r}$ with length m is firstly constructed according to the probability amplitudes of individual p with Q-bit representation, then transform the binary string $\boldsymbol{r}$ to the problem solution (e.g. permutation for scheduling problems) for evaluation. In the first step, for $i = 1,2,...,m$, firstly generate a random number η between [0, 1], if α_i of individual p satisfies $|\alpha_i|^2 > \eta$, then set r_i as 1, otherwise set it as 0.

2.3 Procedure of QGA

The procedure of QGA is summarized as follows.

Step 1: randomly generate an initial population $P_Q(t) = \{p_1^t, \cdots, p_N^t\}$, where p_j^t denotes the j-th individual in the t-th generation with the Q-bit representation $p_j^t = \left[\begin{array}{c|c|c|c} \alpha_1^t & \alpha_2^t & \cdots & \alpha_m^t \\ \beta_1^t & \beta_2^t & \cdots & \beta_m^t \end{array}\right]$.

Step 2: evaluate each solution of $P_Q(t)$, and then record the best one denoted by $\boldsymbol{b}$.

Step 3: if stopping condition is satisfied, then output the best result; otherwise go on following steps.

Step 4: perform selection and quantum crossover, mutation for $P_Q(t)$ to generate $P'_Q(t)$.

Step 5: if catastrophe condition is satisfied, perform catastrophe for $P'_Q(t)$ to generate $P_Q(t+1)$ and go to Step 7; otherwise go to Step 6.

Step 6: applying rotation gate $U(\theta)$ to update $P'_Q(t)$ to generate $P_Q(t+1)$.

Step 7: evaluate every individual of $P_Q(t+1)$, and update the best solution $\boldsymbol{b}$ if possible. Then let $t=t+1$ and go back to step 3.

3 Hybrid QGA for Flow Shop Scheduling

3.1 Representation

In flow shop scheduling, the problem solution is a permutation of all jobs. So, it should convert Q-bit representation to permutation for evaluation. In this paper, Q-bit representation is first convert to binary representation as described in Section 2; then the binary representation is viewed as random key representation [11]; finally, job permutation is constructed based on random key.

For example, consider a 3-job, 3-machine problem and let 3 Q-bits be used to represent a job. Suppose a binary representation is [0 1 1| 1 0 1| 1 0 1] that is converted from Q-bit representation, then the random key representation is [3 5 5]. If two random key values are different, we let smaller random key denote the job with smaller number; otherwise, we let the one first appears denote the job with smaller number. So, the above random key representation is corresponding to job permutation [1 2 3].

Obviously, if enough Q-bits are used to represent a job, any job permutation would be constructed with the above strategy from binary representation based space.

3.2 Permutation-Based GA

Although quantum search has high parallel property, Q-bit representation should be converted to permutation solution for evaluation. As we know, permutation encoding is the most widely used in GA for scheduling problems due to its simplicity and easy implementation [12]. To improve the performance of QGA, permutation-based GA (PGA) is also applied as a complementary search for QGA.

The implementation of PGA is briefly described as follow.

Representation. Job permutation is used as representation in PGA. For example for a 3-job flow shop problem, job permutation [2 3 1] denotes the processing sequence is job-2 first, then job-3, and job-1 last. Obviously, for such an encoding scheme, any TSP-based genetic operators can be applied.

Rank-based Selection. The widely used proportional selection operator of GA is based on fitness value so that one needs to determine a transfer from objective value to fitness value. Here, the rank-based selection [12] is strongly suggested to avoid the transfer. That is, all solutions in the population are ordered from the smallest objective value to the largest one, and let $2^{N-i}/(2^N-1)$ be the selected-probability of the i th solution. And the crossover operator is only performed for these selected solutions.

Crossover. The crossover operator used here is partial mapping crossover (PMX) which may be the most popular one for operating the permutation [12]. In the step of PMX, firstly two crossover points are chosen and the sub-sections of the parents between the two points are exchanged, then the chromosomes are filled up by partial map. For example, for a 9-job flow shop problem, let 3 and 7 be two crossover points for parents (2 6 4 7 3 5 8 9 1) and (4 5 2 1 8 7 6 9 3). Then, the children will be (2 3 4| 1 8 7 6| 9 5) and (4 1 2| 7 3 5 8| 9 6). It has been demonstrated that PMX satisfies the fundamental properties enunciated by Holland, namely that it behaves in such a way that the best schemata reproduce themselves better than the others. Moreover, in order to generate N new temporary solutions for the next mutation operator, it needs to perform such rank-based selection and crossover $N/2$ times.

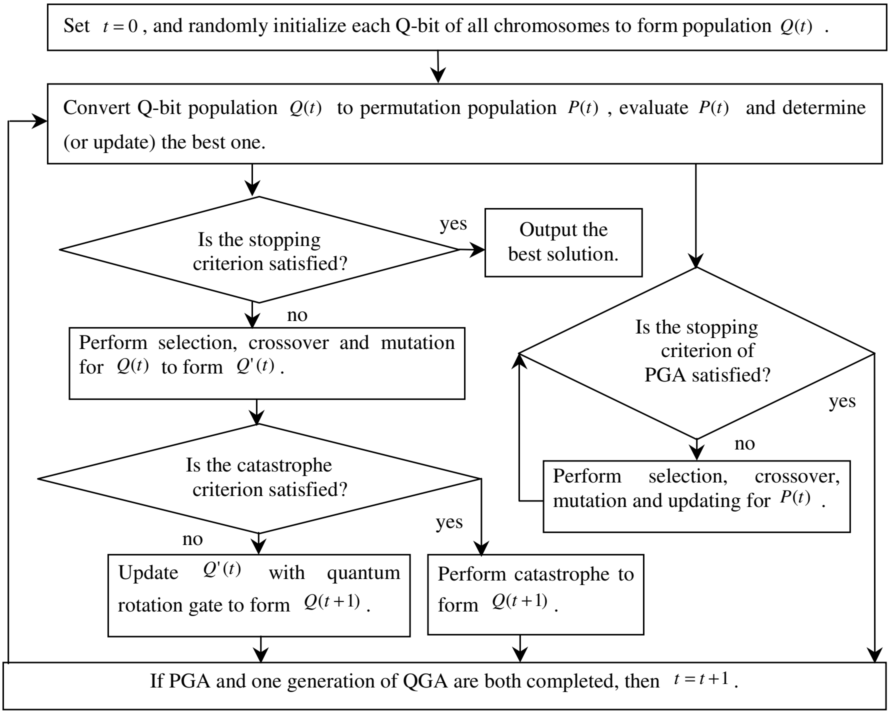

Set $t=0$, and randomly initialize each Q-bit of all chromosomes to form population $Q(t)$.
Convert Q-bit population $Q(t)$ to permutation population $P(t)$, evaluate $P(t)$ and determine (or update) the best one.
Is the stopping criterion satisfied?
yes
Output the best solution.
no
Perform selection, crossover and mutation for $Q(t)$ to form $Q'(t)$.
Is the catastrophe criterion satisfied?
no
yes
Update $Q'(t)$ with quantum rotation gate to form $Q(t+1)$.
Perform catastrophe to form $Q(t+1)$.
Is the stopping criterion of PGA satisfied?
yes
no
Perform selection, crossover, mutation and updating for $P(t)$.
If PGA and one generation of QGA are both completed, then $t=t+1$.

Fig. 1. The framework of HQGA

Mutation. Mutation serves to maintain diversity in population. In this paper, SWAP is used in PGA, i.e., two distinct elements are randomly selected and swapped.

Population Updating. In this paper, ($\mu+\lambda$) strategy [4] is used as updating strategy. That is, the top N solutions among the old population and the new population are selected as the population for the nest generation.

3.3 Hybrid QGA for Flow Shop Scheduling

To achieve good results for solving flow shop problems, we combine the QGA and PGA explained above to develop a hybrid QGA (HQGA), whose framework is illustrated in Fig. 1.

From Fig. 1, it can be seen that the main searching process is Q-bit based QGA, which sequentially performs selection, crossover, mutation and rotation operations for Q-bit based population and applies catastrophe operation to avoid being trapped in local optima. Moreover, permutation population performs classic GA as a supplement search. In brief, the HQGA can be viewed as a fusion of micro-space based search (Q-bit based search) and macro-space based search (permutation based search), thus searching behavior can be enriched, and exploration and exploitation abilities can be well balanced and enhanced.

4 Simulations and Comparisons

4.1 Testing Benchmarks

To test the performance of the proposed HQGA, computational simulation is carried out with some well-studied benchmarks.

In this paper, 29 problems that were contributed to the OR-Library are selected. The first eight problems were called car1, car2 through car8 by Carlier [13]. The other 21 problems were called rec01, rec03 through rec41 by Reeves [14], who used them to compare the performances of simulated annealing, genetic algorithm and neighborhood search and found these problems to be particularly difficult. Thus far these problems have been used as benchmarks for study with different methods by many researchers.

4.2 Simulation Results and Comparisons

We set population size as 40, maximum generation (stopping condition of QGA) as $J\times M$, the length of each chromosome as $10\times J$ (i.e., every 10 Q-bits correspond to a job), crossover probability as 1, mutation probability as 0.05, catastrophe happens in QGA if the best solution does not change in consecutive $J\times M/10$ generations.

Besides, PGA is performed 10 generations (stopping condition of PGA) in every generation of QGA. For fair comparison, pure PGA and pure QGA use the same computational effort as HQGA.

We run each algorithm 20 times for every problem, and the statistical results are summarized in Table 2, where BRE, ARE and WRE denote the best, average and worst relative errors with C^* (lower bound or optimal makespan) respectively.

Table 2. The statistical results of testing algorithms

P	*J, M*	*C**	HQGA			NEH	PGA		QGA	
			BRE	ARE	WRE	RE	BRE	ARE	BRE	ARE
Car1	11,5	7038	0	0	0	0	0	0	0	0
Car2	13,4	7166	0	0	0	2.93	0	0.21	0	1.90
Car3	12,5	7312	0	0	0	1.19	0	0.86	1.19	1.65
Car4	14,4	8003	0	0	0	0	0	0.08	0	0.06
Car5	10,6	7720	0	0	0	1.49	0	0.09	0	0.11
Car6	8,9	8505	0	0	0	3.15	0	0.26	0	0.19
Car7	7,7	6590	0	0	0	0	0	0	0	0
Car8	8,8	8366	0	0	0	2.37	0	0.18	0	0.03
Rec01	20,5	1247	0	0.14	0.16	4.49	1.28	4.45	2.81	6.79
Rec03	20,5	1109	0	0.17	0.72	2.07	0.72	2.18	0.45	3.87
Rec05	20,5	1242	0.24	0.34	2.17	3.14	0.24	1.50	2.25	3.00
Rec07	20,10	1566	0	1.02	2.11	3.83	1.85	2.98	1.15	4.67
Rec09	20,10	1537	0	0.64	2.41	2.99	2.60	4.12	4.03	6.40
Rec11	20,10	1431	0	0.67	2.66	8.32	4.05	6.17	6.08	8.79
Rec13	20,15	1930	0.16	1.07	3.06	3.73	3.01	4.66	5.08	7.98
Rec15	20,15	1950	0.05	0.97	1.64	3.23	2.21	3.96	3.49	5.93
Rec17	20,15	1902	0.63	1.68	2.73	6.15	3.26	5.86	6.51	9.10
Rec19	30,10	2093	0.29	1.43	2.58	4.40	5.88	7.39	7.98	9.80
Rec21	30,10	2017	1.44	1.63	1.83	5.65	5.16	7.05	6.94	10.05
Rec23	30,10	2011	0.50	1.20	2.64	7.16	5.92	7.56	9.10	10.55
Rec25	30,15	2513	0.77	1.87	3.34	5.21	6.29	7.85	7.16	10.06
Rec27	30,15	2373	0.97	1.83	2.82	5.27	6.41	8.14	7.63	11.05
Rec29	30,15	2287	0.35	1.97	3.63	4.55	8.53	10.53	12.42	14.06
Rec31	50,10	3045	1.05	2.50	4.07	4.20	8.83	10.16	9.82	12.68
Rec33	50,10	3114	0.83	0.91	1.19	4.08	5.84	7.07	6.20	9.54
Rec35	50,10	3277	0	0.15	0.34	1.10	2.72	3.88	4.21	6.52
Rec37	75,20	4951	2.52	4.33	5.66	5.58	13.78	15.27	15.54	17.49
Rec39	75,20	5087	1.65	2.71	4.46	4.34	12.04	13.06	13.50	15.49
Rec41	75,20	4960	3.13	4.15	5.73	6.69	13.48	15.86	16.92	18.84

From Table 2, it can be seen that the results obtained by HQGA are much better than not only that of NEH heuristic but also that of PGA and QGA. Even the AVE values resulted by HQGA are better than BRE values resulted by PGA and QGA.

Secondly, the BRE values resulted by HQGA are very near to 0, especially for small-scale problems, which means HQGA is able to obtain good solutions in global sense.

Thirdly, the BRE, ARE and WRE values resulted by HQGA are closer than those by other methods, which means HQGA has good robustness and consistence on initial conditions.

So, it is concluded that HQGA is an effective approach for flow shop scheduling problems.

5 Conclusions

In this paper, we are the first to propose a hybrid QGA for permutation flow shop scheduling problem to minimize makespan. The HQGA can be viewed as a fusion of Q-bit based search and permutation based search. Simulation results demonstrate the effectiveness of the HQGA. The future work is to investigate parameter adaptation and to study QGA for other kinds of scheduling problems, such as job shop [15].

Acknowledgements

This research is partially supported by National Science Foundation of China (60204008, 60374060) and 973 Program (2002CB312200).

References

1. Garey, M.R., Johnson, D.S.: Computers and Intractability: a Guide to the Theory of NP-Completeness. Freeman, San Francisco, (1979)
2. Nawaz, M., Enscore, E.Jr., Ham, I.: A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem. Omega, 11 (1983) 91-95
3. Ogbu, F.A., Smith, D.K.: Simulated annealing for the permutation flowshop problem. Omega, 19 (1990) 64-67
4. Wang, L., Zhang, L., Zheng, D.Z.: A class of order-based genetic algorithm for flow shop scheduling. International Journal of Advanced Manufacture Technology, 22 (2003) 828-835
5. Wang, L., Zheng, D.Z.: A modified evolutionary programming for flow shop scheduling. International Journal of Advanced Manufacturing Technology, 22 (2003) 522-527
6. Nowicki, E., Smutnicki, C.: A fast tabu search algorithm for the permutation flow-shop problem. European J. Operational Research, 91 (1996) 160-175
7. Wang, L., Zheng, D.Z.: An effective hybrid heuristic for flow shop scheduling. International Journal of Advanced Manufacture Technology, 21 (2003) 38-44
8. Han, K.H., Kim, J.H.: Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. IEEE Trans. Evol. Comput., 6 (2002) 580-593
9. Han, K.H., Kim, J.H.: A Quantum-inspired evolutionary algorithms with a new termination criterion, He gate, and two-phase scheme. IEEE Trans. Evol. Comput., 8 (2004) 156-169
10. Han, K.H., Kim, J.H.: Genetic quantum algorithm and its application to combinatorial optimization problem. In: IEEE Proc. of CEC, (2000) 1354-136
11. Bean, J.C.: Genetic algorithms and random keys for sequencing and optimization. ORSA Journal on Computing, 6 (1994) 154-160
12. Wang, L.: Intelligent Optimization with Applications. Tsinghua University & Springer Press, Beijing, (2001)
13. Carlier, J.: Ordonnancements a contraintes disjonctives. R.A.I.R.O. Recherche operationelle/ Operations Research, 12 (1978) 333-351
14. Reeves, C.R.: A genetic algorithm for flowshop sequencing. Computers and Operations Research, 22 (1995) 5-13
15. Wang, L., Zheng, D.Z.: An effective hybrid optimization strategy for job-shop scheduling problems. Computers and Operations Research, 28 (2001) 585-596

Stability and Stabilization of Impulsive Hybrid Dynamical Systems

Guangming Xie, Tianguang Chu, and Long Wang

Center for Systems & Control,
LTCS and Department of Mechanics and Engineering Science,
Peking University, Beijing, 100871, China
xiegming@mech.pku.edu.cn

Abstract. Many practical systems in physics, biology, engineering, and information science exhibit impulsive dynamical behaviors due to abrupt changes at certain instants during the dynamical processes. In this paper, stability analysis and stabilization synthesis problems are investigated for a class of hybrid dynamical systems which consisting of a family of linear constant subsystems and a rule that orchestrates the switching between them. Furthermore, there exist impulses at the switching instants. A switched quadratic Lyapunov function is introduced to check asymptotic stability of such systems. Two equivalent necessary and sufficient conditions for the existence of such a Lyapunov function are established, respectively. The conditions are in linear matrix inequality form and can be used to solve stabilization synthesis problem.

1 Introduction

In recent years, there has been increasing interest in the analysis and synthesis of switched and hybrid systems due to their significance both in theory and applications [1-7]. There have been a lot of studies for switched systems, primarily on stability analysis and synthesis [8-13]. On the other hand, Controllability and observability are the two most fundamental concepts in modern control theory [14-15]. They have close connections to pole assignment, structural decomposition, quadratic optimal control and observer design, etc. Controllability and observability of hybrid systems have been studied by a number of papers[16-25]. [16] first studied the one-period controllability and observability for periodically switched systems, some sufficient and necessary conditions are established. Then [17] introduced the multiple-period controllability and observability concepts naturally extended from the one-period ones, necessary and sufficient criteria are derived. It was also pointed out that controllability can be realized in n periods at most, where n is the state dimension. As to arbitrarily switched linear systems, [18] first gave a sufficient condition and a necessary condition for controllability and proved that the necessary condition is also sufficient for 3-dimensional systems with only two subsystems. Following the work in [18], [19] extended the result to 3-dimensional systems with arbitrary number of subsystems. Then, necessary and sufficient geometric type criteria for controllability

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 645–654, 2005.

and observability are derived in [20] and [21]. The discrete-time counterparts were addressed in [23] and [24]. Furthermore, it was proved that controllability can be realized by a single switching sequence in [21] and [22] , a direct consequence is the criterion given in [20] and [21]. For discrete-time systems, the corresponding result was also derived in [25].

As is well known, many practical systems in physics, biology, engineering, and information science exhibit impulsive dynamical behaviors due to abrupt changes at certain instants during the dynamical processes [26-29]. However, for hybrid and switched systems, as an important model for dealing with complex real systems, there is little work concerning impulsive phenomena.

In [26], impulsive phenomena are introduced into switched systems. necessary and sufficient criteria for controllability and observability of switched impulsive control systems which exhibit impulsive behaviors at switching were established. Moreover, it was shown that the controllability is equivalent to the reachability for such systems under some mild conditions. Similar results were established for observability and determinability of such systems by duality.

In this paper, we focus on the stability analysis and stabilization synthesis for such class of switched impulsive systems. Two equivalent necessary and sufficient conditions for the existence of such a Lyapunov function are established, respectively. The conditions are in linear matrix inequality(LMI) form and can be used to solve stabilization synthesis problem conveniently.

This paper is organized as follows. Section 2 formulates the problem and presents the preliminary results. Section 3 and Section 4 investigate stability and stabilization, respectively. Two numerical examples are given in Section 5. Finally, the conclusion is provided in Section 6.

2 Problem Formulation

Consider the switched impulsive control system given by

$$\begin{cases} \dot{x}(t) = A_{r(t)}x(t) + B_{r(t)}u(t), & \text{if } r(t^+) = r(t) \\ x(t^+) = E_{r(t^+),\, r(t)}x(t^-) + F_{r(t^+),r(t)}u(t^-), & \text{if } r(t^+) \neq r(t) \end{cases} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, the piecewise continuous(p.c.) function $u(t) \in \mathbb{R}^p$ is the input vector; $x(t^+) := \lim_{h \to 0^+} x(t+h)$, $x(t^-) := \lim_{h \to 0^+} x(t-h)$, $x(t^-) = x(t)$, which implies that the solution of the system (1) is left continuous; the left continuous function $r(t) : \mathbb{R}^+ \to \{1, 2, \cdots, N\}$ is the switching signal. Moreover, $r(t) = i$ means that the subsystem (A_i, B_i) is active; $r(t) = i$ and $r(t^+) = j$ mean that the system is switched from the ith subsystem to the jth subsystem at time instant t. At the switching, there exists an impulse described by the second equation of (1). $A_i, B_i, E_{i,j}, F_{i,j}$ are known $n \times n$, $n \times p$, $q \times n$, $q \times p$, $n \times n$ and $n \times p$ constant matrices, $i, j \in \{1, 2, \cdots, N\}$; particularly, $E_{i,i} = I_n$, $F_{i,i} = 0$, $i \in \{1, 2, \cdots, N\}$, which means the switchings between the same subsystem are smooth and without impulse, in other words, there is no impulse when a subsystem is remaining active.

For system (1), the following assumptions are made.

Assumption 1. *The switching signal is arbitrary and not known a priori, but the instantaneous value is available in real time.*

Assumption 2. *The switchings are finite in any finite time interval.*

Here, we are interested in stability analysis and control synthesis problems for such class of switched systems. By stability analysis, we mean stability analysis of the origin for the autonomous switched systems. The control synthesis is related to the design of a switched state feedback control

$$\begin{cases} u(t) = K_{r(t)}x(t), & \text{if } r(t^+) = r(t); \\ u(t^+) = L_{r(t^+),r(t)}x(t^-), & \text{if } r(t^+) \neq r(t). \end{cases} \tag{2}$$

ensuring stability of the closed-loop switched system

$$\begin{cases} \dot{x}(t) = (A_{r(t)} + B_{r(t)}K_{r(t)})x(t), & \text{if } r(t^+) = r(t); \\ x(t^+) = (E_{r(t^+),r(t)} + F_{r(t^+),r(t)}L_{r(t^+),r(t)})x(t^-), & \text{if } r(t^+) \neq r(t). \end{cases} \tag{3}$$

3 Stability Analysis

In this section, we discuss the stability of the origin of the autonomous switched system given by

$$\begin{cases} \dot{x}(t) = A_{r(t)}x(t), & \text{if } r(t^+) = r(t); \\ x(t^+) = E_{r(t^+),\ r(t)}x(t^-), & \text{if } r(t^+) \neq r(t). \end{cases} \tag{4}$$

For system (4), we introduce a switched Lyapunov function defined as

$$V(x(t),t) = x^T(t)P_{r(t)}x(t) \tag{5}$$

with $P_1, \cdots, P_N$ positive definite matrices.

Lemma 1. *If there exists a switched Lyapunov function given by (5) satisfying for any switching signal $r(t)$,*

(i) $\dot{V}(x(t),t) < 0$, if $r(t^+) = r(t)$;

(ii) $V(x(t^+),t^+) \leq V(x(t),t)$, if $r(t^+) \neq r(t)$.

then the system (4) is asymptotically stable under arbitrary switching signals, i.e., given any switching signal $r(t)$, the system trajectory of system (4) along the switching signal $r(t)$ satisfies $\lim_{t\to\infty} x(t) = 0$.

Proof. See Appendix A.

Based on Lemma 1, we present the following LMI-form necessary and sufficient condition for stability of the autonomous system (4).

Theorem 1. *For system (4), the following statements are equivalent:*

(a) there exists a switched Lyapunov function of the form (5) ensuing the asymptotical stability of the system;

(b) there exist positive definite matrices $P_1, \cdots, P_N$ satisfying

$$A_i^T P_i + P_i A_i < 0, i = 1, \cdots, N; \tag{6}$$

$$\begin{bmatrix} P_i & E_{j,i}^T P_j \\ P_j E_{j,i} & P_j \end{bmatrix} \geq 0, i, j = 1, \cdots, N, i \neq j; \tag{7}$$

(c) there exist positive definite matrices $S_1, \cdots, S_N$ satisfying

$$S_i A_i^T + A_i S_i < 0, i = 1, \cdots, N; \tag{8}$$

$$\begin{bmatrix} S_j & E_{j,i} S_i \\ S_i E_{j,i}^T & S_i \end{bmatrix} \geq 0, i, j = 1, \cdots, N, i \neq j; \tag{9}$$

Proof. • To prove a) ⇒ b), suppose that there exists a Lyapunov function of the form (5), whose derivative and difference satisfying

$$\dot{V}(x(t), t) = x^T(t)(A_{r(t)}^T P_{r(t)} + P_{r(t)} A_{r(t)})x(t) < 0, \quad \text{if } r(t^+) = r(t);$$

$$V(x(t^+), t^+) - V(x(t), t) = x^T(t)(E_{r(t^+),r(t)}^T P_{r(t^+)} E_{r(t^+),r(t)} - P_{r(t)})x(t) \leq 0, \quad \text{if } r(t^+) \neq r(t).$$

As this has to be satisfied under arbitrary switching laws, it follows that this has to hold for special configuration $r(t^+) = j$, $r(t) = i$ and for all $x(t) \in \mathbb{R}^n$. Thus, we get (6) and

$$E_{j,i}^T P_j E_{j,i} - P_i \leq 0$$

which is equivalent, by Schur complement, to (7).

• To prove a) ⇒ b), suppose that (6) and (7) are satisfied for all $i, j = 1, \cdots, N, i \neq j$. Then we have

$$E_{j,i}^T P_j E_{j,i} - P_i \leq 0$$

Then, for any switching law $r(t)$, we have

$$\dot{V}(x(t), t) = x^T(t)(A_{r(t)}^T P_{r(t)} + P_{r(t)} A_{r(t)})x(t) < 0, \quad \text{if } r(t^+) = r(t);$$

$$V(x(t^+), t^+) - V(x(t), t) = x^T(t)(E_{r(t^+),r(t)}^T P_{r(t^+)} E_{r(t^+),r(t)} - P_{r(t)})x(t) \leq 0, \quad \text{if } r(t^+) \neq r(t).$$

By Lemma 1, V is the Lyapunov function and the system is stable.

• To prove b) ⇔ c), just let $S_i = P_i^{-1}$ and apply Schur complement.

4 Control Synthesis

In this section, based on the stability analysis result, we established a LMI-form state feedback synthesis condition for stabilization of the system (1).

Theorem 2. *For system (1), there exists a switched state feedback controller (2) ensuing the existence of a switched Lyapunov function (5) which guarantees the asymptotical stability of the closed-loop system (16) if and only if there exist positive definite matrices $S_1, \cdots, S_N$, and matrices $X_1, \cdots, X_N$, $Y_{1,1}, \cdots, Y_{1,N}, \cdots$, $Y_{N,N}$ such that*

$$S_i A_i^T + X_i^T B_i^T + A_i S_i + B_i X_i < 0, i = 1, \cdots, N; \tag{10}$$

$$\begin{bmatrix} S_j & E_{j,i} S_i + F_{j,i} Y_{j,i} \\ S_i E_{j,i}^T + Y_{j,i}^T F_{j,i}^T & S_i \end{bmatrix} \geq 0, i, j = 1, \cdots, N, i \neq j. \tag{11}$$

Furthermore, if (17) and (19) are feasible, then the corresponding switched state feedback matrices can be taken

$$K_i = X_i S_i^{-1}, i = 1, \cdots, N; \tag{12}$$

$$L_{j,i} = Y_{j,i} S_i^{-1}, i, j = 1, \cdots, N, i \neq j. \tag{13}$$

Proof. ($\Leftarrow$) Suppose that there exist S_i, X_i and $Y_{j,i}$ such that (17) and (19) are satisfied. From (21) and (22), we get

$$X_i = K_i S_i, i = 1, \cdots, N;$$

$$Y_{j,i} = L_{j,i} S_i, i, j = 1, \cdots, N, i \neq j.$$

Replacing X_i, $Y_{j,i}$ by $K_i S_i$, $L_{j,i} S_i$, respectively, we get

$$S_i (A_i + B_i K_i)^T + (A_i + B_i K_i) S_i < 0,$$

$$\begin{bmatrix} S_i & S_i (E_{j,i} + F_{j,i} L_{j,i})^T \\ (E_{j,i} + F_{j,i} L_{j,i}) S_i & S_j \end{bmatrix} \geq 0,$$

By Theorem 1, we know that the closed-loop system (16) is asymptotically stable. ($\Leftarrow$) Suppose that there exist switched state feedback controller (2) and switched quadratic Lyapunov function (5) proving the stability of the closed-loop system (16). By Theorem 1, it follows that

$$S_i (A_i + B_i K_i)^T + (A_i + B_i K_i) S_i < 0,$$

$$\begin{bmatrix} S_i & S_i (E_{j,i} + F_{j,i} L_{j,i})^T \\ (E_{j,i} + F_{j,i} L_{j,i}) S_i & S_j \end{bmatrix} \geq 0,$$

by setting $X_i = K_i S_i$, $Y_{j,i} = L_{j,i} S_i$, which is nothing but (17) and (19).

Now, we extend the obtained results for the state feedback case to the switched static output feedback case. Assume that the output of the system (1) is given by

$$y(t) = Cx(t) \tag{14}$$

where C is of full column rank. The switched static output feedback control is to design the following controller

$$\begin{cases} u(t) = R_{r(t)}y(t), & \text{if } r(t^+) = r(t); \\ u(t^+) = W_{r(t^+),r(t)}y(t^-), & \text{if } r(t^+) \neq r(t). \end{cases} \tag{15}$$

ensuring stability of the closed-loop switched system

$$\begin{cases} \dot{x}(t) = (A_{r(t)} + B_{r(t)}R_{r(t)}C)x(t), & \text{if } r(t^+) = r(t); \\ x(t^+) = (E_{r(t^+),r(t)} + F_{r(t^+),r(t)}W_{r(t^+),r(t)}C)x(t^-), & \text{if } r(t^+) \neq r(t). \end{cases} \tag{16}$$

Corollary 1. *For system (1), there exists a switched static output feedback controller (15) ensuing the existence of a switched Lyapunov function (5) which guarantees the asymptotical stability of the closed-loop system (16) if and only if there exist positive definite matrices $S_1, \cdots, S_N$, and matrices $U_1, \cdots, U_N, V_1, \cdots, V_N$, $Q_{1,1}, \cdots, Q_{1,N}, \cdots, Q_{N,N}$, $T_{1,1}, \cdots, T_{1,N}, \cdots, T_{N,N}$ such that*

$$S_iA_i^T + C^TU_i^TB_i^T + A_iS_i + B_iU_iC < 0, i = 1, \cdots, N; \tag{17}$$

$$U_iC = CV_i, i = 1, \cdots, N; \tag{18}$$

$$\begin{bmatrix} S_j & E_{j,i}S_i + F_{j,i}Q_{j,i}C \\ S_iE_{j,i}^T + C^TQ_{j,i}^TF_{j,i}^T & S_i \end{bmatrix} \geq 0, i, j = 1, \cdots, N, i \neq j; \tag{19}$$

$$Q_{i,j}C = CT_{i,j}, i, j = 1, \cdots, N, i \neq j. \tag{20}$$

Furthermore, if (17)- (20) are feasible, then the corresponding switched state feedback matrices can be taken

$$R_i = U_iV_i^{-1}, i = 1, \cdots, N; \tag{21}$$

$$W_{j,i} = Q_{j,i}T_{j,i}^{-1}, i, j = 1, \cdots, N, i \neq j. \tag{22}$$

Proof. It's trivial.

5 Example

In this section, we present two numerical examples to illustrate the utilization of the obtained results.

Example 1. Consider an autonomous switched system (4) with $N = 2$ and

$$\begin{aligned} A_1 &= \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}, A_2 = \begin{bmatrix} -2 & 0 \\ 1 & -1 \end{bmatrix}; \\ E_{1,2} &= \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, E_{2,1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned} \tag{23}$$

By solving the LMIs (6) and (7), we have a solution as follows.

$$P_1 = \begin{bmatrix} 0.9707 & 0.2518 \\ 0.2518 & 0.5776 \end{bmatrix}, P_2 = \begin{bmatrix} 0.5776 & 0.2518 \\ 0.2518 & 0.9707 \end{bmatrix}. \tag{24}$$

As a result, the system is asymptotically stable under arbitrary switching signals.

Example 2. Consider a switched system (1) with $N = 2$ and

$$\begin{aligned} A_1 &= \begin{bmatrix} -2 & 1 \\ 4 & 5 \end{bmatrix}, A_2 = \begin{bmatrix} 5 & 8 \\ 1 & -3 \end{bmatrix}; \\ B_1 &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, B_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}; \\ E_{1,2} &= \begin{bmatrix} 0 & 0 \\ 1 & 0.1 \end{bmatrix}, E_{2,1} = \begin{bmatrix} 1 & 0.1 \\ 0 & 0 \end{bmatrix} \\ F_{12} &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, F_{21} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \end{aligned} \tag{25}$$

By solving the LMIs (17) and (19), we have a solution as follows.

$$\begin{aligned} S_1 &= \begin{bmatrix} 2.1284 & -0.2523 \\ -0.2523 & 2.6997 \end{bmatrix}, S_2 = \begin{bmatrix} 2.3926 & -0.2657 \\ -0.2657 & 3.2403 \end{bmatrix}; \\ X_1 &= [-7.7569 \quad -7.0899], X_2 = [-5.0526 \quad -25.1257]; Y_{12} = Y_{21} = [0 \;\; 0]. \end{aligned} \tag{26}$$

As a result, the system is asymptotically stabilized by (2) with

$$K_1 = X_1 S_1^{-1} = [-4 \quad -3], K_2 = X_2 S_2^{-1} = [-3 \quad -8]$$

and

$$L_{12} = L_{21} = [0 \;\; 0]$$

under arbitrary switching signals.

6 Conclusion

In this paper, stability analysis and stabilization synthesis problems have been investigated for a class of hybrid dynamical systems which consisting of a family of linear constant subsystems and a rule that orchestrates the switching between them. Furthermore, there exist impulses at the switching instants. A switched quadratic Lyapunov function has been introduced to check asymptotic stability of such systems. Two equivalent necessary and sufficient conditions for the existence of such a Lyapunov function have been established, respectively. The conditions are in linear matrix inequality form and can be used to solve stabilization synthesis problem.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 60404001, No. 10372002 and No. 60274001), National Key Basic Research and Development Program (2002CB312200) and National Hi-Tech R&D 863 Project (No. 2002AA755002).

References

1. Williams, S.M., Hoft, R. G.: Adaptive Frequency Domain Control of PPM Switched Power Line Conditioner. IEEE Trans. on Power Electronics. 6 (1991) 665-670
2. Sira-Ramirez, H.: Nonlinear P-I Controller Design for Switch Mode dc-to-dc Power Converters. IEEE Trans. Cir. & Sys. 38 (1991) 410-417
3. Li, Z.G., Wen, C.Y., Soh, Y. C.: Switched Controllers and Their Applications in Bilinear Systems. Automatica 37 (2001) 477-481
4. Ishii, H., Francis, B.A.: Limited Date Rate in Control Systems with Networks. Springer-Verlag, Berlin (2002)
5. Leonessa, A., Hadda, W.M., Chellaboina, V.: Nonlinear System Stabilization Via Hierarchical Switching Control. IEEE Transactions on Automatic Control 46 (2001) 17-28
6. Morse, A.S.: Supervisory Control of Families of Linear Set-Point Controllers, Part 1: Exact matching. IEEE Trans. Auto. Contr. 41 (1996) 1413-1431
7. Narendra, K.S., Balakrishnan, J.: Adaptive Control Using Multiple Models. IEEE Trans. on Auto. Contr. 42 (1997) 171-187
8. Agrachev, A.A., Liberzon, D.: Lie-algebraic Stability Criteria for Switched Systems. SIAM J. on Contr. & Opt. 40 (2002) 253-269
9. Liberzon, D., Morse, A.S.: Basic Problems in Stability and Design of Switched Systems. IEEE Control Systems 19 (1999) 59-70
10. Brannicky, M.S.: Multiple Lyapunov Functions and Other Analysis Tools for Switched and Hybrid Systems. IEEE Trans. Auto. Contr. 43 (1998) 475-482
11. Johansson, M. Rantzer, A.: Computation of Piecewise Quadratic Lyapunov Functions for Hybrid Systems. IEEE Transactions on Automatic Control 43 (1998) 555 -559
12. Dayawansa, W.P., Martin, C.F.: A Converse Lyapunov Theorem for a Class of Dynamical Systems Which Undergo Switching. IEEE Transactions on Automatic Control 44 (1999) 751-760
13. Liberzon, D., Hespanha, J.P., Morse, A.S.: Stability of Switched Systems: a Lie-algebraic Condition. Syst. Contr. L. 37 (1999) 117-122
14. Kailath, T.: Linear Systems. Englewood Cliffs, N. J., Prentice-Hall (1980)

15. Wonham, W.M.: Linear Multivariable Control: A Geometric Approach. Springer-Verlag, New York (1985)
16. Ezzine, J. Hadda, A.H.: Controllability and Observability of Hybrid Systems. Int. J. Control 49 (1989) 2045-2055
17. Xie, G., Zheng, D.: Research on Controllability and Reachability of Hybrid Systems. Proc. of Chinese Contr. Conf. (2000) 114-117
18. Sun, Z., Zheng, D.: On Stabilization of Switched Linear Control Systems. IEEE Transactions on Automatic Control 46 (2001) 291-295
19. Xie, G. Zheng, D., Wang, L.: Controllability of Switched Linear Systems. IEEE Transactions on Automatic Control 47 (2002) 1401 -1405
20. Sun, Z., Ge, S.S., Lee, T.H.: Controllability and Reachability Criteria for Switched Linear Systems. Automatica 38 (2002) 775-786
21. Xie, G., Wang, L.: Necessary and Sufficient Conditions for Controllability of Switched Linear Systems. Proc. of American Control Conference (2002) 1897-1902
22. Xie, G., Wang, L.: Controllability and Stabilizability of Switched Linear-Systems. Systems and Control Letters 48 (2003) 135-155
23. Ge, S.S., Sun, Z., Lee, T.H.: Reachability and Controllability of Switched Linear Discrete-Time Systems. IEEE Transactions on Automatic Control 46 (2001) 1437-1441
24. Ge, S.S., Sun, Z., Lee, T.H.: Reachability and Controllability of Switched Linear Systems. Proceedings of the American Control Conference (2001) 1898-1903
25. Xie, G., Wang, L.: Reachability Realization and Stabilizability of Switched Linear Discrete-time systems. J. Math. Anal. Appl. 280 (2003) 209-220
26. Xie, G., Wang, L.: Necessary and Sufficient Conditions for Controllability and Observability of Switched Impulsive Control Systems. IEEE Transactions on Automatic Control 49 (2004) 960-966
27. Deo, S.G., Pandit, S.G.: Differential Systems Involving impulses. Springer-Verlag, New York (1982)
28. Lakshmikantham, V., Bainov, D.D., Simenov, P.S.: Theory of Impulsive Differential Equations. World Scientific, Singapore (1989)
29. Guan, Z.H., Qian, T.H., Yu, X.: Controllability and Observability of Linear Time-Varying Impulsive Systems. IEEE CAS I. 49 (2002) 1198-1208

Appendix A

Proof (Proof of Lemma 1.). Let $t > t_0$ and $r\left(t^+\right) = r(t) = i$. Then by condition (i),

$$\begin{aligned}\dot{V}\left(t\right) &= x^T\left(P_iA_i + A_i^TP_i\right)x\\ &\le -\sigma V\left(t\right),\end{aligned}$$

where $V\left(t\right) = V\left(x(t), t\right)$,

$$\sigma = -\max_{1\le i\le N}\left\{\frac{\lambda_{\min}\left(P_iA_i + A_i^TP_i\right)}{\lambda_{\max}\left(P_i\right)}\right\} > 0,$$

and $\lambda_{\min}\left(\cdot\right)$ (resp. $\lambda_{\max}\left(\cdot\right)$) is the minimum (maximum) eigenvalue of a (symmetric) matrix.

Now suppose system (4) undergoes k times switching at $t_1 < t_2 < \cdots < t_k \leq t$ on interval $[t_0, t]$. From above and conditions (i) and (ii) we have

$$\begin{aligned} \dot{V}(\tau) &\leq -\sigma V(\tau), \; t_{i-1} < \tau < t_i, \\ V\left(t_{i-1}^{+}\right) &\leq V(t_{i-1}), \; \tau = t_{i-1}^{+}, \end{aligned}$$

for $i = 1, ..., k$. It thus follows that

$$\begin{aligned} V(t) &\leq V(t_k)\, e^{-\sigma(t-t_k)} \\ &\leq V\left(t_{i-1}^{+}\right) e^{-\sigma[(t-t_k)+(t_k-t_{k-1})]} \\ &\quad \vdots \\ &\leq V(t_0)\, e^{-\sigma[(t-t_k)+(t_k-t_{k-1})+\cdots+(t_1-t_0)]} \\ &= V(t_0)\, e^{-\sigma(t-t_0)}. \end{aligned}$$

On the other hand, it is clear from definition of $V(t)$ that

$$\mu \left\| x(t) \right\|^2 \leq V(t) \leq \eta \left\| x(t) \right\|^2,$$

where $\mu = \min_{1 \leq i \leq N} \{\lambda_{\min}(P_i)\} > 0$ and $\eta = \max_{1 \leq i \leq N} \{\lambda_{\max}(P_i)\} > 0$. From this and above we can obtain

$$\left\| x(t) \right\| \leq \kappa \left\| x(t_0) \right\| e^{-\rho(t-t_0)}$$

for any $t \geq t_0$, where $\kappa = \sqrt{\eta/\mu} \geq 1$, $\rho = \sigma/2 > 0$. So the system is globally exponentially stable.

Fault Tolerant Supervisory for Discrete Event Systems Based on Event Observer

Fei Xue and Da-Zhong Zheng

Department of Automation, Tsinghua University, Beijing 100084 China
xuefei00@mails.tsinghua.edu.cn

Abstract. The fault tolerant supervisory problem for discrete event systems is addressed in this paper. The proposed approach is based on the state avoidance control theory and observer-based control for Petri net. The key idea of the authors is to use a simple linear algebraic formalism to estimate system states and generate diagnostic information. Hence, the state explosion problem is avoided and the observer-based fault diagnosis algorithm can be made on-line.

1 Introduction

As complexity of the modern man-made systems has increased, concerns for safety and reliability have grown. Thus system-theoretic methods on failure diagnosis [1], [2], [3] and fault-tolerant control [6], [7], [8] have been developed. In [6], the quantitative definitions of faults, failures, and fault tolerant systems are proposed in the Ramadge–Wonham framework for control of discrete event systems [10], and the supervisory control for *fault tolerable event sequences* are developed under the assumption that all transitions of system are observable. In [7], [8], this problem is extended to systems with uncertain model and partial observation. However, these are either mainly concerned with the fault isolation problems, or the fault transitions and the normal unobservable transitions of system are treated identically, thus the fault tolerant supervisory problem degrade to supervisory problem with unobservable transitions [9]. In real world, the fault transitions are sparse in system and occur sparsely. If the fault transitions and normal unobservable transitions are treated identically, it will make the restriction of system behaviors too rigorously and make unnecessary waste. Hence, it is required to establish a whole framework including state identification, diagnostics and fault-tolerant supervision.

In this paper, the fault-tolerant supervisory for discrete event systems modeled in Petri net is addressed. The proposed approach is based on the state avoidance control theory [9] and observer-based control for Petri net [4], [5]. The fault-tolerant supervisory framework in this paper includes two parts. Firstly, the supervisor is constructed based on the normal behaviors of system and restrict system to avoid forbidden states. Secondly, an event observer is constructed to estimate system states and generate diagnostic information. Different with the approaches in [7], [8], in this paper, the occurrences of fault transitions are assumed to be sparse haphazard. Hence, if and only if the occurrence of a fault transition is detected by the event observer, the supervisor will be reconfigured according to the diagnostic information.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 655 – 664, 2005.

On the other hand, the state explosion problem in on-line estimation of the system states is a critical problem for diagnosis [2], [3]. The information about possible states is also required when the supervisor is reconfigured. This makes the reconfiguration of the supervisor much more complex and difficult to operate on-line. We show that the sparse assumption of faults allows us to use a simple linear algebraic formalism to generate diagnostic information. An observer-based fault diagnosis algorithm is firstly given in this paper. Secondly, to ensure the supervisor is permissible after a fault occurrence, it is required that the occurrence of fault can be detected before the system reach forbidden states, namely *tolerant detectable*. We also explain how to restrict the system's normal behavior to ensure the *tolerant detectable* property.

2 System Model and Assumptions

In this section, we recall some basic definitions of Petri nets. For a more comprehensive introduction to Petri nets, see [11]. A Petri net is a bipartite graph

$$N = (P, T, Pre, Post) \tag{1}$$

where P is a set of places; T is a set of transitions; $Pre: P \times T \to Z$ and $Post: T \times P \to Z$, where Z is the set of nonnegative integers, are the pre-incidence and post-incidence function respectively. A net is *pure* if no place is both input place and output place of the same transition. A pure Petri net can be represented by the *incidence matrix* defined as $C(p,t) = Post(p,t) - Pre(p,t)$. A Petri net is said to be *acyclic* if and only if there is not directed cycle in it. A *state* of Petri net is a mapping $M : P \to Z$ that assigns to each place a nonnegative integer number of so-called tokens. A Petri net (N, M_0) is a net N with initial state M_0. A transition t is enabled at a state M iff $M \geq Pre(\bullet, t)$. An enabled transition t may fire yielding a new state M' such that $M' = M + C(\bullet, t)$, and this will be denoted as $M[t > M'$. A state M' is reachable from a state M if there exists a firing sequence $\sigma = t_1 t_2 \cdots t_n$ transforming M into M'. This is denoted as $M[\sigma > M'$. Any state M reachable from M_0 by firing a sequence σ satisfies the following state equation: $M = M_0 + C\vec{\sigma}$, where $\vec{\sigma}: T \to Z$ is a vector of nonnegative integers, called the *occurrence vector* whose *i*-th entry denotes the number of occurrences of transition t_i in σ.

The discrete event systems in this paper are modeled in Petri net $N = (P, T, Pre, Post)$. The transitions of system are classified into two subsets, normal transitions and fault transitions, as following.

$$T = T_n \cup T_f \text{ and } T_n \cap T_f = \emptyset \tag{2}$$

Denote the sub-system $N_o = (P, T_n, Pre', Post')$ without fault transitions as *original Petri net*, where Pre' and $Post'$ are the pre-incidence function and post-incidence function respectively, which restricted to normal transitions set T_n. It is assumed that following conditions are hold for the system.

A1) System $N = (P, T, Pre, Post)$ is pure and the normal behaviors of system are live. I.e. for any reachable state M, $\exists t \in T_n$ that $M[t >$.

A2) The initial state M_0 is known exactly.

A3) The normal transitions are observable and controllable. By contraries, the fault transitions are unobservable and uncontrollable.

A4) The fault transitions are sparse in system behaviors.

3 Fault Tolerant Supervisory Based on Event Observer----Single Fault Case

One of the most challenging problems for on-line diagnosis is the state explosion problem in on-line estimation of the system states [2], [3]. This makes the reconfiguration of supervisor much more complex and difficult to operate on-line. We show that the sparse assumption of faults allows us to use a simple linear algebraic formalism to estimate system states and generate diagnostic information. The fault tolerant supervisory algorithm will be firstly given in this section for single fault case. For the multiple fault case, it will be studied in the next section.

3.1 Single Fault Assumption

In this section, the assumption A4 is replaced by the more restricted one as follows, namely single fault assumption. This will avoid enumerating all possible states in the diagnosis process and make the generation of diagnostic information much simpler.

A4') The *single fault* assumption is hold. I.e. in any transition sequence of the system, at most one failure transition is included.

3.2 Fault Detection in Single Fault Case

Before studying the fault tolerant supervisory problem, we will first recall the definition of detectability and then study the fault detection in the single fault case.

Definition 1: Petri net (N, M_0) is said to be detectable with respect to fault transitions set T_f if there exist $n \in Z$, such that for $\forall s \in T^*$ is defined at M_0, if $\exists t_f \in T_f$ that is included in s and at least n steps occur after t_f in s, then

$$l(s') = l(s) \Rightarrow t'_f \in T_f, t'_f \in s' \tag{3}$$

where $l(s)$ is the observable sequence of transition sequence s, which only contain the observable transitions in s.

The above definition means that Petri net (N, M_0) is detectable if and only if every fault transition $t_f \in T_f$ can be detected if the system operates sufficient long time (at least n steps) after the fault occurred. That is, if a transition sequence s' gives the same observable sequence as the faulty sequence s, then s' must also contain some fault transition $t'_f \in T_f$.

Theorem 1: In the single fault case, Petri net (N, M_0) is detectable with respect to fault transitions set T_f if and only if the following condition holds.

C1) For the original Petri net N_o with initial state M_0, it holds that

$$\begin{aligned}&(\exists n \in Z)(\forall t_f \in T_f)(\forall M \in R(N_o, M_0) \wedge M[t_f >)\\ &\qquad\qquad [(\forall s \in T_n^*, M[s >) \wedge Q \Rightarrow \| s \| \leq n]\end{aligned} \tag{4}$$

where condition Q =: for any state M' in the trace $M[s >$, t_f is enabled.

Proof: For $\forall t_f \in T_f$, we will prove the inverse and negative proposition of Theorem 1, instead of prove it directly. The negation of definition of detectability is that for any $n \in Z$, there exist transition sequences as follows.

s with fault: $M_0[t_1 > M_1 \cdots M'_0[t_f > M''_0[t'_1 > M''_1[t'_2 > M''_2 \cdots M''_n[t'_n >$

s' without fault: $M_0[t_1 > M_1 \cdots M'_0[t'_1 > M'_1[t'_2 > M'_2 \cdots M'_n[t'_n >$

The negative position for condition C1 is that for any $n \in Z$, there exists a trace

$$(M^*): M_0[t_1 > M_1[t_2 > M_2 \cdots M'_0[t'_1 > M'_1[t'_2 > M'_2[t'_3 > M'_3 \cdots M'_n[t'_n >$$

and $\exists t_f \in T_f$ that $M'_i[t_f >$, $\forall i = 0, 1, 2, \cdots n$. We will prove that the negation of detectability definition and the negation of condition C1 are equal to each other.

Necessity (from detectable to C1): Firstly, we will prove that if $M'_0[t'_1 > M'_1$ and $M'_0[t_f > M''_0[t'_1 > M''_1$, then $M'_1[t_f >$.

$$M'_0[t'_1 > M'_1 \Rightarrow M'_0 \geq Pre(\bullet, t'_1) \text{ and } M'_1 = M'_0 + C(\bullet, t'_1)$$

$$M'_0[t_f > M''_0[t'_1 > \Rightarrow M'_0 \geq Pre(\bullet, t_f)$$

$$M''_0 = M'_0 + C(\bullet, t_f) \geq Pre(\bullet, t'_1)$$

Hence, $M'_0 - Pre(\bullet, t'_1) + Post(t_f, \bullet) \geq Pre(\bullet, t_f)$. Based on the assumption that system is pure, we can get that

$$M'_1 = M'_0 - Pre(\bullet, t'_1) + Post(t'_1, \bullet) \geq Pre(\bullet, t_f) \Rightarrow M'_1[t_f >$$

$$M''_1 = M''_0 + C(\bullet, t'_1) = M'_0 + C(\bullet, t_f) + C(\bullet, t'_1) = M'_1 + C(\bullet, t_f)$$

We get that $M_1'[t_2' > M_2'$ and $M_1'[t_f > M_1''[t_2' > M_2''$. Similarly, we can prove that $M_i'[t_f >$, $i = 2,\cdots n$. The negation of condition C1 is proved.

Sufficiency (from C1 to detectable): Firstly, we will prove that if $M_0'[t_1' > M_1'$, $M_0'[t_f > M_0''$ and $M_1'[t_f >$, it is satisfied that $M_0''[t_1' >$ and $M_1'[t_f > M_1''$.

$$M_0'[t_1' > M_1' \Rightarrow M_1' = M_0' + C(\bullet, t_1') \text{ and } M_0' \geq Pre(\bullet, t_1')$$

$$M_0'[t_f > \Rightarrow M_0'' = M_0' + C(\bullet, t_f) \text{ and } M_0' \geq Pre(\bullet, t_f)$$

$$M_1'[t_f > \Rightarrow M_1' \geq Pre(\bullet, t_f)$$

Hence, $M_1' = M_0' + C(\bullet, t_1') \geq Pre(\bullet, t_f)$. As the proof in necessity, we can get that

$$M_0'' = M_0' + C(\bullet, t_f) \geq Pre(\bullet, t_1') \Rightarrow M_0''[t_1' >$$

$$M_1'' = M_0' + C(\bullet, t_f) + C(\bullet, t_1') = M_1' + C(\bullet, t_f)$$

It is similarly to prove that $M_i''[t_{i+1}' >$ and $M_{i+1}'[t_f > M_{i+1}''$, $i = 1,\cdots,n$. The negation of detectability definition is proved.

3.3 Diagnostic Information Generated by Event Observer

Based on the proof of Theorem 1, we find that the diagnostic information will not lose before the fault is detected in the single fault case. It allows us to use simple linear algebraic formalism based on event observer to generate diagnostic information.

Algorithm 3.1(generation of the diagnostic information based on event observer):

1. Initial estimation $\mu_0 = M_0$, possible fault set $F = \emptyset$, let $i = 1$;
2. Wait until $t_i \in T_n$ fires;
3. If t_i is enable at μ_{i-1}, update $\mu_i = \mu_{i-1}[t_i >$, $i = i+1$ and go to step 2;
4. Else, $F = \{t_f \mid t_f \in T_f$ that $\mu_{i-1}[t_f >$ and $\mu' = \mu_{i-1} + C(\bullet, t_f)[t_i >\}$;
5. Send the reconfigure message and diagnostic information (μ_{i-1}, F, t_i) to system supervisor.

3.4 Fault Tolerant Supervisory

Algorithm 3.1 gives a simple method to detect fault and generate diagnostic information based on the event observer. Once a fault is detected by Algorithm 1, it will send the reconfigure message and diagnostic information to system supervisor. And then, the system supervisor will be reconfigured and adjust its supervisory rules. In the state avoidance supervisory [9], the specification of system is defined as a set of forbidden states Q. The system supervisor may disable transitions to prevent the plant from entering forbidden states. In the fault tolerant supervisory framework, it requires that the system supervisor can be reconfigured after fault occurrence and can ensure the system behaviors in safe region to avoid forbidden states. In this subsection, we will study how the system supervisor should adjust its supervisory rules based on the diagnostic information. Here, we assume the system supervisor has been well constructed for normal system behaviors and only focus on the supervisory rules after a fault is detected.

Algorithm 3.2(supervisor adjusts based on diagnostic information):

1. System supervisor run based on the original Petri net (N_o, M_0);
2. Wait until reconfigure message and information (μ, F, t) is received;
3. Reconfigure system supervisor by update the possible states set
$$\bar{M} := \{\mu' \mid \exists t_f \in F, \mu' = \mu + C(\bullet, t_f) + C(\bullet, t)\};$$
4. Transition $t \in T_n$ is disabled if $\exists M \in \bar{M}$, $M[t >$ is a forbidden state;
5. Wait until next transition $t' \in T_n$ fires;
6. Update $\bar{M}' := \{M' \mid \exists M \in \bar{M}, M[t > M'\}$ and go to step 4.

In the single fault case, the maximal number of possible states in $\bar{M}$ is equal to $|T_f|$. Hence, the complexity of algorithm 2 is acceptable by online fault tolerant supervisory. Moreover, in above algorithm, the step 5 and step 6 are not only supervisor steps, but also fault diagnosis steps. The fault will be located more exactly step by step. This process can be regarded as transient region of system supervisor after its reconfiguration. If the system is diagnosable with respect to fault transitions set in the single fault case, the exact state of system can be determined in bounded steps after a fault occurrence. Correspondingly, the system supervisor will reach new stabilization (i.e. will run based on the original Petri net (N_o, M_0) with new initial state) in bounded steps after it is reconfigured. Hence, the single fault assumption can be relaxed as following.

A4) The occurrences of fault transitions are enough spares in the system behaviors, i.e. the interval between occurrences of two fault transitions is long enough that the supervisor can reach new stabilization after its reconfiguration.

4 Fault Tolerant Supervisory Based on Event Observer----Multiple Fault Case

In above section, we study the event observer-based fault tolerant supervisory in single fault case. In real world, one fault always will result some new faults and it will propagate in the system. Hence, the single fault assumption is often too strict in real industry application. In this section, we will try to relax this assumption and study the more complex multiple fault case.

4.1 Sparse Assumption in Multiple Fault Case

As shown in above section, the single fault assumption ensures that diagnostic information will not lose before the fault is detected. However, this property is not held in general multiple fault case. Thus, it will require enumerating all possible states in the diagnosis process step by step. That will make the fault tolerant supervisor algorithm too complexity to be operated online. Looking for simpler case, we consider the multiple fault case which satisfying following sparse fault assumption.

A4″) The fault transitions are distributing sparsely in system structure, i.e. the sub-system $N_f = (P, T_f, Pre', Post')$ is an acyclic Petri net, where Pre' and $Post'$ are the pre-incidence and post-incidence function respectively, which restricted to fault transitions set T_f. Moreover, the occurrence of fault transitions is sparse in system behaviors, i.e. once if fault is detected, there will be no new fault transition is detected before the exactly state of system can be determined.

4.2 Fault Detection in Multiple Fault Case

In this sub-section, we will study the fault detection in multiple fault case under sparse fault assumption A4″. The occurrence of fault transitions will be detected if and only if the observable transition sequence can not be explained by system normal behaviors. Based on assumption A3 that all normal transitions are observable, fault will be detected if and only if the observable sequence is not enable in original Petri net $N_o = (P, T_n, Pre', Post')$ with initial state M_0.

Theorem 2: In multiple fault case under sparse fault assumption A4″, give a fault sequence s in Petri net (N, M_0). If the observable sequence $s' = l(s)$ is enable in original Petri net (N_o, M_0), then transition vector $\vec{\sigma}$ is enable at state $M[s' > M'$, where $\vec{\sigma}$ is the occurrence vector of fault transitions in s.

Proof: Firstly, we will prove that if $M_0'[t_1' > M_1'$ and $M_0'[\sigma_f > M_0''[t_1' > M_1''$, there exist σ_f' that $M_1'[\sigma_f' >$ and $\vec{\sigma}_f' = \vec{\sigma}_f$. Here σ_f and σ_f' are fault transition sequences. We have

$$M_0'[t_1' > M_1' \Rightarrow M_0' \geq Pre(\bullet, t_1') \text{ and } M_1' = M_0' + C(\bullet, t_1'),$$

$$M_0'[\sigma_f > M_0''[t_1' > \Rightarrow M_0' + C\vec{\sigma}_f \geq 0 \text{ and } M_0'' = M_0' + C\vec{\sigma}_f \geq Pre(\bullet, t_1').$$

Hence, we get $M_0' - Pre(\bullet, t_1') + C\vec{\sigma}_f \geq 0$. Because $Post(t_1', \bullet) \geq 0$, we have

$$M_0' - Pre(\bullet, t_1') + Post(t_1', \bullet) + C\vec{\sigma}_f = M_1' + C\vec{\sigma}_f \geq 0$$

Based on assumption A4” that $N_f = (P, T_f, Pre', Post')$ is acyclic, it holds that occurrence vector $\vec{\sigma}_f$ is enable at state M_1', i.e. $\exists \vec{\sigma}_f = \vec{\sigma}_f'$ that $M_1'[\sigma_f' >$. Similarly, the proof can repeat in the fault sequence s. The theorem 2 is proved.

4.3 Diagnostic Information and Fault Tolerant Supervisory

As shown in the above subsection, theorem 2 shows that under sparse fault assumption A4″, the diagnostic information will not lose before the fault is detected in multiple fault case. Moreover, based on the assumption A4″, the sub-net N_f is acyclic, i.e. the unobservable part of system is acyclic. Thus, it allows us to use a simple linear algebraic formalism based on event observer to estimate system states. The fault detection and fault tolerant supervisory algorithm is given as following.

Algorithm 4.1(generation of the diagnostic information based on event observer):

1. Initial estimation $\mu_0 = M_0$, $i = 1$;
2. Wait until $t_i \in T_n$ fires;
3. If t_i is enable at μ_{i-1}, update $\mu_i = \mu_{i-1}[t_i >$, $i = i + 1$ and go to step 2;
4. Else, send the reconfigure message and diagnostic information (μ_{i-1}, t_i) to system supervisor.

Algorithm 4.2(supervisor adjusts based on diagnostic information):

1. System supervisor run based on the original Petri net (N_o, M_0);
2. Wait until reconfigure message and information (μ, t) is received;
3. Reconfigure system supervisor by update the possible states set
$$\bar{M} := \{\mu' = \mu + C\vec{\sigma} + C(\bullet, t) \mid \exists \sigma \in T_f^*, \mu + C\vec{\sigma} \geq Pre(\bullet, t)\};$$
4. Transition $t \in T_n$ is disabled if $M \in \bar{M}$ and $M[t >$ is a forbidden state;
5. Wait until next transition $t' \in T_n$ fires;
6. Update $\bar{M}' := \{M' \mid \exists M \in \bar{M}, M[t > M'\}$ and go to step 4.

Algorithm 4.1 gives a simple method to detect fault and generate diagnostic information based on the event observer. Once fault is detected, it will send the reconfigure message and diagnostic information to system supervisor. Then, the system supervisor will be reconfigured as algorithm 4.2. If system state can be determined in bounded steps after fault occurrence, the system supervisor will reach new stabilization in bounded steps after it is reconfigured

5 Tolerable Detectability

In the fault tolerant supervisory framework, to ensure the supervisor after fault occurrence is feasible, it is required that the occurrence of fault should be detected before the system reach forbidden states. This constraint is also required to prevent faults from developing into failures that could cause safety hazards. In [13], this problem, namely *safety diagnosability*, is studied in the framework of finite state machine with permissive event strings. In this section, we will study this problem, namely *tolerant detectability*, in the framework of state avoidance supervisory. The formally definition of tolerable detectability is given as follows.

Definition 2 (Tolerable Detectability): Given Petri net (N, M_0) and forbidden states set Q. It is said to be tolerable detectable with respect to fault transitions set T_f and forbidden states set Q if it holds that 1) It is detectable with respect to fault transitions set T_f ; 2) For the shortest prefix that fault can be detected, it has never reached forbidden states before the fault is detected.

Considering the above definition, firstly it requires the normal system behaviors have been well restricted, i.e. it will not reach forbidden states in original Petri net. Secondly, it requires Petri net (N, M_0) is detectable with respect to fault transitions set T_f. Hence, based on the discussion in above sections, we know that under sparse fault assumption A4″, it is tolerant detectable with respect to forbidden states set Q if and only if for reachable state M in the original Petri net (N_o, M_0), it will not reach forbidden states by firing fault transitions sequence. For general Petri net with forbidden states, the tolerant detectable problem is much more complex. The verifying condition for tolerant detectability can be seen in another paper [12].

6 Discussions and Conclusion

In this paper, we propose a fault-tolerant supervisory framework for discrete event systems modeled in Petri net. The proposed approach is based on the state avoidance control theory [9], [10] and observer-based control for Petri net [4], [5]. Different with the previous approaches, in this paper, the occurrences of fault transitions are assumed to be sparse in system behaviors. If the occurrence of fault transitions is detected, the supervisor will be reconfigured according to diagnostic information generated by the observer. We show that under the sparse fault assumption, a simple linear algebraic formalism can be used to estimate system states and generate diagnostic information. Hence, the state explosion problem is avoided and the fault tolerant supervisory algorithm can be run online. To ensure the supervisor after fault occurrence is permissible, the *tolerant detectable* property is also studied.

Acknowledgements

This research is partially supported by National Science Foundation of China (60074012, 60274011).

References

1. M.Sampath, R.Sengupta, S.Lafortune, K.Sinnamohideen and D.Teneketzis: Failure Diagnosis Using Discrete-event Models. IEEE Transactions on Control System Technology, (1996) **4** (2): 105-24

2. Albert Benveniste, Eric Fabre, Stefan Haar, Claude Jard: Diagnosis of Asynchronous Discrete-event Systems: A Net Unfolding Approach. IEEE Transactions on Automatic Control, (2003) **48** (5):714-727
3. Laurence Roze, Marie-Odile Cordier: Diagnosing Discrete-event Systems: Extending the 'Diagnoser Approach' to Deal with Telecommunication Networks. Discrete Event Dynamic Systems: Theory and Applications, (2002) **12** (1): 43-81
4. Alessandro Giua, Carla Seatzu: Observability of Place/Transition Nets. IEEE Transactions on Automatic Control, (2002) **47** (9):1424-1437
5. Alessandro Giua, Carla Seatzu, Fransesco Basile: Observer-based State-feedback Control of Timed Petri Nets with Deadlock Recovery. IEEE Transactions on Automatic Control, (2004) **49** (1):17-29
6. Kwang-Hyun Cho, Jong-Tae Lim: Synthesis of Fault-tolerant Supervisor for Automated Manufacturing Systems: A Case Study on Photolithographic Process. IEEE Transactions on Robotics and Automation, (1998) **14** (2):348-351
7. Kwang-Hyun Cho, Jong-Tae Lim: Fault-tolerant Supervisory Control under C, D Observability and Its Application. International Journal of Systems Science, (2000) **31** (12): 1573-1583
8. Seong-Jin Park, Jong-Tae Lim: Fault-tolerant Robust Supervisor for Discrete Event Systems with Model Uncertainty and Its Application to a Workcell. IEEE Transactions on Robotics and Automation, (1999) **15** (2):386-391
9. L. E. Holloway, X. Guan, and L. Zhang: A Generalization of State Avoidance Policies for Controlled Petri Nets. IEEE Transactions on Automatic Control, (1996) **41** (6):804-816
10. P. J. Ramadge, W. M. Wonham: Supervisory control of a class of discrete event processes. SIAM Journal of Control and Optimal, (1987) **25** (1):206-230
11. Tadao Murata: Petri Nets: Properties, Analysis and Applications. Proceedings of IEEE, (1989) **77** (4):541-579
12. Fei Xue, Da-Zhong Zheng: Tolerable Fault Detectability for Discrete Event Systems accepted by ICCA05
13. A.Paoli, S.Lafortune: Safe Diagnosability of Discrete Event Systems. Proceedings of 42nd IEEE Conference on Decision and Control, Maui, Hawaii USA, (2003) 2658-2664

A Study on the Effect of Interference on Time Hopping Binary PPM Impulse Radio System

YangSun Lee[1], HeauJo Kang[1], MalRey Lee[2], and Tai-hoon Kim[3]

[1] Division of Computer & Multimedia Content Eng. 800, Doan-Dong, Taejon,
Mokwon University, Taejon, 305-729, Korea
{yslee, hjkang}@mokwon.ac.kr
[2] School of electronics & information Eng. 664-14, 1 Ga, Deokjin-Dong, Jeon ju, ChonBuk
Chonbuk National University , Jeonju-si, 561-756, Korea
mrlee@chonbuk.ac.kr
[3] San-7, Geoyeo-Dong, Songpa-Gu, Seoul, Korea
taihoonn@empal.com

Abstract. In this paper, the effects of the interference environments on the performance of the time hopping (TH) binary PPM impulse radio (IR) system are presented. Based on the monocycle pulses available within the frequency of 3.1~10.6 GHz permitted for application by FCC, a PPM modulated TH IR system simulator was designed and followed by the analysis of the monocycle pulse characteristics as well as the system performance. Particularly for the evaluation of the system performance, the multiple access interference and the narrowband system interference signals were considered as the interference signals. Since the narrowband system interference signal has very narrow bandwidth and very large amplitude compared with those of IR system, the analysis of the IR system performance was implemented by considering the interference power and band fraction ratio of the narrowband interference signal.

1 Introduction

The problem for frequency resources is very serious even though world each country's effort including advanced nation to maximize efficiency for confined frequency resources in present wireless communication areas. By sharing and using the existent communication system and frequency spectrum to solve this problem, Ultra Wideband(UWB or Impulse Radio : IR) communication method that can use frequency resources more efficiently appeared. Such as Ultra Wideband communication method does not give interference to other communication system existing in form such as baseband noise. It is profitable than existing communication system in which transmission speed because it can adopt bandwidth by wideband. And because of unuse carrier wave was used compulsorily in existing system, It can reduce power dissipation of transceiver and make simplify transceiver[1], [2].

On February 2002, Federal Communications Commission (FCC) partially was allowed UWB technology to be used for commercial applications and thus the UWB based communication systems has been developed [3].

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 665 – 674, 2005.

Unfortunately, each conventional study has been done under each own system specification and environment. And the system specification can not used in the frequency range permitted by FCC in the year of 2002 [3], [4], [5]. Also, it must consider interference problem from existent narrow band system when it shares frequency, because it is a low power transmitting mode.

In this paper, the effects of the interference environments on the performance of the time hopping (TH) binary PPM impulse radio (IR) system is presented. Based on the monocycle pulses available within the frequency of 3.1~10.6GHz [3] permitted for application by FCC, a PPM modulated TH IR system simulator was designed in an AWGN environment. Using the simulator, the monocycle pulse characteristics and system performance were analyzed. Particularly for the evaluation of the system performance, the multiple access interference (MAI) and the narrowband system interference signal were considered as the interference signals. Since the narrowband system interference signal has very narrow bandwidth and very large amplitude compared with those of IR system, the IR system performance was analyzed and evaluated by considering the interference power and band fraction ratio of the narrowband interference signal.

2 TH Binary PPM IR System

2.1 Characteristic of Monocycle Pulse

IR pulse $p_{RX}(t)$ has duration T_p and energy $\exp\left(-2\pi[t/t_n]^2\right)$. For $p_{RX}(t)$, we consider a IR pulse that can be modeled by the second derivative of a Gaussian function $\exp\left(-2\pi[t/t_n]^2\right)$ properly scaled[4]. In this case, the transmitted pulse is

$$P_{TX}(t) = t\exp\left(-2\pi\left[\frac{t}{t_n}\right]^2\right) \tag{1}$$

Where t_n is a parameter which determines the temporal width of the pulse, and it is approximately 0.0326 nanoseconds(ns) in this work. In this paper the received signal is modeled as the second derivative of a Gaussian signal like in (2) :

$$p_{RX}(t) = \left(1-4\pi\left[\frac{t}{t_n}\right]^2\right)\exp\left(-2\pi\left[\frac{t}{t_n}\right]^2\right) \tag{2}$$

Where, the spectral and temporal characteristics of the signal depend on the parameter t_n.

The conventional IR receiver correlates the received signals with a template signal and makes binary decisions depending on the sign of correlation values, The normalized autocorrelation function of $p_{RX}(t)$ is given by [5].

$$\gamma_p(\tau) = \frac{1}{E_p}\int_{-\infty}^{+\infty} p_{RX}(t)p_{RX}(t+\tau)dt$$

$$= \left[1 - 4\pi\left[\frac{\tau}{t_n}\right]^2 + \frac{4\pi^2}{3}\left[\frac{\tau}{t_n}\right]^4\right]\exp\left(-\pi\left[\frac{\tau}{t_n}\right]^2\right) \tag{3}$$

Where, the template signal is given by

$$v(t) = p_{RX}(t) - p_{RX}(t-\delta) \tag{4}$$

2.2 Design of Binary PPM IR System Simulator

In this section, we designed PPM modulated IR system simulator. and we using a simulator to analyze the characteristics of this system.

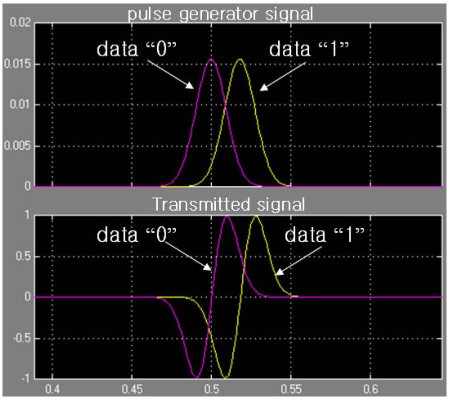


Fig. 1. Monocycle pulse signal of transmitter

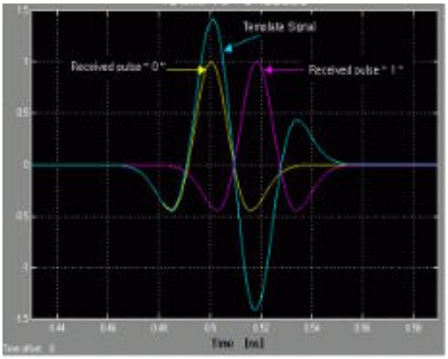


Fig. 2. Received signal waveform “0” , ”1” and template signal of correlator output

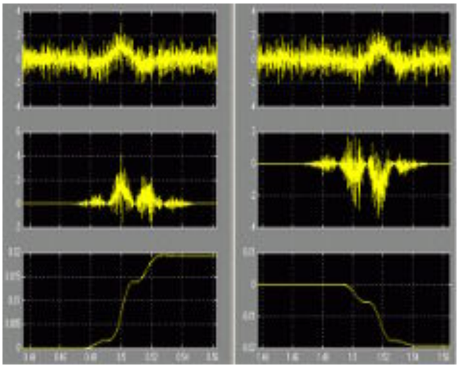


Fig. 3. Correlator output according to received signal waveform “0” , “1”

Fig. 1 ~ Fig. 3 illustrate signal waveform of transceiver and the process of the correlator output. In this modulation method, when the data symbol is “0”, no additional time shift is modulated on the monocycle, but a time shift δ is added to a monocycle when the symbol is “1”. In Fig. 3 and Fig. 4, when the data symbol is “0”, correlator output would become a positive value, and when the data symbol is “1”, correlator output would become a negative value.

The correlation receiver, under hypothesis of perfect synchronization, computes:

$$\beta_i = \sum_{j=iN_s}^{(i+1)N_s} \int_{\tau+iT_f}^{\tau+(j+1)T_f} r(t)v\left(t - jT_f - c_jT_c - \tau\right)dt \tag{5}$$

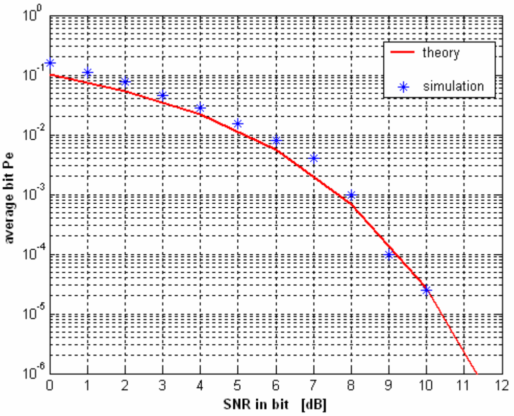


Fig. 4. Error probability of Binary PPM IR system in AWGN(N_s =1, no MAI)

Decision is made according to the following rule:

$$\hat{\alpha}_i = \begin{cases} 0 & if \quad \beta_i \geq 0 \\ 1 & if \quad \beta_i < 0 \end{cases} \tag{6}$$

Fig. 4 show the theoretical BER performance and simulation of Binary PPM IR system.

3 Performance Evaluation of IR System

3.1 C-Channel Interference(CCI) Cancellation Technique

When N_u transmitters are active, the received signal can be modelled as,

$$r(t) = A_s s^{(1)}(t-\tau_1) + n_{tot}(t) \tag{7}$$

$$n_{tot}(t) \equiv n(t) + \sum_{k=2}^{N^u} A_k s^{(k)}(t-\tau_k)$$

$$= N_s \sigma_a^2 \sum_{k=2}^{N_u} A_k^2 \tag{8}$$

Where, $\sigma_n^2 = N_0(1-\gamma(\delta))$

$$\sigma_a^2 = T_f^{-1} \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \sqrt{E_p}\, v(t) p_{RX}(t-s)dt \right]^2 ds$$

$$= \frac{E_p}{T_f} \int_{-\infty}^{\infty} \left[\gamma(s) - \gamma(s+\delta)^2 ds \right]$$

$$= \frac{E_p}{T_f} G(\delta) \tag{9}$$

and $G(\delta) = \int_{-\infty}^{\infty} [\gamma(s) - \gamma(s+\delta)]^2 ds$

$n_{tot}(t)$ is assumed to be a mean-zero Gaussian random process. Standard techniques [8] can then be used to show that the probability of error P_b, when using the decision procedure of (6), is given by [8]

$$P_b = \frac{1}{2} erfc\left(\sqrt{\left\{(SNR)_1^{-1} + 2R_s P(\delta)\sum_{k=2}^{N_u}\left(\frac{A_k}{A_1}\right)^2\right\}^{-1}}\right) \tag{10}$$

Where, $(SNR)_1 = A_1^2 \frac{N_s E_p}{2N_0}(1-\gamma(\delta))$,

$$R_s = \frac{1}{N_s T_f}(bps) \text{, and } P(\delta) = \frac{G(\delta)}{(1-\gamma(\delta))^2} .$$

The simulation parameters are used to performance evaluation that is presented in table 1. and, Fig. 5 Show BER performance of IR system using parameters of table 1.

Table 1. Parameter for simulation

Parameter	Symbol	Value	
Waveform width parameter	t_n	0.0326 ns	
Waveform width	T_p	0.072 ns	
Binary PPM parameter	δ	0.0176 ns	
Autocorrelation value	$\gamma(\delta)$	-0.6183	
Pulses per bit	N_s	Variable(1 , 2, 4 , 10)	
Frame width	T_f	10 ns	5 ns
Bit rate	R_s	Variable(100, 50, 25, 20,10) Mbps	
P function value	$P(\delta)$	0.586814 X 10^{-10}	

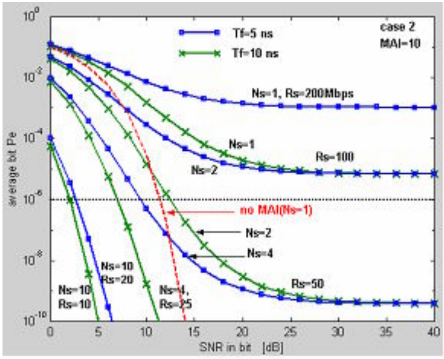


Fig. 5. BER performance of IR system according to variable bit rate in presence of MAI

In the case of no MAI, IR system is satisfied reference BER(10^{-6}) of the data service more than SNR 11.45dB, but it is decreased in presence of MAI. Especially, we observed that proposed IR systems decreased more than 50Mbps by MAI. Then, IR systems that effect by MAI is serious in a high speed transmission environment. Therefore, when we design high speed IR system, it must need to design a considering MAI. And, Variable of bit rate using T_f can improve performance more than 50Mbps. But it would decrease maximum user's number. Besides, increase N_s means the bit rate decrease. As a result, we can satisfy required reception SNR and bit rate by we specify maximum T_f according to user number in indoor environment and design N_s according to get quality of service.

3.2 Narrowband Interference Model

Fraction bandwidth of IR system is 6.8GHz in simulation parameters of table 1. and Narrow band signal of interferers can consider of 802.11a WLAN system and different narrow band systems in fraction bands of IR system. Bandwidth of 802.11a is very small such as fraction ratio 0.002941 in IR bandwidth. Moreover, it has very bigger amplitude than IR system. Thus, in this paper, we analyze effect of the narrow band interference is considering a signal power and fraction ratio of interference on the TH binary PPM IR system is an amplitude of narrow band signal. If W_{ss} is a bandwidth of IR system and W_I is bandwidth of interference signals, $\rho = W_I / W_{ss}$. Where, ρ is a fraction ratio of interference signal. Therefore, error performance of IR system adds error probability of AWGN environment and error probability of narrow band interference environment. It is possible to evaluate the probability of error in presence of narrow band interference.

$$P_e = (1-\rho)P_b + \rho P_b$$

$$= \frac{1-\rho}{2} erfc\left(\sqrt{(SNR)_1}\right) + \frac{\rho}{2} erfc\left(\sqrt{\frac{1}{\frac{1}{(SNR)_1} + \frac{1}{SIR}\frac{1}{\rho}}}\right) \tag{11}$$

Where, SIR is signal power to interference power ratio.

From Fig. 6 and Fig. 7, we can know that the system performance is reduced greatly by narrow band interference. In the case of SIR ≤30dB and ρ ≥0.01, system performance is improved. And the case ρ ≥0.01, performance of system consist more than SIR =20dB. But generally, System performance could know that standard service is not effective, even if the received power is increased. Therefore, IR system could know that fraction bandwidth of the interference reduces performance much more than electric power of the interference.

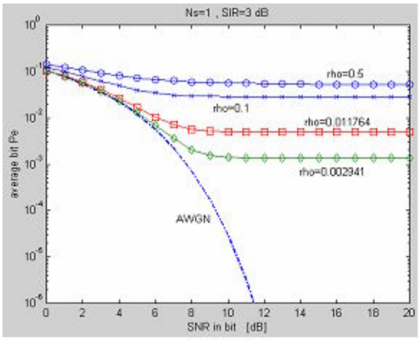


Fig. 6. Performance of IR system by narrowband interference(rho= ρ)

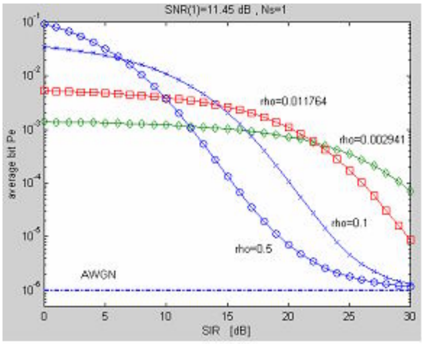


Fig. 7. BER according to variable SIR under narrowband interference environment

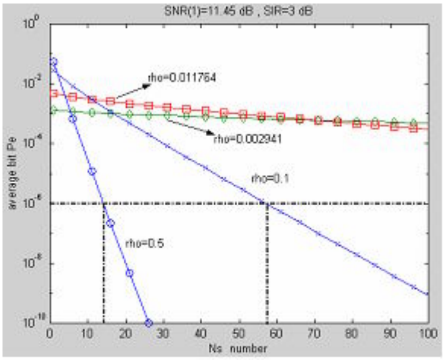


Fig. 8. Performance of IR system by fraction ratio of narrowband interference

Fig. 8 shows performance of IR system according to N_s and ρ, especially when $(SNR)_1$=11.45dB and SIR =3dB.

The case of no MAI in Figure 4., By increase of N_s, performance improvement in AWGN environment has been achieved. On the contrary, in presence of the narrow band interference such as Figure 10, performance improvement has been achieved by increasing N_s in $\rho \geq$0.1. but, the case of $\rho \leq$0.1, the narrow band's interference did not influence in the system performance, even if increase N_s.

4 Conclusion

In this paper, the effects of the interference environments on the performance of the time hopping (TH) binary PPM impulse radio (IR) system are presented. Based on the monocycle pulses available within the frequency of 3.1□10.6GHz[3] permitted for application by FCC, a PPM-modulated TH IR system simulator was designed in an AWGN environment. Using the simulator, the monocycle pulse characteristics and system performance were analyzed. Particularly for the evaluation of the system performance, the MAI and the narrowband system interference signals were considered as the interference signals. Since the narrowband system interference signal has a very narrow bandwidth and very large amplitude, compared with those of IR system, the IR system performance was analyzed and evaluated by considering to the interference power and band fraction ratio of the narrowband interference signal.

According to the results, we could know that we must need a duration of Monocycle pulse and setting of T_f according to multiple user's number and design proper pulse repetition number by bit rate. Besides, we observed that IR system is decreased performance by MAI. The case of performance analysis of IR system is considering to the interference power and bandwidth fraction ratio in presence of narrow band interference environment, it reduced more system performances than MAI. and We could know that Performance is decreased clearly by fraction bandwidth of interference more than amplitude of interference signal. Also, performance improvement has been achieved by increasing N_s in $\rho \geq$0.1. but, in the case of $\rho \leq$0.1, narrow band interference did not influence in system performance even if it increased a N_s. By the way, Increase of N_s reduces bit rate. Therefore, the design of IR system in interference environment that occupy more than IR band's 10% selects proper N_s parameter and coding technique to compensate errors by interference that should be applied. And, When $\rho \leq$0.1, The systems design that have optimum transmission efficiency would apply robust suppression's techniques about the interference of narrowband more than increase of N_s parameter may be available.

References

1. P. Withington: UWB Regulation & Applications. UWB Workshop. www.timedomain.com, 11. 2001
2. Aetherwire & Location Inc.. www.aetherwire.com
3. FCC Notice of Proposed Rule Making: Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission System. ET-Docket, pp. 98-153. Feb. 2002
4. M. Z. Win and R. A. Scholtz: Impulse Radio: How It Works. IEEE Comm. Lett., vol. 2, pp. 36-38, Feb. 1998
5. F. Ramirez-Mireles: On Performance of Ultra Wideband Signals in Gaussian Noise and Dense Multipath. accepted for publication in IEEE Trans. On Veh. Tech. 1999
6. J. T. Conroy, J. L. Locierto, D. R. Ucci: Communication Techniques using Monopulse Waveforms. MILCOM, vol. 2, pp. 1181-1185, 1999
7. R. A. Scholtz: Multiple Access with Time Hopping Impulse Modulation. Proc. IEEE MILCOM '93, pp. 447-450, Oct. 1993
8. J. M. Wozencraft and I. M. Jacobs: Principles of Communication Engineering. John Wiley, 1965, Chapter 4
9. M. K. Simon, et al.: Spread Spectrum Communication Handbook. McGraw-Hill, 1994. 4

A Study on the Enhanced Detection Method Considering the Channel Response in OFDM Based WLAN

Hyoung-Goo Jeon[1], Hyun Lee[2], Won-Chul Choi[2], Hyun-Seo Oh[2],
and Kyoung-Rok Cho[3]

[1] Dong Eui University, Busan, Korea
hgjeon@deu.ac.kr
[2] Electronics and Telecommunications Research Institute, Daejeon, Korea
{hyunlee, wcchoi, hsoh5}@etri.re.kr
[3] School of Electrical Engineering Chungbuk Nat'l University, Chungbuk, Korea
krcho@cbu.ac.kr

Abstract. In this paper, we proposed a channel estimation method by impulse signal train in OFDM. In order to estimate the channel response, 4 impulse signals are generated and transmitted during one OFDM (Orthogonal Frequency Division Multiplexing) symbol. The intervals between the impulse signals are all equal in time domain. At the receiver, the impulse response signals are summed and averaged. And then, the averaged impulse response signal is zero padded and fast Fourier transformed to obtain the channel estimation. The BER performance of the proposed method is compared with those of conventional channel estimation method using the long training sequence in fast fading environments. The simulation results show that the proposed method improves by 3 dB in terms of Eb/No, compared with the conventional method.

1 Introduction

Recently, OFDM has been effectively used for transmitting high speed data in multipath fading environment. In OFDM, the high speed data is parallel processed and transmitted by N(=power of 2) orthogonal sub-carriers. One OFDM symbol duration is N times longer than the duration of the original data bit. Therefore, OFDM has robustness over multi-path fading environment. For coherent detection in OFDM, we have to estimate the channel frequency response information. In OFDM systems, channel estimation is very important, because the performance of receivers is very dependant on the precise of the channel estimation.

There have been studies on time domain and frequency domain channel estimation methods. In recent, DFT(Discrete Fourier Transform) based channel estimation method was proposed to reduce the affection by AWGN, using zero padding[2]. The method takes into account the maximum length of the impulse channel response. Two DFT based channel estimation methods were proposed in Ref. [3]. One is MMSE method, the other one is LS method. The MMSE method has a high complexity to implement. Moreover, if a prior knowledge of noise variance and channel covariance is not available, the MMSE method can not be used. The LS method has a low complexity, but not a good performance.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 675 – 684, 2005.

To increase the precise of the channel estimation, in this paper, we proposed a channel estimation method by impulse signal train. In OFDM, an impulse signal can be easily made by performing inverse FFT (fast Fourier transform) over all one data on all N sub-carriers. If the impulse signal is transmitted as a OFDM symbol, the channel estimation becomes very simple. In the proposed method, four equal interval impulse signals are generated and transmitted during one OFDM symbol period. The impulse signal train generation and channel estimation method will be described further in section III.

This paper is organized as follows. Section II introduces the conventional channel estimation method. In section III, we propose the channel estimation method by impulse signal train. In section IV, we describe the simulation to evaluate the proposed method. Finally, section V presents our conclusions.

2 Channel Estimation in Frequency Domain

Channel estimation is required for coherent detection in OFDM systems. Long training sequence is used to estimate the channel before start transmitting the data. Fig. 1 shows the frame structure of IEEE 802.11a physical layer.

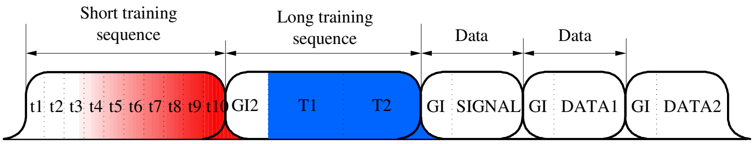


Fig. 1. Frame structure of IEEE 802.11a physical layer

The long training sequence is a known data to receivers. When the known data x_n passed through the channel, the received signal y_n can be expressed as equation (1).

$$y_n = x_n * h_n + w_n \tag{1}$$

where h_n is the impulse channel response, w_n is AWGN, * symbol denotes convolution. Equation (1) becomes equation (2) in frequency domain.

$$Y_k = X_k H_k + W_k, \quad 0 \le k \le N-1 \tag{2}$$

From equation (2), the channel response can be estimated in frequency domain as equation (3).

$$\hat{H}_k = Y_k / X_k + W_k / X_k, \quad 0 \le k \le N-1 \tag{3}$$

3 Proposed Channel Estimation Method by Impulse Signal Train

The length of channel impulse response is much shorter than that of an OFDM symbol. Therefore, more than one channel impulse response can exist during one OFDM symbol. In this paper, impulse signal train in time domain is proposed as the training sequence for channel estimation. We consider IEEE 802.11a OFDM modem and assumed that the maximum impulse response length L = 16 and the sub-carrier number N = 64. Fig. 2 shows the conceptual block diagram of the proposed channel estimation method.

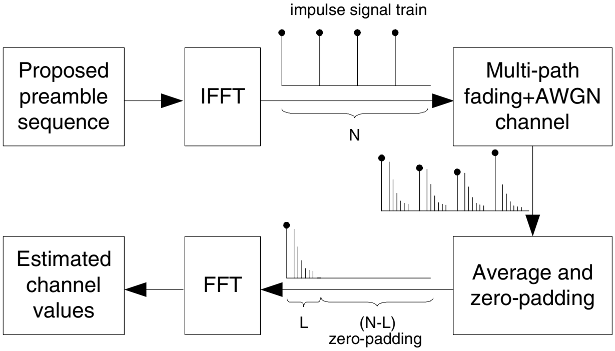


Fig. 2. The proposed channel estimation method

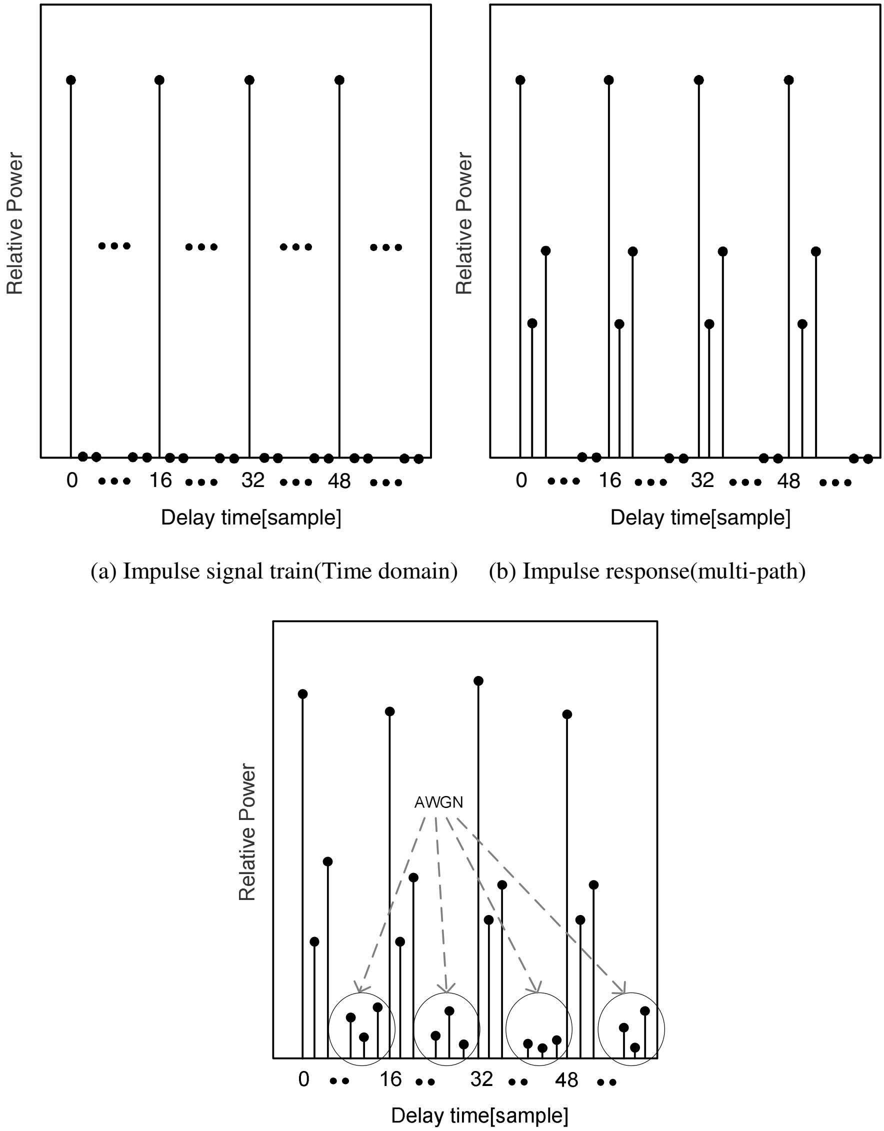
In the proposed method, one OFDM symbol period is equally divided into 4 sections. We designed a preamble sequence in frequency domain so that one impulse signal can be generated in each section. A total of 4 impulse signals exist in one training OFDM symbol as shown in Fig. 2. The proposed preamble signal in frequency domain can be expressed as equation (4).

$$new\_preamble[n] = \begin{cases} 4, & if \;\; n = 4k, \;\; k = 0,1,2,...15 \\ 0, & \qquad otherwise \end{cases} \tag{4}$$

The time domain signal can be obtained by performing IFFT over the proposed preamble signal and be expressed as equation (5).

$$\begin{aligned} TS(n) &= IFFT(new\_preamble[n]) \\ &= \delta(n) + \delta(n-L) + \delta(n-2L) + \delta(n-3L), \quad 0 \le n \le N-1 \end{aligned} \tag{5}$$

where $\delta(n)$ is unit impulse function which is one only when n = 0. TS(n) signal is an impulse signal train as shown in Fig. 3-(a). When the impulse signal train passed through a wireless multi-path channel, an example of the received signal is shown in Fig. 3-(c) and (d). It can be assumed that during one OFDM symbol the channel is invariant with time.



(a) Impulse signal train(Time domain) (b) Impulse response(multi-path)

(d) Impulse response(multi-path with AWGN)

Fig. 3. Impulse response examples

Received signal $r(n)$ can be expressed as equation (6).

$$r(n) = TS(n) * h(n) + n(n) \tag{6}$$

where $n(n)$, $h(n)$, and * symbol denote AWGN noise, channel impulse response, and convolution operation, respectively. Equation (6) can be expressed as equation (7)

$$\begin{aligned} r(n) &= \delta(n) * h(n) + \delta(n-L) * h(n) + \delta(n-2L) * h(n) + \delta(n-3L) * h(n) + n(n) \\ &= h(n) + h(n-L) + h(n-2L) + h(n-3L) + n(n) \end{aligned} \tag{7}$$

We consider only a causal system. And the maximum channel response length is assumed to be L sample duration. In this case, $h(n) = 0, if\ n < 0$ or $n > L-1$. Therefore, from equation (7), it can be seen that a series of 4 channel responses with L sample interval are there. In the OFDM receiver, 4 received impulse signals can be regarded as an identical impulse channel response so that they are averaged to make one received impulse signal as expressed in equation (8). Additionally, the effect of AWGN noise is reduced by the averaging operation.

$$\begin{aligned} h_a(n) &= \frac{1}{4}\sum_{i=0}^{3} r(n+iL) \\ &= h(n) + \frac{1}{4}\sum_{i=0}^{3} n_i(n), \qquad 0 \le n \le L-1 \end{aligned} \tag{8}$$

where $n_i(n) = n(n+iL), i = 0,1,2,3$. Each random process $n_i(n)$ is independent and has an identical probability distribution. Therefore, noise variance in equation (8) is reduced as much as 1/4 factor. That means an increased accuracy in the channel estimation. where $h(n)$ is the channel response by the impulse signal train and $n_i(n)$ is AWGN. Fig. 4 is an example of an averaged impulse response obtained by equation (8) under the condition of multi-path and AWGN of Fig 3-(d). In this paper, the maximum length of impulse channel response, L, is assumed to be 16. Therefore, the sample over the maximum length is set to be zero.

The affection of AWGN in the channel estimation can be reduced by averaging the received impulse response signal. To make N point FFT (Fast Fourier Transform) and eliminate AWGN components included in (N-L) samples after the maximum impulse response length, (N-L) zeros are padded at the sample positions of more than L. The zero padded impulse response signal $\tilde{h}(n)$ is expressed as equation (9).

$$\tilde{h}(n) = \begin{cases} h_a(n) & , \quad 0 \le n \le L-1 \\ 0 & , \quad L \le n \le N-1 \end{cases} \tag{9}$$

We can estimate the channel by performing FFT over the zero padded impulse response signal $\tilde{h}(n)$ as shown in equation (10).

$$H(k) = FFT\{\tilde{h}(k)\}, \quad 0 \le k \le N-1 \tag{10}$$

In indoor environment, channel response length is shorter than that of outdoor environment. According to the channel response length, the zero padding length should be determined. How to determine the channel response length is beyond our topic. In this paper, we assume that L = 5 in indoor wireless channel and L = 16 in outdoor wireless channel.

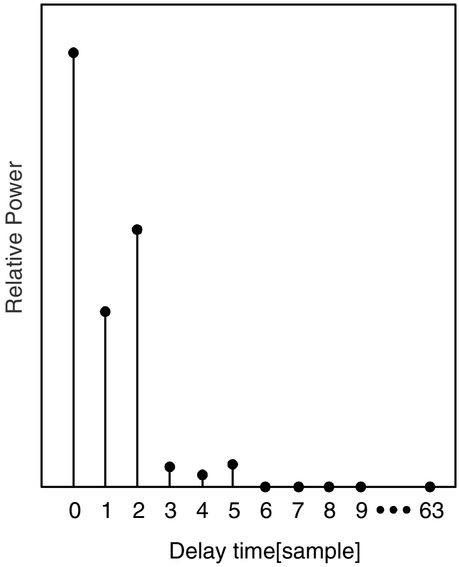


Fig. 4. Impulse response with zero padding

4 Simulation

In order to evaluate the performance of the proposed method, a computer simulation is carried out under the condition of multi-path with 3 paths. In indoor environment, Doppler frequency is set to 0 Hz, and maximum delay is assumed to be 4 samples. In outdoor environment, Doppler frequency is set to 200 Hz, and maximum delay is assumed to be 16 samples, considering the worst case. The simulation condition is described in table 1. The MMSE method is not considered in this simulation, because the method needs a prior knowledge about the channel to be estimated.

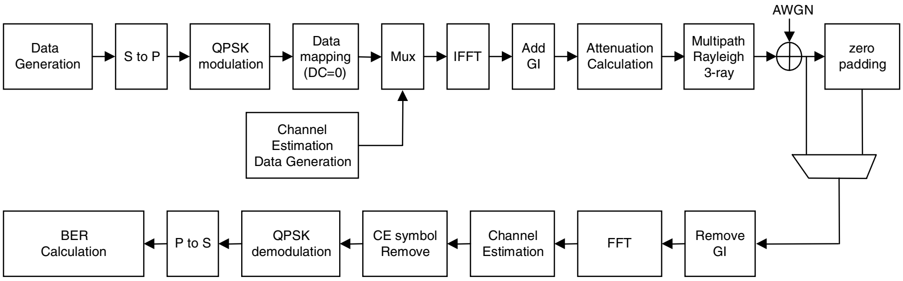


Fig. 5. Simulation model block

Fig. 5 shows the OFDM system model block for the simulation. In this simulation, we assumed a perfect synchronization in the OFDM system. The BER performance of the proposed method is compare with those of the long preamble method and the LS method.

Table 1. Simulation parameters

Item	Value
Fading channel	Multipath Rayleigh fading channel(3ray)
Mean power	3-ray (0, 10, 20) dB 4-ray (0, 20, 10, 30) dB
Arrival time delay	3-ray (0, 2, 4) samples 4-ray (0, 2, 3, 4) samples
Modulation	QPSK
Doppler Frequency	0 Hz(indoor), 200Hz(outdoor)
Loop	10000

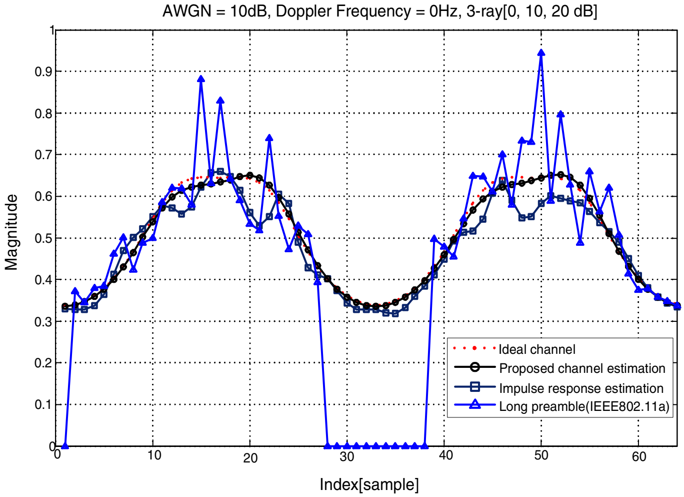
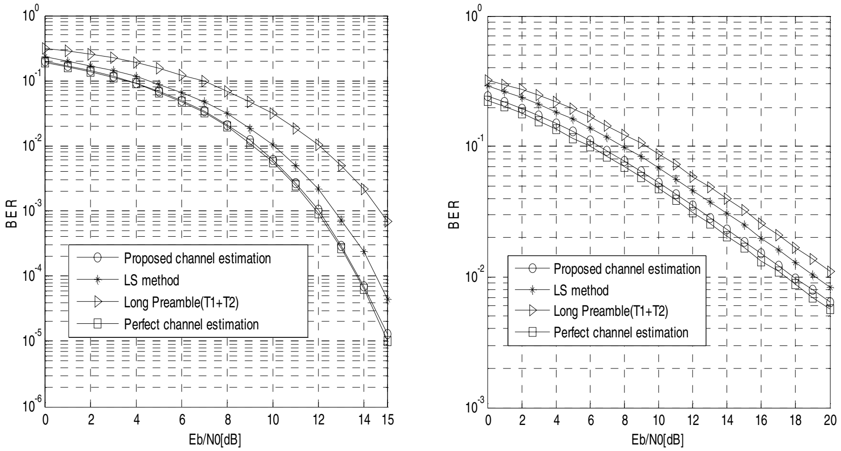


Fig. 6. Comparison of channel estimation values

Fig. 6 shows the channel estimation results obtained by long preamble method, LS method, and the proposed method in indoor environment (L = 5). As we can see from Fig. 6, the proposed method is very close to the ideal channel estimation. When L is a small value (L<< N), we can see that the proposed method performance is not different from that of the LS estimation. However, it is expected that the averaging effect in the proposed method will become larger at outdoor environments in which L is N/4. Long preamble method shows the worst performance in Fig. 6.

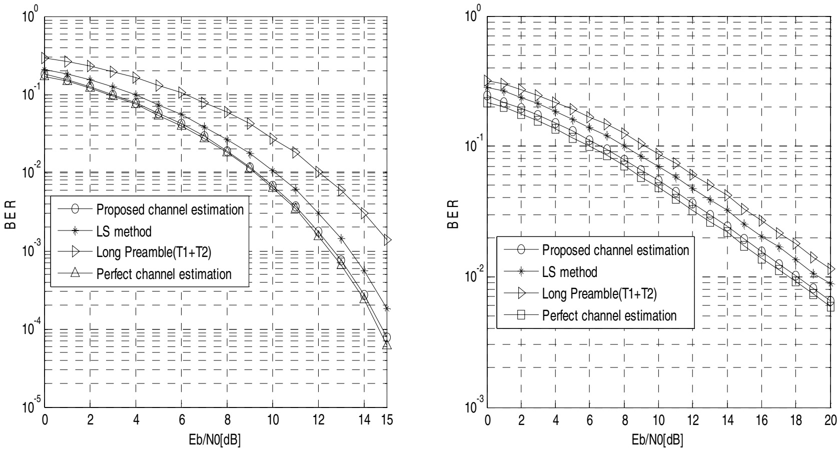
Fig. 7-(a) shows the BER curves obtained from the long preamble method, LS method, and the proposed method under the condition of 3 multi-path, the channel response length L = 5, and 0 Hz Doppler frequency in indoor environment. The proposed method has about 2 dB and 0.4 dB gains in terms of Eb/No, compared with those of long preamble method and the LS method, respectively. In indoor environments in which the channel response length is very short, the zero padding length becomes longer. In this case, the effect of averaging impulse responses does not make a large difference in the BER performance. Fig. 7-(b) shows the BER curves obtained from the long preamble method, LS method, and the proposed method under the condition of 3 paths and 200 Hz Doppler frequency of an outdoor high speed mobile vehicle environment. In high speed mobile vehicle environments, the channel is variant fast with time. Therefore, the long preamble method using two training symbols (L_{T1} and L_{T2}) is expected to be ineffective, because of its long estimation time. Since only one training symbol is needed in the proposed method, it is expected to be effective in high speed mobile vehicle environment. In most cases, we do not know the channel impulse response length in outdoor environment. Therefore, we have to assume L = 16, considering the worst case. In this case, the effect of averaging four impulse signal response is expected to be large.

Fig. 8 shows the BER curves obtained from the long preamble method, LS method, and the proposed method under the condition of 4 paths and 0, 200Hz Doppler frequency.



(a) Indoor environment (b) outdoor environment(f_d =200Hz)

Fig. 7. BER curves in 3 path indoor, outdoor environments



(a) Indoor environment

(b) outdoor environment(f_d=200Hz)

Fig. 8. BER curves in 4 path indoor, outdoor environments

As expected, the proposed method has about 3 dB gain in terms of Eb/No, compared with the long preamble method, and about 1.5 dB gain, compared with the LS method.

5 Conclusions

The performance of OFDM receivers is highly dependant on the precise of channel estimation. In this paper, we proposed a channel estimation method by using impulse signal train in OFDM. In order to estimate the channel response, 4 impulse signals are generated and transmitted during one OFDM symbol. The intervals between the impulse signals are all equal in time domain. At the receiver, the impulse response signals are summed and averaged. And then, the averaged impulse response signal is zero padded and fast Fourier transformed to obtain the channel response information.

The BER performance of the proposed method is evaluated in a fast fading environment and an indoor environment, comparing with those of other channel estimation methods such as the long preamble method and the LS method. The simulation results show that the proposed method improves by 3 dB in terms of Eb/No, compared with the long preamble method, and 1.5 dB, compared with the LS method.

In cases where the mobile speed is high and the OFDM symbol length is relatively long, like WiBro, it is required to estimate the channel response rapidly. The proposed method guarantees a rapid channel estimation and requires no calculation complexity when estimating the channel response. From this point of view, the proposed method would be more effective in fast fading channels rather than in slow fading channels.

References

1. John Terry and Juha Heiskala, OFDM Wireless LANs: A Theoretical and Practical Guide. Sams Publishing, 2002
2. A. M Saleh and R. A. Valenzuela: A Statistical Model for Indoor Multipath Propagation. IEEE J. on Selected Areas in commum., vol. 5, no. 2, pp. 128-137, Feb, 1987
3. O. Edfors, M. Sandell, and J.-J van de Be et al.: OFDM Channel Estimation by Singular Value Decomposition. IEEE Trans. Commun., vol. 46, pp. 931-939. July 1998
4. J.-J van de Beek and O. Edfors, et al.: On Channel Estimation in OFDM Systems. in IEEE VTC'95., 1995, pp. 815-819
5. Hlating Minn and Vijay K. Bhargava: DFT-based Channel Estimation in 2D-pilot-symbol-aided OFDM Wireless Systems. in IEEE VTC'01., 2001, pp. 810-814
6. F.Tufvesson and T. Maseng: Pilot Assisted Channel Estimation for OFDM in Mobile Cellular Systems. in Proc. VTC, pp. 1639-1643, 1997
7. N. Weste and D. J. Skellern: VLSI for OFDM. IEEE Commun, Mag, vol. 36, pp. 127-131, Oct. 1998

Block Error Performance Improvement of DS/CDMA System with Hybrid Techniques in Nakagami Fading Channel

Heau Jo Kang[1] and Mal Rey Lee[2]

[1] Division of Computer & Multimeda Content Eng. 800, Doan-Dong, Taejon
Mokwon University, Taejon, 305-729, Korea
hjkang@mokwon.ac.kr
[2] School of Electronics & Information Eng. 664-14, 1 Ga, Deokjin-Dong, Jeon ju, ChonBuk
Chonbuk National University, Jeonju-si, 561-756, Korea
Mrlee@chonbuk.ac.kr

Abstract. As results of study, the coding techniques provide more efficient improvement than a diversity technique, but coding techniques are required the adding bandwidth as many coding rate. Also, when the system is combined MRC diversity technique with coding techniques, the amount of improvement is dramatically increased.

1 Introduction

The Direct-Sequence (DS) CDMA in wireless and cellular mobile communications is attractive to the view point of random access capability and resistance against multipath fading [1].

The transmission of information over radio channels with multiple changing propagation paths is subject to fading, i.e., random time variations of the receiver signal strength. For digital transmission over a fading channel, time variation causes a changing error probability with the effect of clustering errors at the receiver output.

The present work was motivated by the need to evaluate block data transmission performance under fading conditions for application to sequential polling of vehicles in a fleet. These systems commonly employ some form of error control, such as error detection block coding, so that evaluation of performance requires the analysis of the probabilities distribution of the number of errors in code blocks transmitted under fading conditions. In data communication applications, the expressions of error probabilities are important in evaluating system performance.

Previous work on error probabilities have been conducted considering mobile radio channels with Rayleigh [2], [3], [4], Rician [5] and m-distribution [6] fading characteristics. The m-distribution is chosen to characterize the fading channel because it takes the Rayleigh distribution as a special case, approximates the Rician distribution well, models fading conditions which are more or less severe than those of Rayleigh, and more importantly, fits experimental data better than Rayleigh or Rician distributions [7], [8].

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 685 – 694, 2005.

In this paper, we analyze the error performance improvement of the DS/CDMA BPSK system combining with repetition transmission, FEC code, and MRC diversity technique in mobile communication channel which is characterized m-distribution fading.

2 Performance of A DS/CDMA System in Nakagami Fading Channel

This section is concerned with the calculation of the error probability of the DS/CDMA communications in a multipath faded channel that is modeled by a discrete set of m-distribution faded paths. The system, which is illustrated in Fig. 1, consists of K users. Each user is assumed to be used BPSK along with DS spread-spectrum modulation. The k th user first generates data bits at a rate of $1/T$ bits per second. Its binary data signal $b_k(t)$ and signature sequence signal $a_k(t)$ are given by [9]

$$b_k(t) = \sum_{i=-\infty}^{\infty} b_i^{(k)} p_T(t - iT) \tag{1}$$

$$a_k(t) = \sum_{i=-\infty}^{\infty} a_i^{(k)} p_{T_c}(t - iT_c) \tag{2}$$

where $p_\tau(t)$ is a unit rectangular pulse on $(0,\tau), b_i^{(k)}$ is one symbol of the k th transmitted signal and $a_i^{(k)}$ is the k th user's code sequence. Both of these symbols take on values in $\{-1,1\}$ and we have $a_i^{(k)} = a_{i+N}^{(k)}$ for all i and k and for some integer $N = T/T_c$ where T is the bit interval duration and T_c is the chip length.

The modulated BPSK signals of k users are transmitted to channel, and then they are distorted by fading which is characterize a m-distribution.

A m-distribution characterizes channels with different fading depths through a parameter called amount of fading. The signal envelope, R is a random variable with a m-distribution probability density function (pdf) [10] i.e.,

A m-distribution characterizes channels with different fading depths through a parameter called amount of fading. The signal envelope, R is a random variable with a m-distribution probability density function (pdf) [10] i.e.,

$$p(R) = \frac{2m^m R^{2m-1}}{\Gamma(m)\Omega^m} \exp(-\frac{mR^2}{\Omega}) \tag{3}$$

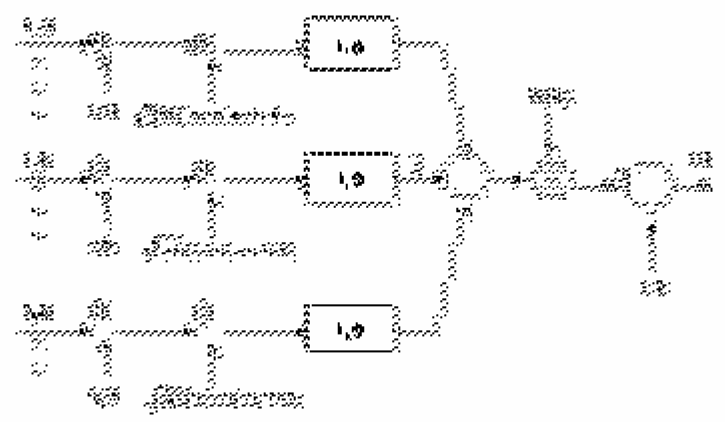


Fig. 1. DS/CDMA system model

where $\Gamma(\cdot)$ is the Gamma function, $\Omega/2=\overline{R^2}/2$ is the mean power of the enveloped signal by fading, and is fading index($m=\Omega^2/(\overline{R^2-\Omega})\geq 1/2$). if we are represented to equation (3) by $(=\dfrac{R^2}{2N})$, the pdf of γ is found to be

$$p(\gamma)=\frac{m^m\gamma^{m-1}}{\Gamma(m)\gamma_0}\exp\left(-\frac{m\gamma}{\gamma_0}\right) \tag{4}$$

where $\gamma_0=\Omega/2N$ is average signal to noise power ratio. Note that (3) and (4) is represented to Rayleigh and AWGN pdf, when m is 1, and m is infinity, respectively.

The pdf of Nakagami fading which changes equation (4) into average signal to noise ratio of DS/CDMA system, is [2].

$$p_{cdma}(\gamma)=\frac{m^m\gamma^{m-1}}{\Gamma(m)\Gamma'}\exp\left(-\frac{m\gamma}{\Gamma'}\right) \tag{5}$$

where Γ' is average signal to noise power ratio in DS/CDMA system and is given by

$$\Gamma'=\frac{1}{\dfrac{2(L\cdot K-1)}{3PG}+\dfrac{N_0}{2E_b}} \tag{6}$$

where L is the multipath number of channel, K is the number of multiple access user, and PG is processing gain.

The bit error probability of BPSK with a signal to noise power ratio (SNR) γ is

$$P_{SM}(\gamma) = \frac{1}{2} erfc(\sqrt{\gamma}) \tag{7}$$

The bit error probability for BPSK signaling over a m-distribution fading can be found by averaging the bit error probability of (7) with respect to the pdf of fading.

$$P_{SMP} = \int_0^\infty P_{SM}(\gamma) p_{cdma}(\gamma) d\gamma \tag{8}$$

3 Performance of Diversity and Block Coded Schemes in Nakagami Fading Channel

A MRC Diversity Technique

Assuming a maximal ratio combining approach with L identical branches, the diversity effect is examined as follows. The output signal to noise power ratio after maximal ratio combining is equal to the sum of the signal to noise power ratio of the various combining branches. It can be shown that pdf of the resulted signal to noise power ratio is

$$p_{mrc(\gamma)} = \frac{\gamma^{mL-1}}{\Gamma(mL)} \left(\frac{m}{\Gamma'} \right)^{mL} \exp\left(-\frac{m\gamma}{\Gamma'} \right) \tag{9}$$

The average error probability of BPSK signal over the m-distribution fading channel with MRC diversity reception is

$$P_{MRC} = \int_0^\infty P_{SM}\, p_{mrc}(\gamma) d\gamma \tag{10}$$

B Repetition Transmission

In this technique each message is sent an odd number of times and at the reception, a bit by bit majority decision is applied. If s is the number of repeats, a $(s+1)/2$ out of s majority voting process is used to determine each valid bit in the message.

The probability of bit error, P_{TB}, after an $(s+1)/2$ out of s majority voting is the probability of at least $(s+1)/2$ errors occurring, i.e.,

$$P_{TB} = \sum_{i=(s+1)/2}^{s} \binom{s}{i} (P)^i (1-P)^{s-i} \tag{11}$$

where P_{TB} is the bit error probability of a single transmission. The probability of having exactly j corrupted bits in an N-bit message is

$$P_{TB}(N,j)=\begin{pmatrix} N \\ j \end{pmatrix}(1-P_{TB})^{N-j}(P_{TB})^{j} \tag{12}$$

when errors generate more than t bits, the probability of message error is

$$P_{eM}=1-\sum_{j=0}^{t} P_{TB}(N,j) \tag{13}$$

where $P_{TB}(N,j)$ is given by equation (12).

Repetition transmission is also a type of code, known as repetition code. The repetition code is represented as (s ,1), having a minimum distance equal to s. Therefore, it is able to correct up (s -1)/2 errors in an s -bit message, having a rate $R=1/s$.

C Error Correcting Code

A linear block code having k information bits and $n-k$ redundancy bits is described as (n,k). The ratio $R=k/n$ is called the code rate. Moreover, if its minimum distance is $d_{\min}$, this code is able to correct up to $t=(d_{\min}-1)/2$ bits out of n bits.

It is easy to see that, for a code that can correct up to t bits, the probability of message error is

$$P_{eM}=1-\sum_{j=0}^{t} p(N,j) \tag{14}$$

$$P(N,j)=\begin{pmatrix} \frac{N}{j} \end{pmatrix}(1-P)^{(N-j)}P^{j} \tag{15}$$

where $P(N,j)$ is the probability of having exactly m bits error in an N bit message, and P is error probability according to modulation technique.

The bit error probability $P(N,j)$ used in Equation (15) varies according to be the modulation technique. It is important to note that, in order to keep the same transmitted power, the SNR per bit in the encoded message is multiplied by the code rate, R . In other words, the average SNR per bit of the encoded message is $(k/N)\Gamma'$.

As we consider another error correction code ,one special subclass of the BCH codes is the particularly useful nonbinary set called Reed-Solomon codes. RS codes achieve the largest possible code minimum distance for any linear code with same encoder input and output block lengths. RS codes are particularly useful for burst-error correction; that is, they are effective for channels that have memory. Also, they can be used efficiently on channels where the set of input symbols is large. For nonbinary codes, the distance between two code words is defined as the number of nonbinary symbols in which the sequences differ. For RS codes the code minimum distance is given by

$$d_{\min} = n - k + 1 \tag{16}$$

where k is the number of information symbols being encoded, and n is the total number of code symbols in the encoded block. The code is capable of correcting any combination of t or fewer symbol errors, as follows

$$t = \frac{d_{\min} - 1}{2} = \frac{n-k}{2} \tag{17}$$

A -error-correcting RS code with an alphabet of 2^m symbols has $n = 2^m - 1$ and $k = 2^m - 1 - 2t$, where m =2, 3, ...

The RS decoded symbol error probability, P_E [10], can be written in terms of the channel symbol error probability, p

$$P_E = \frac{1}{n} \sum_{j=t+1}^{n} j \binom{n}{j} p^j (1-p)^{n-j} \tag{18}$$

The bit error probability can be upper bounded by the symbol error probability for specific modulation types.

4 Comparative Performance and Combined Techniques and Numerical Results

In this paper, we now consider the performance of the DS/CDMA BPSK system in m-distribution fading channel. As a technique for the performance improvement, diversity, coding, and repetition transmission have been used, and their performance have been compared and analyzed.

To compare the performance of both coding and repetition transmission, we set the error correcting capability at the same t=3 bits. Accordingly, we select 2 branch MRC diversity, (23,12) Golay code, (7,5) RS code, and (5,1) Majority voting.

Fig. 2 is shown the effects of diversity according to the number of multiple access user when m is equal to 3. As user number is increased, interval between curves is decreased by degrees. it is shown that, with no diversity, voice service is available to less than 5 users. With, however, diversity, voice service is reachable up to 20 users, even data service is possible with less than 5 users.

Fig. 3 show BER performance related to fading figure m and the number of multiple access user at E_b / N_0 =18 dB. As fading is deeper(m decreases), improvement of diversity decreases. Also even though user increases, slope of curve is not large. But it is shown that in the case of m =3, improvement of diversity is large, and sensitive to user.

In the case of user=10, and m =3, fig 4 represents degree of improvement according to improvement techniques(diversity, code, and repetition transmission). In order to compare coding techniques with repetition transmission, we adopt to be equal to error correcting ability(t =3). we are shown that when use only one of improvement techniques, at higher E_b / N_0 , Golay code is better than the others, and when use diversity combined with codes, also Golay code is the best.

Fig. 5 shows relation of the required E_b / N_0 according to the number of user to satisfy data service quality(BER= 10^{-5}). In the case of only diversity, user is able to up to 8. In the case of only coding or repetition transmission, it is shown that Golay, RS, and Majority voting is restricted within 14, 10, and 12 respectively. But diversity combined with codes techniques can be reduced restriction of users.

When use only Coding or diversity, coding is better than diversity at high E_b / N_0 , but the former has problem in aspect of increase of bandwidth. When it is combined MRC diversity techniques with coding techniques, the amount of improvement is increased. But Bit error performance enhancement is accomplished at the expense of increasing system complexity and cost. Some techniques can yield better results than others, but can be more costly. Therefore, the decision for one or another technique depends on the analysis of cost versus effectiveness.

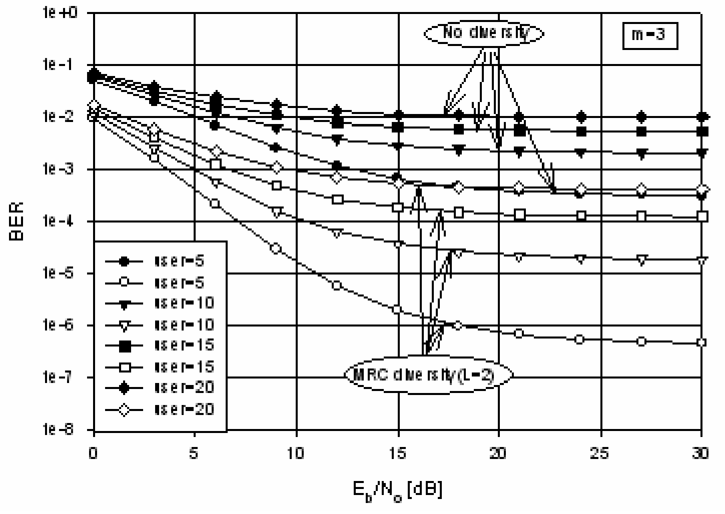


Fig. 2. Error rates for DS/CDMA BPSK system according to MRC diversity and user($L = 2, m = 3$)

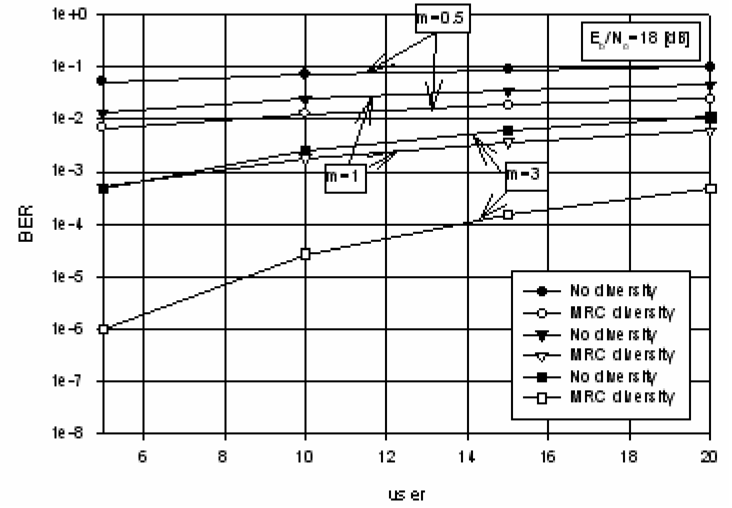


Fig. 3. Error rates for DS/CDMA BPSK signal not using and using maximal ratio diversity according to user and fading index m ($L = 2$)

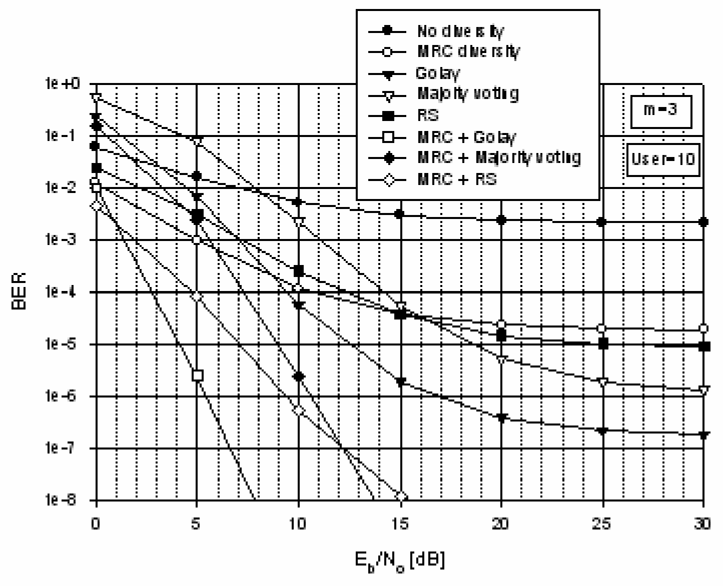


Fig. 4. Comparison of the techniques using diversity combined with coding and repetition transmission (user=10, m =3)

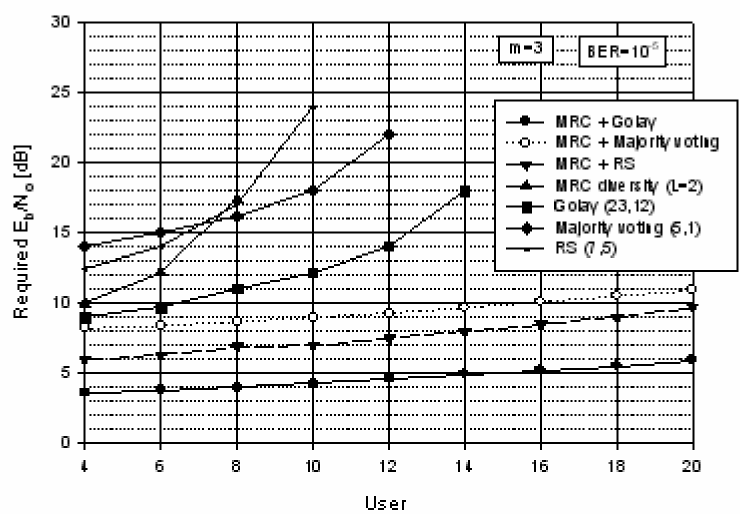


Fig. 5. Comparison of the techniques using diversity combined with coding and repetition transmission (BER=10^{-5}, m =3)

5 Conclusions

In this paper, we analyze the bit error performance improvement of the DS/CDMA BPSK signal with combining repetition transmission, FEC (Forward Error Correction) code and MRC diversity techniques in mobile communication channel which is characterized by m-distribution fading. The results that we have obtained in this paper are that the coding techniques provides more efficient improvement over diversity techniques, but coding techniques are required the adding bandwidth as many coding rate. When it is combined MRC diversity techniques with coding techniques, the amount of improvement is dramatically increased. Bit error performance enhancement is accomplished at the expense of increasing system complexity and cost. Some techniques can yield better results than others, but can be more costly. Therefore, the decision for one or another technique depends on the analysis of cost versus effectiveness. In the following, we shall be going to investigate this trade-off.

References

1. Akira Ogawa, et al.: S/SS/GMSK with Differential Detection over Multipath Rayleigh Fading Channels. IEEE ISSSTA '96, pp. 399-403, Sep. 1996
2. R. E. Eaves and A. H. Levesque: Probability of Block Error for Very Slow Rayleigh Fading in Gaussian Noise. IEEE Trans., on Commun, vol. COM-25, pp, 368-373, Mar. 1977

3. C. E. Sundberg: Block Error Probability for Noncoherent FSK with Diversity for Very Slow Rayleigh Fading in Gaussian Noise. IEEE Trans., on Commun, vol. COM-29, pp, 57-60, Jan. 1981
4. B. Maranda and C. Leung: Block Error Performance of Noncoherent FSK Modulation Rayleigh Fading Channels. IEEE Trans, on Commun, vol. COM-32, pp, 206-209, Feb. 1984
5. F. Adachi and K. Ohno: Block Error Probability for Noncoherent FSK with Diversity Reception in Mobile radio. Electron Lett., vol. 24, pp. 1523-1525, Nov. 1988
6. Y. D. Yao, T. Le-Ngoc and U. H. Sheik: Block Error Probabilities in A M-distribution Fading Channel. IEEE Trans., on Commun., vol. COM-29, pp, 130-133, 1993
7. H. Suzuki: A Statistical Model for Urban Radio Propagation. IEEE Trans. Commun., vol. COM-25, pp. 673-680, July 1977
8. U. Charash: Reception through M-distribution Fading Multipath Channel with Random Delays. IEEE Trans., Commun., vol. COM-27, pp. 657-670, Apr. 1979
9. K. Ben Letaief and M. Hamdi: Efficient Simulation of CDMA Systems in Wireless Mobile Communications. IEEE GLOBECOM'95, pp. 1799-1803, Nov. 1995
10. J. G. Proakis, Digital Communications: New York, McGraw-Hill. 1983

Performance Evaluation of Convolutional Turbo Codes in AWGN and ITU-R Channels

Seong Chul Cho[1], Jin Up Kim[1], Jae Sang Cha[2], and Kyoung Rok Cho[3]

[1] ETRI, Mobile Telecommunication Research Division,
161 Gajeong-dong Yuseong-gu, Daejeon, 305-350, Korea
{sccho, jukim}@etri.re.kr
http://www.etri.re.kr
[2] Seokyeong University, Department of Information and Communication Engineering,
16-1 Jungnung-dong, Sungbuk-gu, Seoul, 136-704, Korea
chajs@skuniv.ac.kr
[3] Chungbuk National University, Department of Computer
and Communication Engineering,
48 Gaeshin-dong, Cheongju-si, Chungcheongbuk-do, 361-763, Korea
krcho@cbu.ac.kr

Abstract. In this paper, the performance of a non-binary convolutional turbo codes is evaluated through computer simulations. Especially, The influence of the various frame sizes and code rate are discussed. Also, in this paper the symbol-by-symbol MAX-Log-MAP algorithm is derived for this coding scheme. We present simulation results for the performance of the non-binary convolutional turbo coded system with QPSK, 16QAM, and 64QAM over an AWGN and the ITU-R channel models.

1 Introduction

In the wireless mobile communication systems, a channel coding scheme is essential for improving reliability of communication. The turbo codes proposed by Berrou et al. [1] are a recent development in the field of error control coding. It has been confirmed that the turbo code achieves near Shannon capacity performance under AWGN. Recently, many studies are continued for applying this powerful coding technique to the wireless communication systems, and significant improvements in the performance of the turbo code are shown even in Rayleigh fading channel [2],[3],[4]. The turbo codes involve iterative decoding of concatenated codes separated by an internal interleaver. Depending upon the codes used for concatenation, turbo codes are classified as convolutional turbo codes and block turbo codes [5].

In the DVB-RCS (Digital Video Broadcasting standard for Return Channel via Satellite) [6] standard and the IEEE 802.16a [7], double-binary convolutional turbo code was adopted due to its outstanding performance. In recent years, the non-binary convolutional turbo code has attracted much attention. Accordingly, we investigate the performance of non-binary turbo code over an AWGN and the ITU-R channel environments. And the symbol-by-symbol Max-Log-MAP algorithm is derived for the iterative decoding.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 695 – 703, 2005.

The organization of this paper is as follows. Section 2 presents the structure of convolutional turbo code considered in this paper. In section 3, the MAP and the Max-Log-MAP iterative decoding algorithm are discussed. Section 4 provides the simulation results and finally, our investigations are summarized and concluded in section 5.

2 Convolutional Turbo Code

The CTC encoder, including its constituent encoder, is depicted in fig. 1. It uses a double binary CRSC (Circular Recursive Systematic Convolutional) code. The bits of the data to be encoded are alternatively fed to A and B, starting with the MSB of the first byte being fed to A, followed by the next bit being fed to B. The encoder is fed blocks of k bits or N couples (k=2N bits), where k is a multiple of 8 and N is a multiple of 4. The polynomials defining the connections are described in octal and symbol notations as follows:

- For the feedback branch : $1+D+D^3$
- For the Y parity bit : $1+D^2+D^3$
- For the W parity bit : $1+D^3$

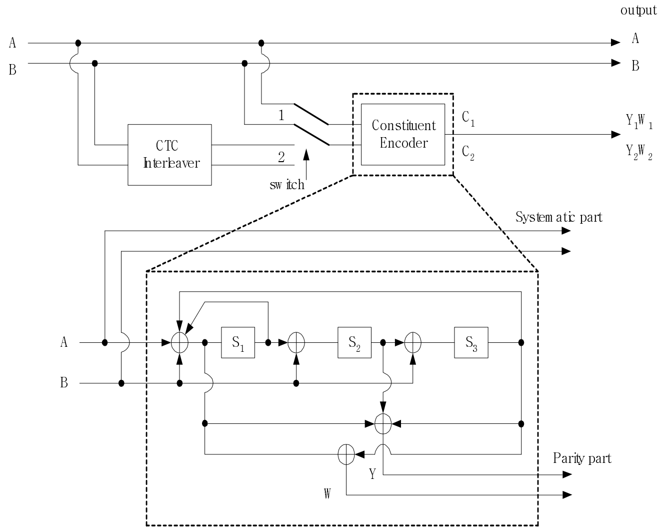


Fig. 1. Encoder Block Diagram

First, the encoder (after initialization by the circulation state S_{C_1}) is fed the sequence in the natural order (switch 1 in Figure 7) with incremental address $i=1, 2, \cdots, N$. This first encoding is called C$_1$ encoding. Then the encoder (after initialization by the circulation state S_{C_2}) is fed by the interleaved sequence (switch 2 in Figure 7) with incremental address $i=1, 2, \cdots, N$. This second encoding is called C$_2$ encoding.

In this system, the overall code rate of CTC is 1/3. The code rate is selected to be 1/3, 1/2, 2/3, or 5/6. The relevant code rates are achieved through selectively deleting the parity bits Y and W. The puncturing patterns are shown in table 1.

Table 1. Puncturing Pattern

Coding Rate	Y						W	
	0	1	2	3	4	5	0	1
1/3	1	1					1	1
1/2	1	1					0	0
2/3	1	0	1	0			0	0
5/6	1	0	0	0	0	1	0	0

The CTC internal interleaver requires the parameters P_0, P_1, P_2, P_3, shown in table 2.

Table 2. CTC Interleaver Parameters

N	P_0	P_1	P_2	P_3
144	17	74	72	2
192	11	96	48	144
240	17	120	60	180
480	17	240	120	360
720	13	360	180	540
2400	29	1200	600	1800

The two-step interleaver shall be performed by:

Step 1 : Switch alternate couples

for $j = 1, 2, \cdots, N$

$if\,(j_{\mathrm{mod}\,2} == 0)\ let\ (B, A) = (A, B)$

Step 2 : $P_i(j)$

The function $P_i(j)$ provides the interleaved address i of the considered couple j.

for $j = 1, 2, \cdots, N$

$switch\ \ j_{\mathrm{mod}\,4}$:

$case\ \ 0: \ i = (P_0 \cdot j + 1)_{\mathrm{mod}\,N}$

$case\ \ 1: \ i = (P_0 \cdot j + 1 + N/2 + P_1)_{\mathrm{mod}\,N}$

$$case\ 2: i = (P_0 \cdot j + 1 + P_2)_{\mathrm{mod}\,N}$$

$$case\ 3: i = (P_0 \cdot j + 1 + N/2 + P_3)_{\mathrm{mod}\,N}$$

The state of the encoder is denoted S with the value calculated by $S = 4 \times s_1 + 2 \times s_2 + s_3, 0 \le S \le 7$. The circulation state S_{C_1} and S_{C_2} are determined by the following operations:

- Initialize the encoder with state 0.
- Encode the sequence in the natural order for determination of S_{C_1} or in the interleaved order for determination of S_{C_2}.
- In both cases the final state of the encoder is S^0_{N-1}.
- According to the length N of the sequence, use table 3 to find S_{C_1} or S_{C_2}.

Table 3. Circulation State Lookup Table (S$_c$)

$N_{\mathrm{mod}\,7}$	S^0_{N-1}							
	0	1	2	3	4	5	6	7
1	0	6	4	2	7	1	3	5
2	0	3	7	4	5	6	2	1
3	0	5	3	6	2	7	1	4
4	0	4	1	5	6	2	7	3
5	0	2	5	7	1	3	4	6
6	0	7	6	1	3	4	5	2

3 Decoding Algorithm

3.1 MAP Algorithm

Let S_k be the encoder state at time *k*. The information symbol d_k is associated with the transition from time *k-1* to time *k*. The trellis states at stage *k-1* and at stage *k* are indexed by the integers S_{k-1} and S_k, respectively. When the MAP algorithm is adopted for every constituent decoder, the soft output each decoded symbol d_k is determined from the log-likelihood ratio as follows:

$$\Lambda(d_k) = \ln \frac{P_r(d_k = i | observation)}{P_r(d_k = 0 | observation)} \tag{1}$$

Where $P_r(d_k = i | observation),\ i = 0,1,2,3$ is the APP (A Posteriori Probability) of the information symbol d_k, and the *observation* is the sequence applied to the turbo decoder at the receiver side. This sequence is denoted $R_1^N = \{R_1 ... R_k ... R_N\}$, where $R_k = \{x_k, y_k\}$. The APP of the information symbol d_k is derived from the joint probability of Eq. (2). Therefore, the APP of d_k becomes Eq. (3).

$$\lambda_k^i(S_k) = P_r\{d_k = i, S_k \mid R_1^N\} \tag{2}$$

$$\Pr\{d_k = i \mid R_1^N\} = \sum_{S_k} \lambda_k^i(S_k) \tag{3}$$

Finally, we can write the LLR for d_k as follows:

$$\Lambda(d_k) = \ln \frac{\sum_{S_k}\sum_{S_{k-1}} \gamma_k^i(R_k, S_{k-1}, S_k)\cdot\alpha_{k-1}(S_{k-1})\cdot\beta_k(S_k)}{\sum_{S_k}\sum_{S_{k-1}} \gamma_k^0(R_k, S_{k-1}, S_k)\cdot\alpha_{k-1}(S_{k-1})\cdot\beta_k(S_k)} \tag{4}$$

In Eq.(4), α_k and β_k are the probability of forward and backward recursion, respectively. And γ_k^i is the branch transition probability.

$$\alpha_k(S_k) = \frac{\sum_{S_{k-1}}\sum_{i=0}^{3} \gamma_k^i(R_k, S_{k-1}, S_k)\cdot\alpha_{k-1}(S_{k-1})}{\sum_{S_k}\sum_{S_{k-1}}\sum_{i=0}^{3} \gamma_k^i(R_k, S_{k-1}, S_k)\cdot\alpha_{k-1}(S_{k-1})} \tag{5}$$

$$\beta_k(S_k) = \frac{\sum_{S_{k+1}}\sum_{i=0}^{3} \gamma_k^i(R_{k+1}, S_k, S_{k+1})\cdot\beta_{k+1}(S_{k+1})}{\sum_{S_k}\sum_{S_{k+1}}\sum_{i=0}^{3} \gamma_k^i(R_{k+1}, S_k, S_{k+1})\cdot\alpha_k(S_k)} \tag{6}$$

$$\begin{aligned}\gamma_k^i(R_k, S_{k-1}, S_k) = q(d_k = i \mid S_k, S_{k-1})\cdot p(x_k \mid d_k = i)\cdot \\ p(y_k \mid d_k = i, S_k, S_{k-1})\cdot P_r\{S_k \mid S_{k-1}\}\end{aligned} \tag{7}$$

3.2 Max-Log-MAP Algorithm

First, find the logarithm of the branch metrics as

$$\begin{aligned}\overline{\gamma_k^i}((x_k, y_k), S_{k-1}, S_k) &= \ln \gamma_k^i((x_k, y_k), S_{k-1}, S_k) \\ &= \frac{2x_k b_x(i, S_{k-1}, S_k)}{N_0} + \frac{2y_k b_y(i, S_{k-1}, S_k)}{N_0} \\ &\quad + \ln P_r\{S_k \mid S_{k-1}\} + K\end{aligned} \tag{8}$$

Next, compute $\alpha_k(S_k)$ and $\beta_k(S_k)$ as

$$\begin{aligned}\overline{\alpha_k}(S_k) &= \ln \alpha_k(S_k)\\ &\approx \max_{(S_{k-1},i)}(\overline{\gamma_k^i}((x_k,y_k),S_{k-1},S_k)+\overline{\alpha_{k-1}}(S_{k-1}))\\ &\quad - \max_{(S_k,S_{k-1},i)}(\overline{\gamma_k^i}((x_k,y_k),S_{k-1},S_k)+\overline{\alpha_{k-1}}(S_{k-1}))\end{aligned} \tag{9}$$

$$\begin{aligned}\overline{\beta_k}(S_k) &= \ln \beta_k(S_k)\\ &\approx \max_{(S_{k+1},i)}(\overline{\gamma_k^i}((x_{k+1},y_{k+1}),S_k,S_{k+1})+\overline{\beta_{k+1}}(S_{k+1}))\\ &\quad - \max_{(S_k,S_{k+1},i)}(\overline{\gamma_k^i}((x_{k+1},y_{k+1}),S_k,S_{k+1})+\overline{\alpha_k}(S_k))\end{aligned} \tag{10}$$

Therefore, the log-likelihood ratio becomes

$$\begin{aligned}\Lambda(d_k) &\approx \max_{(S_k,S_{k-1})}(\overline{\gamma_k^i}((x_k,y_k),S_{k-1},S_k)+\overline{\alpha_{k-1}}(S_{k-1})+\overline{\beta_k}(S_k))\\ &\quad - \max_{(S_k,S_{k-1})}(\overline{\gamma_k^0}((x_k,y_k),S_{k-1},S_k)+\overline{\alpha_{k-1}}(S_{k-1})+\overline{\beta_k}(S_k))\\ &\mathit{for}\ i=1,2,3\end{aligned} \tag{11}$$

4 Simulations

In order to analyze the performance, we consider AWGN, ITU-R pedestrian-B, and vehicular-A channel models [8]. Monte Carlo simulations are carried out on these channel environments with respect to increasing signal-to-noise ratio. The parameters employed to analyze the performance of using convolutional turbo code are listed in table 4.

Table 4. Parameters for Simulation

Parameters		Value
Channel model		AWGN, ITU-R
Modulation		QPSK, 16QAM, 64QAM
Channel coding		CTC 1/3, 1/2, 2/3, 5/6 (Fig. 1)
Decoding algorithm		Max-Log-MAP
Number of iteration		8
Mobility	3km/h	Pedestrian-B
	60km/h	Vehicular-A

Respectively, fig. 2 and fig. 3 show the BER performances of the CTC with QPSK-1/2 rate and 64QAM-5/6 rate for different lengths. For QPSK, data block sizes of 288, 480, 960, and 4800 bits are simulated. And for 64QAM, data block sizes of 480, 960 1440, and 4800 bits are considered.

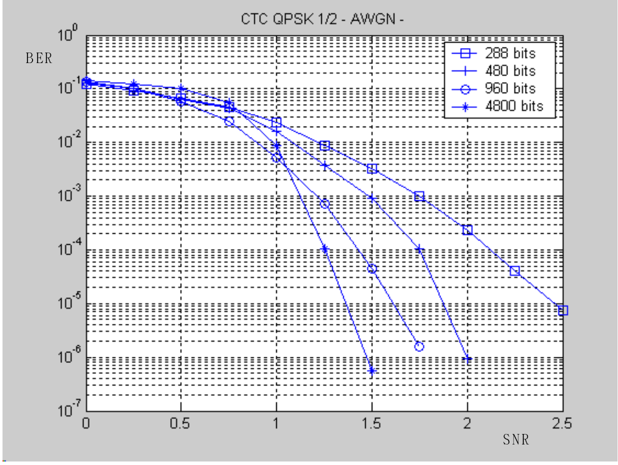


Fig. 2. Performance of the double-binary CTC with QPSK-1/2 rate

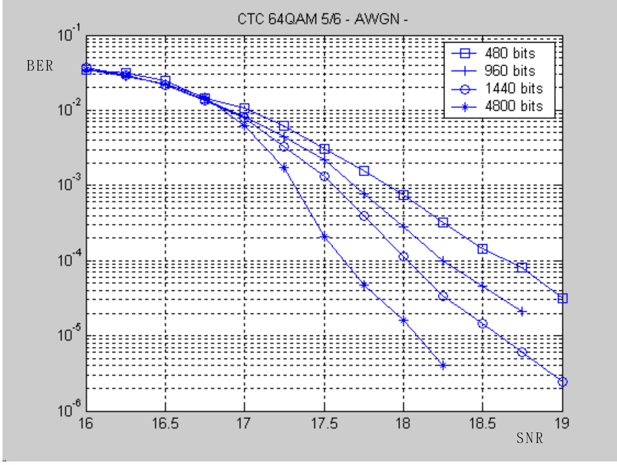
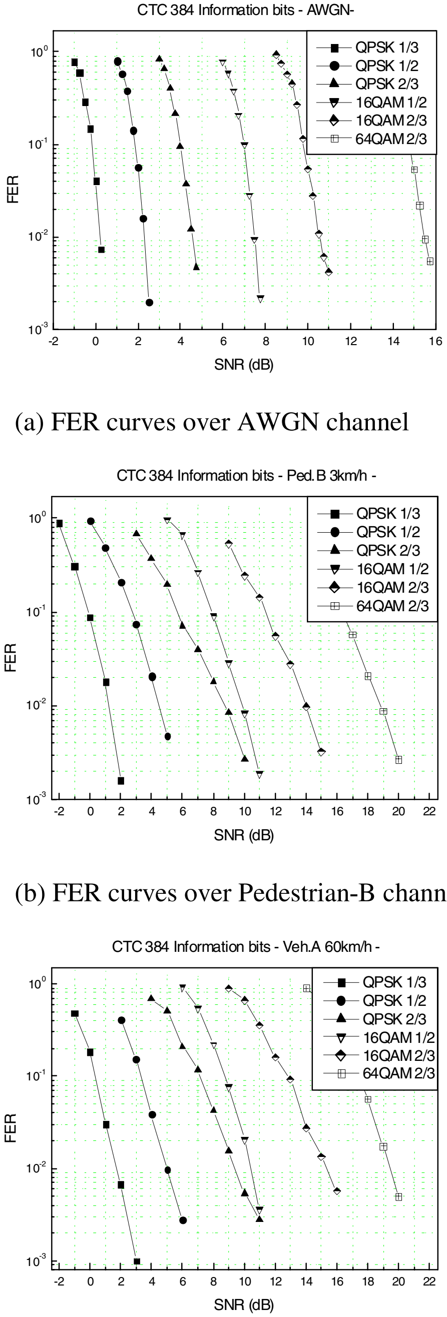


Fig. 3. Performance of the double-binary CTC with 64QAM-5/6 rate

In fig. 4, the different modes are analyzed in terms of FER performance. The simulation modes are QPSK-1/3, QPSK-1/2, QPSK-2/3, 16QAM-1/2, 16QAM-2/3, and 64QAM-2/3.

In fig. 4, each figure shows the FER performance of CTC with the data block size of 384 bits over an AWGN and the ITU-R channel models in the case of the perfect channel estimation. As a main result, we can see that as the modulation level increase the performance becomes worse. We also see that as the coding rate is high the performance degradation due to Rayleigh fading channel is larger. In table 5, we present the required SNR for a FER of 10^{-2}.



(a) FER curves over AWGN channel

(b) FER curves over Pedestrian-B channel

(c) FER curves over Vehicular-A channel

Fig. 4. Performance of the double-binary CTC for block size 384 bits

Table 5. Required SNR for a FER of 10^{-2}

Modulation & Coding Rate	Data bits	AWGN	Ped.B 3km/h	Veh.A 60km/h
QPSK 1/3	384	-0.60	0.60	1.75
QPSK 1/2	384	1.65	3.85	5.00
QPSK 2/3	384	3.95	7.70	9.40
16QAM 1/2	384	6.95	9.40	10.40
16QAM 2/3	384	9.75	13.65	15.35
64QAM 2/3	384	15.00	18.85	19.45

(dB)

5 Conclusion

In this paper, the performances of double-binary convolutional turbo codes were presented. The BER and FER performances were studied via computer simulations by various data block size and the modulation and coding schemes. For CTC decoding, the symbol-by-symbol Max-Log-MAP algorithm was derived and employed. It is confirmed from the simulation results that CTC offers considerable performance improvement.

The non-binary convolutional turbo code has been strongly recommended as error correction code for high speed mobile communication systems, so we expect that these results can be useful as basic data for the implementation of convolutional turbo coded system

References

1. C. Berrou, A. Glavieux, and P. Thitimajshima: Near Shannon Limit Error-correcting Coding and Decoding: Turbo-codes. IEEE ICC’93, Genva, Switzerland, vol. 2. pp. 1064-1070, May 1993
2. E. K. Hall and S. G.. Wilson: Design and Analysis of Turbo Codes on Rayleigh Fading Channels. IEEE J. Select. Areas Commun., vol. 16, No. 2, pp. 160-174, Feb. 1998
3. P. Frenger: Turbo Decoding on Rayleigh Fading Channels with Noisy Channel Estimates. Proc. IEEE VTC’99, pp. 884-888, May. 1999
4. P. Komulainen, K. Pehkonen: Performance Evaluation of Superorthogonal Turbo Codes in AWGN and Flat Rayleigh Fading Channels. IEEE Journal on Selected Areas in Communications, vol. 16, no. 2, pp. 196-205, Feb. 1998
5. S. Dave, J. Kim, and S. C. Kwatra: An Efficient Decoding Algorithm for Block Turbo Codes. IEEE Transactions on Communications, vol. 49, No. 1, pp. 41-46, Jan. 2001
6. European Telecommunications Standards Institute (ETSI), Digital Video Broadcasting (DVB): Interaction cahnnel for satellite distribution systems, http://www.etsi.org, EN 301 790, Dec. 2000
7. IEEE Standards 802.16a, Part 16: Air Interface for Fixed Broadband Wireless Access Systems- Amendment 2: Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11GHz, Apr. 2003
8. Recommendation ITU-R M.1225; Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000, 1997

Adaptive Modulation Based Power Line Communication System

Jong-Joo Lee[1], Jae-Sang Cha[2,*], Myong-Chul Shin[1], and Hak-Man Kim[3]

[1] Dept. of Electronic and Electrical Eng. Sungkyunkwan Univ. Suwon, Korea
[2] Dept. of Information and Communication Eng. Seokyeong Univ. Seoul, Korea
[3] KERI, 28-1 Sengju-dong, Changwon, Kyung-Nam 641-120, Korea
chajs@skuniv.ac.kr

Abstract. In this paper, we present an adaptive modulation/demodulation methods based on each load fluctuation after measuring and modeling of noise pattern due to the various load fluctuations of PLC (power line communication) transmission channel. Additionally, we proposed adaptive modulation for PLC using Decision Making Algorithm that could select the optimum modulation/demodulation methods adaptively according to the various load fluctuation. And we certified the availability of adaptive modulation for PLC using the computer simulation and hardware implementation. The newly proposed adaptive modulation can be widely used to attain the high performance PLC implementation in the severe noise environment due to the load fluctuations.

Keywords: Power line communication; Adaptive Modulation; Decision Making.

1 Introduction

Recently Power Lines Communications utilizing except supplying the power by increasing of the smart electronic and electrical devices and constructing home network are studied and applied as useful scheme in home network field. However those are affected by load fluctuation properties different from the general phone lines or cable since those are not communication lines but for transmission the power. Hence power lines have poor transmission line properties with occurred noise of communications on transmission lines at all times [1].

Nevertheless, studies of PLC with transmission line properties are preceded still as beginning. Also various modulation/demodulation techniques are presented, but the performances of them hadn't being compared and analyzed. And most simulations used transmission line properties without load fluctuations properties [2], [3].

Therefore, power line communication techniques based on conventional uniformed modulation/demodulation can't be maintained flexible system properties that have resistance property for various load fluctuations on power

[*] Corresponding author.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 704 – 712, 2005.

lines. Also study with improved performance is required efficient transmit information, calculating of optimized transmit capacity on transmission lines, and various transmission property as key points of communication techniques

Hence, in this paper we will propose and certify the adaptive modulation technique applying optimal modulation/demodulation scheme according to various power lines with load fluctuations. At first, after measuring and analyzing noise properties in each case of load fluctuations, we applied various modulation/demodulation based in each case of load fluctuations. By these simulations, we will calculate transmission capacity and analysis BER (Bit Error Rate) performances to get optimal modulation/demodulation methods. Moreover we will certify the availability of the adaptive modulation as proving the property for maintaining flexible communication performance by applying the optimum modulation/demodulation methods of each property of resulted load fluctuations to the DMA (Decision Making Algorithms).

2 Measurement of Noise Property According to Load Fluctuations

In this section, we selected the loads by considering various home appliance properties in home network power line communication environment and classified them according to electric properties. We will also use the results classified in each case of changes and types of transmission lines by load fluctuations as data for optimal modulation/demodulation scheme.

Especially, we classified the load types as passive loads that consist of passive components and active loads that consist of power components to get noise environment of PLC according to load fluctuations. Passive loads are classified as resistive, inductive, and capacitive of loads according to the properties.

In the mean time, active loads are classified as power supply devices used actually. And we measured each load fluctuation by selecting dimmer loads controlling power through phase control of SCR (Silicon Controlled Rectifier) or Thyristor or SMPS (Switching Mode Power Supply) which is a typical power equipment occurring fast switching noises and harmonics. Following Table1 shows measured load fluctuation properties case by case.

From analysis of measured load fluctuation data in the case of passive load, aperiodic impulse noise is happened in transient state such as cut-off, injection. And we conformed that periodic impulse is happened in the case of active loads, according to power control system.

3 Selection of Optimal Modulation/Demodulation Methods for Load Fluctuation

In this paper, superior modulation/demodulation methods were selected for selecting optimal modulation / demodulation methods at the base of measured load model that is described in Table1 as follow [4], [5], [6].

1) Spread Spectrum (SS) modulation/demodulation
   a) M-ary Spread Spectrum (SS) method
   b) CCK (Complementary Code Keying) method
2) Multi-Carrier modulatipon/demodulation
   a) OFDM (Orthogonal Frequency Division Multiplexing)

Load model based on measured data in multiple communication methods selected earlier is used, and we analyzed and compared communication capacity and BER (BER; Bit Error Rate) performance of proposed system.

Figure1 showed E_b/N_0 property of each modulation method according to change of load models as follow.

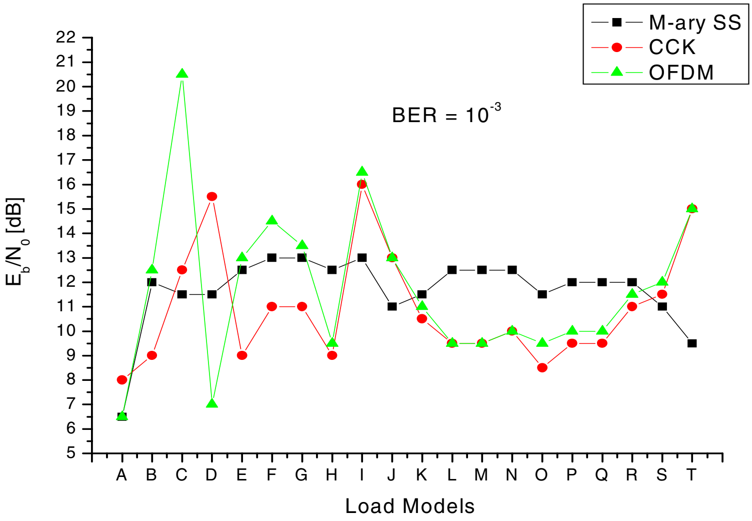


Fig. 1. The comparison of communication efficiency by change of load models

Values of Figure1 are E_b/N_0 of fixed BER as 10^{-3} according to each modulation method. Performance becomes more superior when the values are smaller.

As seeing Figure 1, in case of the C model that is inductivity load model has 11.5dB of E_b/N_0 using M-ary SS, and has 20.5dB of E_b/N_0 using OFDM. Difference of the BER performances between modulation methods is 9dB at most. Also difference of E_b/N_0 of both modulation methods is 8.5dB in the case of D model simulated as resistance load. 4dB differenced for E model simulated as resistance load and inductivity load. For T model simulated load attenuation of SMPS is 5.5dB. So those differences were considerable deviation.

Therefore we conformed that optimal communication performance can be maintained when modulation/demodulation methods with superior communication performance were selected according to multiple communication methods.

4 Presentations of the Decision Making Algorithms for the Adaptive Modulation

Various load devices exist on home network power lines, and these are operated randomly in time domain. So, home network power line communications are affected by various noises. And it is hard to satisfy that estimating those noises and degrading interference.

In this paper we will investigate and analyze the Dempster-Shafer theory that is a typical one of conventional uncertainty inference methods to implement Decision Making Algorithms of the adaptive modulation considered uncertain PLC environments and will prove the properties from specific cases [7], [8].

First, Shafer marked measure of trust about some hypothesis H as following interval than probability that is expressed by a number.

$$[Bel(H),\ Pl(H)] \tag{1}$$

This is called by 'evidence interval'. Here, Bel shows the information (belief) receiving random hypothesis by given evidence. And Pl shows possibility (plausibility) based on the evidences that hypothesis exists not deny, that is trusted at most.

As seeing the following picture, *Bel* has dimensions from 0 (there is no entirely evidence) to 1 (there is conclusive evidence), and Pl has [0, 1] as it since it can be defined as follow.

$$Pl(H) = 1 - Bel(H) \tag{2}$$

Shafer's idiosyncrasy can express the information for amount of evidence that could not know when Bayes arrangement was used [9].

Assuming that communication performances are decreased while virtual PLC environment is selected and data is transmitted and received. Here, we could reasoning the factor that makes changeable various communication channel environment assignment probability values. That is, assuming that motor loads and SMPS (Switching Mode Power Supply) exist, 0.5 probability value can be assigned in each load.

But, this way does not make to recognize whether assign as much as 0.5 for each case with some evidences we have. In this case each probability value can be expressed as [0,1], if Shafer's expression was used. That is we have perfect uncertainty with no supporting evidence and denying evidence. Especially this uncertainty is happened when the channel environment is varying or a lot of interference components exist on PLC. It is hard to get stochastic calculations or statistics or making the properties according to these interference elements that are combined. Therefore, in this paper we supposed power line communication environment as "Unforeseeable uncertain situation" to using reasoned techniques that are based on uncertainty and we collected the information. From the information, we get the result that Decision Making Algorithms which can maintain the optimum communication performance in each case. The algorithms were written based on Dempster-Shafer theory and applied to the adaptive modulation for PLC. Operation characteristics of the adaptive modulation scheme are follows.

5 Estimation of Adaptive Modulated PLC Operation Characteristic Based on DMA

We compared BER Performance by the adaptive modulated PLC and noise property of load models that was divided in table 1.

Table 1. Noise quality of Load model in Power line communication

Dividing	Load models	Noise quality		Remarks
A	AWGN	-		Noise
B	Resistive	In Phase	Steady State	Passivity Load
C	Inductive	Current delay		
D	Capacitive	Voltage delay		
E	Resistive + Inductive	Non-linear current increase	Transient State	Composition Passivity Load
F	Resistive + Capacitive	Impulsive noise		
G	Resistive + Inductive + Capacitive	Non-linear current increase & Impulsive noise		
H	Radiator	In Phase	Steady State	Resistive Load
I	SMPS #1	Periodic impulsive noise		Activity Load
J	SMPS #3	Periodic impulsive noise		
K	Dimmer	Periodic impulsive noise		
L	Compressor	Voltage delay		Capacitive Load
M	Vacuum cleaner	Current delay		Inductive Load
N	Compressor + SMPS #3	Periodic impulsive noise		Composition Load
O	Radiator + Dimmer	Periodic impulsive noise		
P	Vacuum cleaner + Dimmer	Periodic impulsive noise		

Table 1. (*Continued*)

Q	Radiator (Low→High)	Asynchronous impulsive noise	Variable Load
R	SMPS (Low→High)	Periodic impulsive noise	
S	SMPS (#5→#5+ #1)	Asynchronous impulsive noise + Periodic impulsive noise	
T	SMPS (#6→#6 - #1)	Asynchronous impulsive noise + Periodic impulsive noise	

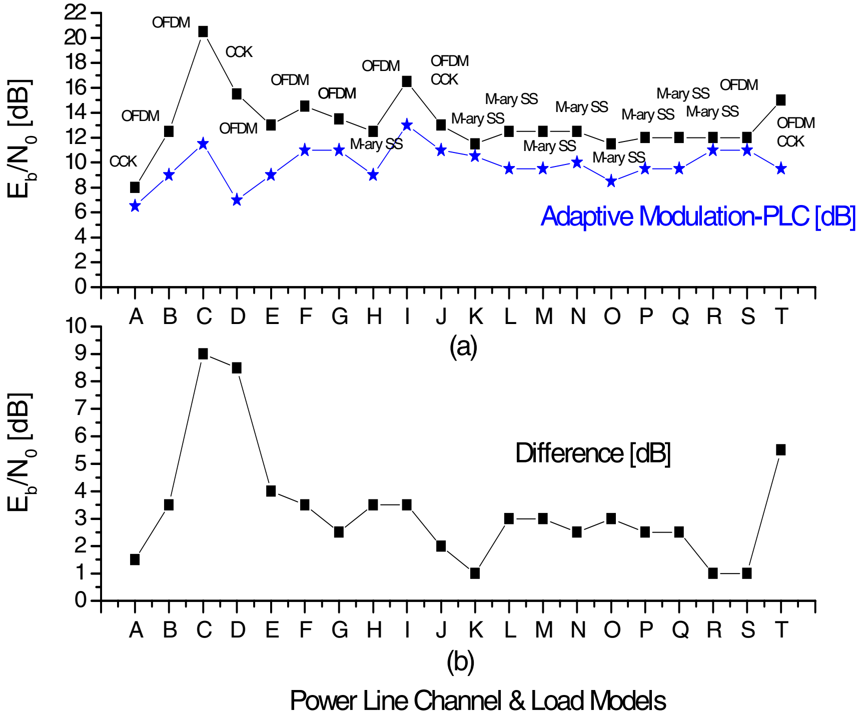


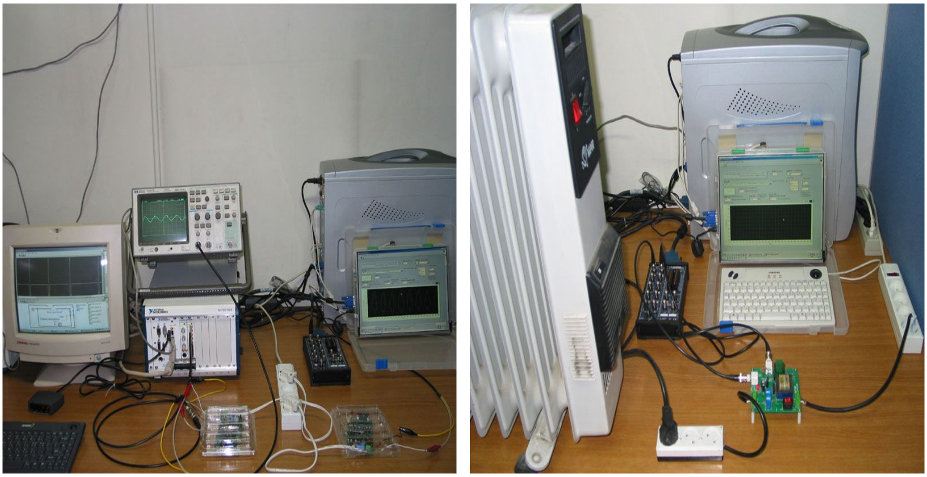
Fig. 2. The comparison of communication efficiency by change of power line channel

In Figure 2, (a) compared modulation/demodulation methods with lower communication performance in optimal modulation/demodulation methods and communication methods according to load fluctuations. In Figure2, (b) showed differences of E_b/N_0 between optimal adaptive modulation/demodulation methods on suggested adaptive modulation and conventional fixed modulation/ demodulation methods in same channel environment.

From the difference of E_b/N_0, we proved superior communication performance is maintained in various load fluctuations using modulation/demodulation methods different from conventional modulation/demodulation methods. This fact is also showed from the examples of components of base band hardware and simulating result in Figure 3, 4.

Table 2. Specification Example for the adaptive modulation PLC

Item	Content
Modulation	M-ary SS, CCK, OFDM
Channel Model	A~T Load models (Table 1. Ref.)
BER	10^{-3}
Eb/N_0	0~22 dB

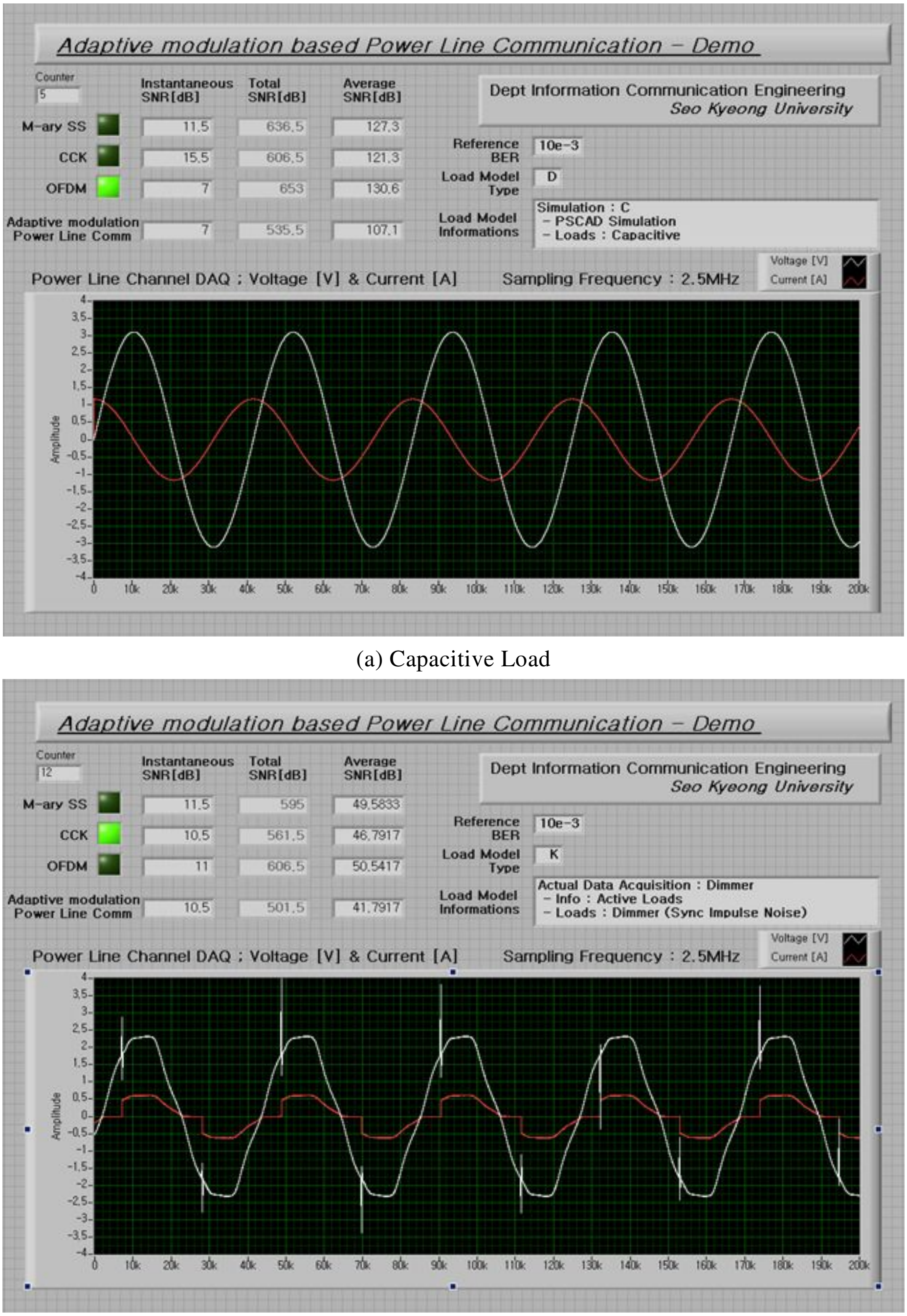


(a) Base-band hardware structure block

(b) Load connection quality of demo module PLC

Fig. 3. Hardware structure of adaptive modulation PLC

As you see, in case of operating system by selecting optimal communication methods from Decision Making algorithms has 22dB~38dB of E_b/N_0 different from conventional operating system applied to uniformly modulation/ demodulation methods. From the results we certified superiority of the proposed adaptive modulated PLC.



Adaptive modulation based Power Line Communication – Demo

Counter: 5

	Instantaneous SNR[dB]	Total SNR[dB]	Average SNR[dB]
M-ary SS	11.5	636.5	127.3
CCK	15.5	606.5	121.3
OFDM	7	653	130.6
Adaptive modulation Power Line Comm	7	535.5	107.1

Dept Information Communication Engineering
Seo Kyeong University

Reference BER: 10e-3

Load Model Type: D

Load Model Informations: Simulation : C
- PSCAD Simulation
- Loads : Capacitive

Power Line Channel DAQ ; Voltage [V] & Current [A]   Sampling Frequency : 2.5MHz

(a) Capacitive Load

Adaptive modulation based Power Line Communication – Demo

Counter: 12

	Instantaneous SNR[dB]	Total SNR[dB]	Average SNR[dB]
M-ary SS	11.5	595	49.5833
CCK	10.5	561.5	46.7917
OFDM	11	606.5	50.5417
Adaptive modulation Power Line Comm	10.5	501.5	41.7917

Dept Information Communication Engineering
Seo Kyeong University

Reference BER: 10e-3

Load Model Type: K

Load Model Informations: Actual Data Acquisition : Dimmer
- Info : Active Loads
- Loads : Dimmer (Sync Impulse Noise)

Power Line Channel DAQ ; Voltage [V] & Current [A]   Sampling Frequency : 2.5MHz

(b) Dimmer Load

Fig. 4. An Adaptive Modulation PLC Demo Simulation program

6 Conclusions

In this paper, we get the results from Modulation/demodulation methods that have optimal performances of each load fluctuation after measuring and modeling noise property by load fluctuation of transmission channel. Also we proposed the adaptive modulation Power Line Communication systems using selectable Decision Making method adaptively and certified the availability by implementing testbed (demo) and simulating. The proposed adaptive modulation PLC has high performance with load fluctuation property applying optimal modulation/ demodulation methods. Sine we applied future-oriented optimal PLC scheme, we expect it is utilized usefully and practical use in related filed.

Acknowledgement

This work was supported by KESRI (R-2003-B-289), which is funded by MOCIE (Ministry of Commerce, Industry and Energy).

References

1. Ferreira, H.C. Grove, H.M. Hooijen, O. Han Vinck, A.J.: Power Line Communi-cation: An Overview. IEEE AFRICON 4th, Vol.2, pp.558-563 199
2. Manfred Zimmermann and Klaus Dostert: Analysis and Modeling of Impulsive Noise in Broad-Band Power Line Communications. IEEE Trans. Electromagnetic Compatibility, Vol.44, NO. 1, pp. 249-258, FEBRUARY 2002
3. L. T. Tang, P. L. So, Member, IEEE, E. Gunawan, Y. L. Guan, S. Chen, and T. T. Lie: Characterization and Modeling of In-Building Power Lines for High-Speed Data Transmission. IEEE transactions on power delivery, VOL.18, No. 1, January 2003
4. Tadahiro WADA, Takaya YAMAZATO, Masaaki KATAYAMA, Akira OGAWA: A New Mary Spread-Spectrum Multiple Access Scheme in the Presence of Carrier Frequency Offset. IEICE Trans, Fundamentals, vol. E79-A, No.9, September, 1996
5. Carl Andren: CCK Modulation Delivers 11Mbps for High Rate IEEE 802.11 Extension. Wireless Symposium, 1999
6. Tlili. F, Rouissi. F, Ghazel. A: Precoded OFDM for Power Line Broadband Communication. IEICE Trans, Fundamentals, vol. E79-A, No.9, September 1996
7. P. Dempster: A Generalization of Bayesian Inference. Journal of the Royal Statistical Society, Series B, 30, 1968
8. G. Shaper: A Mathematical Theory of Evidence. Princeton University Press, Princeton, 1976

A Novel Interference-Cancelled Home Network PLC System Based on the Binary ZCD-CDMA

Jae-Sang Cha[*], Myong-Chul Shin[**,©], Jong-Joo Lee[**]

[*]Dept. of Information and Communication Eng. Seokyeong Univ.
16-1 Jung-nung dong Sungbuk-ku, Seoul, 136-704, Korea
Dept. of Electronic and Electrical Eng. Sungkyunkwan Univ. Suwon, Korea
mcshin@yurim.skku.ac.kr

Abstract. The transmission channel of Home network PLC (power line communication) are characterized by various noise components and delayed waves generated by load fluctuation and multi-path transmission. In this paper, a novel Interference-cancelled Home Network CDMA-PLC system based on the binary ZCD (zero correlation duration) spreading code are proposed as one solution to overcome the previous problems. The properties of the proposed ZCD-PLC systems are effective for MPI(multi-path interference) and MAI (multiple access interference) cancellation in the CDMA-PLC (code division multiple access-PLC) systems. By BER performance simulation, we certified the availability of proposed ZCD-CDMA-PLC system.

1 Introduction

Recently, industry for constructing home-network connected with information electronic appliances has been enlarged than before. At the aftermath of that situation, discussion about PLC technology for home-network at local area is going in progress in academic and industrial circles. However, in case of PLC, the transmission line for communication is not leased and load fluctuations related to power consumption arises very frequently. Besides, PLC system has many poor conditions like various noise, reflected wave caused by time varying line impedance and multi-path transmission.

But, the progress of study for PLC has been in the early stage until now, so there have been only modeling and trial to analyze efficiency about modeling systems and BER (Bit Error Rate) emphasizing simply noise characteristics out of impulse response characteristics of transmission lines [1], [2]. Therefore, in this paper, echo characteristics by reflected waves is contained for channel modeling including impulse noise on the previous PLC channel, at the same time ZCD-PLC system is proposed, that is new modulation-demodulation and multiple access skills to overcome the previous problems, as communication system only for

[©] Corresponding author.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 713 – 722, 2005.

home-networking based on spreading code.

ZCD-PLC system proposed in this paper is a new modulation-demodulation and multiple access skills using CDMA system based on a new spreading code that has consecutive ZCD (zero correlation duration) characteristics. In this paper, first of all, as basic study to apply ZCD-PLC system, property of multi-path will be analyzed very precisely; this is caused by various interference or echo from load fluctuation of PLC channel under home-network system. Also, spreading coding technique with ZCD process that has excellent capability to relax these interferences under home-network system will be explained more clearly. Moreover, we will develop PLC system based on low coherent modulation-demodulation and multiple access system using ZCD-CDMA that carry out CDMA by binary ZCD spreading code, and then through simulation at link level, the utility of this system will be proved by analyzing and estimating BER.

2 Property of PLC Channel Under Home-Network System

In this clause, noise type and interference sources were selected like below thinking over properties of noise and interference sources [3].

1) Background noise

- Background noise (nearly white Gaussian)

2) Narrow band interference

- Narrow band (single tone) interference, mostly generated by broadcast radio transmitters

3) Unexpected noise source (impulse noise)

- Impulse noise, either periodical and synchronous with the mains voltage, or stochastic, from various kinds of switching equipment

4) Interferences between signals by multi-path and attenuation

As mentioned above, PLC channel is different from previous one under wired or wireless system. So, in this paper, we made modeling changes by interference source on the channel and noise or many kinds of types to expect or analyze accurately PLC system for home-network, and then we will put upper results to practical use for noise and interference data to develop PLC system for low coherent home-network based on binary ZCD-CDMA suggested in this paper.

2.1 Analysis for Property of Multi-path Channel Under Home-Network PLC

In this clause, we have analyzed the properties of multi-path channel by simulation, after modeling home-network channel based on simple PLC thinking over problems under above-mentioned system.

And property of Delay Profile will be abstracted by building virtual home-network system based on echo model [4], [5] that has channel on the multi-path.

In this paper, we use PN (Pseudo Noise) spreading code among techniques for analyzing property of channel by abstracting Delay Profile, then induce peak value of auto correlation function (ACF) on PN spreading code by correlation circuit at receiver, apply new skills about analyzing channel for abstracting Delay Profile using technique mentioned above.

We abstracted many channel models for PLC that has characteristic about multi-path through analyzing skill mentioned above, and property of channel model presented in this paper is in table 1.

Delay time is about 0.33 uses at intervals of 0.1 km on the assumption that transmission speed is equal to light velocity in table 1, when transformed to frequency is equal to one cycle delay of 3MHz signal.

In this paper, we apply channel-analyzing skills using PN spreading code to derivate property of multi-path on PLC channel system. Delay Profile characteristic from this simulation is used to demonstrate superiority of PLC system based on binary ZCD-CDMA proposed in this paper with impulse noise.

Table 1. Delay Profile characteristic of each channel model

Model Property	Channel model I	Channel model II	Channel model III	Channel model IV	Channel model V
Total Line Length	1.4 km	2.6 km	7.8 km	6.2 km	8.4 km
Path Number	2	3	4	6	7
User Number	3	4	4	5	5
Delay time per 0.1 km	0.33 sec				
Decrease Num. per 0.1 km	0.95	0.95	0.97	0.97	0.97
Max. Magnitude	Normalization: 1				
Max. delay time	4sec	8sec	26sec	20sec	28sec

3 Spreading Code Using Binary ZCD to Solve Multi-path and Noise Problems

3.1 Binary Spreading Code Having Zero Correlation

When binary series of random periodic N, $S_N^{(x)} = (s_0^{(x)},\cdots,s_{N-1}^{(x)})$ and $S_N^{(y)} = (s_0^{(y)},\cdots,s_{N-1}^{(y)})$ exist, periodic correlation function about shift T is defined like below.

$$\theta_{x,y}(\tau) = \sum_{n=0}^{N-1} s_n^{(x)} s_{(n+\tau,\mathrm{mod}\,N)}^{(y)} \tag{1}$$

Here, when x=y, equation (1) become ACF, when $x \neq y$ that become CCF. Maximum values of side lobe in periodic autocorrelation function and maximum value of periodic interactive correlation function were proved to have theoretical threshold level by Trade off [6]. But it is possible to make binary code which have θ_{as} and θ_{c} zero continuously within specific part around t=0. We define continuous local duration as zero correlation duration (ZCD). It was cleared up in some papers [7], [8], [9], [10] that upper ZCD characteristic establishes sub-synchronous section which doesn't have MAI, especially in up-link of CDMA system.

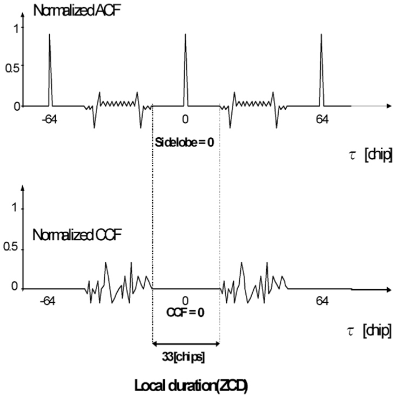


Fig. 1. ZCD characteristic of binary ZCD code

3.2 Binary ZCD Spreading Code

In this paper, we use binary ZCD spreading code that I recommend as generation skill [7] , [8], [9], [10] to configure ZCD-CDMA system. Binary ZCD spreading code is very easy to embody code generator, and has maximum ZCD section as (0.5N+1) chip. ZCD characteristic of 64chip-period binary ZCD spreading code is showed in figure 1. As you see in figure 1, when N=64, we can confirm that property of ZCD, that is, side-lobe of ACF and CCF become zero consecutively for 33-chip applicable to (0.5×64+1), is maintained. In the following, among diverse ZCD codes, for instance, $\left\{S_{64}^{(a)}, S_{64}^{(b)}\right\}$ is expressed, which is binary ZCD code pair of 64-chip used in simulation. Here, + means 1, - means -1.

Ex1) Binary ZCD code pair of periodic 64-chip

$$\left\{ \begin{array}{l} S_{64}^{(a)} = c \;\; d \;\; c \;\; -d \;\; c \;\; d \;\; -c \;\;\; d \;\; c \;\; d \;\; c \;\; -d \;\; -c \;\; -d \;\; c \;\; -d \\ S_{64}^{(b)} = u \;\; v \;\; u \;\; -v \;\; u \;\; v \;\; -u \;\;\; v \;\; u \;\; v \;\; u \;\; -v \;\; -u \;\; -v \;\;\; u \;\; -v \end{array} \right\}$$

$$\begin{array}{l} \mathit{Here},\;\; u = \left(- \;\; - \;\; - \;\; + \right),\;\; v = \left(- \;\; - \;\; + \;\; - \;\; \right), \\ \qquad\quad c = \left(- \;\; + \;\; - \;\; - \right),\;\; d = \left(- \;\; + \;\; + \;\; + \;\; \right) \end{array}$$

4 Modeling ZCD-CDMA PLC for Home-Network Based on Binary ZCD-CDMA

Under transmission channel of actual CDMA-PLC system, MPI (multi-path interference) is generated on the transmission line by impulse noise related to impedance variation from load fluctuation and echo characteristics, at the same time orthogonal characteristics decay is happened inevitably among spreading codes by MAI (multiple access interference) originated from multiple access environment.

In case of applying matched filter about k-th user within Gaussian noise channel at ZCD-CDMA-PLC system, theoretical property of BER [11] is like equation (2).

$$P^{k}(\sigma) = \frac{1}{2} \sum_{e1 \in \{-1,1\}} \cdots \sum_{\substack{ej \in \{-1,1\} \\ j \neq k}} \cdots \sum_{ej \in \{-1,1\}} Q\left(\frac{C_k}{\sigma} + \sum_{j \neq k} e_j \frac{c_k}{\sigma} \beta \rho_{jk} \right) \tag{2}$$

Here, Q(x) is complementary cumulative distribution function related to unit normalization variables. In addition to Gaussian noise, when we express property of BER considering multi-path channel, equation (3) is like below.

$$Q(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{3}$$

In this paper, capability of BER in ZCD-CDMA PLC system doesn't have MAI, because orthogonal characteristics is maintained consecutively for zero correlation duration among spreading codes which were used for multiple access. Therefore, capability of BER under multi-path system by Gaussian noise and echo characteristics can be expressed like equation (4) and (5) below.

$$p^{k}(\sigma) = \frac{1}{2} Q\left(\frac{C_k}{\sigma}\right) \tag{4}$$

$$p^{Fk}(\sigma) = \frac{1}{2}\left(1 - \frac{C_k}{\sqrt{\sigma^2 + C_k}}\right) \tag{5}$$

Because function Q(x) is a simply decreasing function about variable x, BER value of system applying proposed algorithm, if comparing and supposing CDMA-PLC system using Walsh-code which has orthogonal characteristics at only one point establishing synchronous among codes under MAI or MPI system, has always larger value than CDMA-PLC system based on ZCD code.

5 Simulations and Analyzing Capability

Here, as a mean for verifying capability on basis of modeling ZCD-CDMA PLC system mentioned-above, after composing MonteCarlo simulator performed in link-level by Matlab, we have carried out simulation comparing ZCD-CDMP PLC under various simulation mentioned-below with CDMA-PLC based on Walsh code for BER.

5.1 Conditions for BER Simulation

In this simulation, we have carried out simulation by selecting 4 conditions below and related to comparing capability, abstracting BER with ZCD-CDMA-PLC system and Walsh-CDMA-PLC system.

1) Impulse noise surroundings by Gaussian noise and load fluctuation.
2) Multi-path channel by echo characteristics
3) Channel including Gaussian, impulse noise, multi-path simultaneously.
4) Abstracting capability of BER under variable MAI by increasing user.

In figure 2, we applied maximum amplitude V_{pk}=1, normalized unit impulse wave for estimating communication capacity under impulse noise. Every simulation to consider property of impulse period from diverse load fluctuations classifies the period of normalized impulse noise expressed in figure 2.

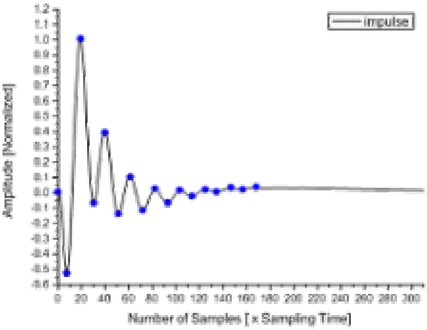


Fig. 2. Normalized impulse noise model

5.2 Simulations and Analyzing Capability

In this clause, for verifying capability of BER in ZCD-CDMA-PLC proposed in clause 4, we took a simulation about capability of BER using computer under transmission channel assumed in clause 5.1, and the results of that simulation are showed in figure 3, 4, 5, 6, 7.

Figure 3 is the case of assuming and simulating patterns about impulse noise from Gaussian noise and load fluctuation, that means it can be verified that if repeating samples of impulse periodically twenty times, capability of BER is satisfied with less than 10^{-3}, but if repeating more than 20 times, is not satisfied with less than 10^{-3}.

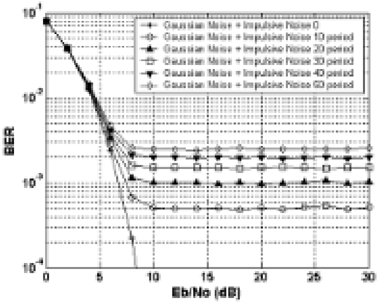


Fig. 3. BER performance comparison in Gaussian noise and impulse noise channel

So, we have to try to improve capability of BER using error correction coding mechanism or specific signal processing skills for problems mentioned above.

In figure 4, we have compared CDMA-PLC system based on Walsh code with CDMA-PLC system based on ZCD code about capability of BER under channel model 4 of table 1. In this figure, when BER value is 10^{-3}, we can't get exact value because E_b/N_0 value of CDMA-PLC system based on Walsh code is divergent. On the other hand, E_b/N_0 value of CDMA-PLC system based on ZCD code is roughly 17dB. We can verify that CDMA-PLC system based on ZCD code under violent multi-path condition is more superior to CDMA-PLC system based on Walsh code.

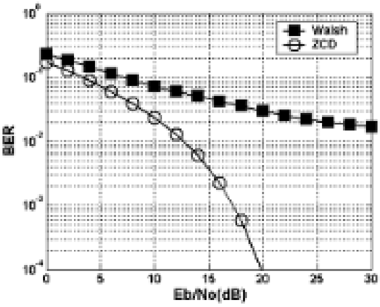


Fig. 4. BER performance comparison in multi-path Channel (Channel model IV)

In the simulation of figure 5, we have thought over Gaussian noise, impulse noise and multi-path channel to reflect actual PLC channel system. As a result of analyzing a property of BER in this figure, E_b/N_0 value of CDMA-PLC system based on ZCD code is approximately 18dB on the basis of 10^{-3}. But we cannot get precise value under same condition because CDMA-PLC system based on Walsh code is divergent. So, we can understand that CDMA-PLC system based on ZCD code under channel condition that make impulse noise through the result of these simulation.

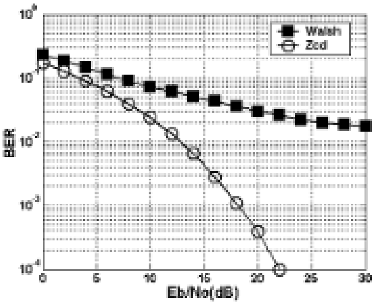


Fig. 5. BER performance in Gaussian noise, impulse noise (#10) and multi-path channel

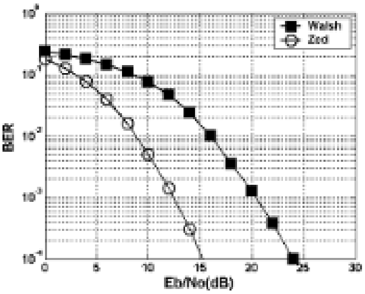


Fig. 6. BER performance in downlink environment (User: #3)

Property of diverse noises and channels at only one user were considered in prior simulation. But, in this simulation, we have applied Gaussian noise and MAI condition to testify superiority of ZCD-PLC system in multi-user system. We have made a simulation including downlink channel of three users in figure 6, and uplink channel of five users in figure 7. In figure 6, E_b/N_0 value of CDMA-PLC system based on ZCD code is roughly 12dB, and E_b/N_0 value of CDMA-PLC system based on Walsh code is about 20dB on the basis of value, 10^{-3}. Therefore, CDMA-PLC system based on ZCD code is more superior as 8dB as another one.

Meanwhile, when BER is 10^{-3} in figure 7, we can't get precise value because E_b/N_0 value of CDMA-PLC system based on Walsh code is divergent but E_b/N_0 value of CDMA-PLC system based on ZCD code is about 12dB. So to speak, we can realize that it is very similar to capability of downlink BER in figure 6.

Finally, we can verify that the factor of delay wave by multi-path and MAI is not affected by the factor of delay wave by multi-path and MAI when it is inside ZCD section.

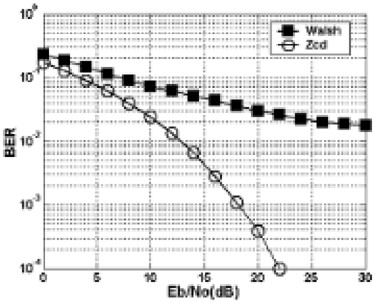


Fig. 7. BER performance in downlink environment (User: #5)

6 Conclusions

In this paper, we have proposed and proved the capability of new PLC system for low coherent home-network system based on binary ZCD-CDMA to overcome many problems generated by interference-wave from multi-path or performance-degradation by MAI from multi-user in CDMA-PLC system. I think that PLC system for low coherent home-network based on binary ZCD-CDMA system proposed in this paper will effectively solve the problem related to interference of PLC system for home-network. Moreover, at the point of time not to complete finally standard or basis of PLC for home-network, the outcome of this paper will expect to be used for references to standardize or innovative commercial use of future PLC technique.

Acknowledgement

This work was supported by KESRI (R-2003-0-516), which is funded by MOCIE (Ministry of Commerce, Industry and Energy).

References

1. Manfred Zimmermann and Klaus Dostert: Analysis and Modeling of Impulsive Noise in Broad-and Power line Communications. IEEE Trans. Electromagnetic Compatibility, Vol.44, NO. 1, pp. 249-258, February 2002
2. L. T. Tang, P. L. So, Member, IEEE, E. Gunawan, Y. L. Guan, S. Chen, and T. T. Lie: Characterization and Modeling of In-Building Power Lines for High-Speed Data Transmission. IEEE Trans. Power Delivery, VOL.18, No. 1, January 2003
3. Klaus M. Dostert: Power Lines as High Speed Data Transmission Channels - modeling the Physical Limits. IEEE Communication Magazine, pp. 585-589 April 1998
4. H.Philipps: Modelling of Powerline Communication Channels. Proc. 3rd Int'l. Symp. Power-Line Commun. and its Applications, Lancaster UK, 1999, pp.14-21
5. M.Zimmermann and K.Dostert: A Multipath Model for the Power Line Channel. IEEE Trans. Commun., vol. 50, no.4, Apr. 2002, pp. 553-59
6. P. Fan, M. Darnell: Sequence Design for Communications Applications. Research Studies Press, 1997
7. Cha,J.S., Kameda,S., Takahashi,K., Yokoyama,M., Suehiro,N., Masu,K. and Tsubouchi, K: Proposal and Implementation of Approximately Synchronized CDMA System Using Novel Biphase Sequences. Proc. ITC-CSCC 99, Vol. 1, pp.56-59, Sado Island, Japan, July13-15, 1999
8. Cha, J.S., Kameda, S., Yokoyama, M., Nakase, H., Masu, K., and Tsubouchi, K.: New Binary Sequences with Zero-correlation Duration for Approximately Synchronized CDMA. Electron. Lett., 2000, Vol. 36, no.11, pp.99,1993
9. Cha, J.S. and Tsubouchi, K: Novel Binary ZCD Sequences for Approximately Synchronized CDMA. Proc. IEEE 3G Wireless01, Sanfransisco, USA, Vol. 1, pp.810-813, May 29, 2001
10. Cha, J.S: Class of Ternary Spreading Sequences with Zero Correlation Duration. IEE Electronics Letters, Vol. 36, no.11, pp. 991-993, 2001.5.10
11. Sergio Verdu: Multi-user Detection. Cambridge university press, 1998

Securing Biometric Templates for Reliable Identity Authentication

Muhammad Khurram Khan and Jiashu Zhang

Research Group for Biometrics (RGB), Sichuan Province Key
Lab of Signal & Information Processing,
School of Computer & Communication Engineering,
Southwest Jiaotong University,
Chengdu, Sichuan, P.R China
Khurram.khan@scientist.com, itp@home.swjtu.edu.cn

Abstract. The large-scale implementation and deployment of biometric systems demand the concentration on the security holes, by which a reliable system can loose its integrity and acceptance. Like the passwords or PIN codes, biometric systems also suffer from inherent security threats and it is important to pay attention on the security issues before deploying a biometric system. To solve these problems, this paper proposes a novel chaotic encryption method to protect and secure biometric templates. To enhance the security of the templates, this research uses two chaotic maps for the encryption/decryption process. One chaotic map generates a pseudorandom sequence, which is used as private key. While on the other hand, another chaotic map encrypts the biometric data. Experimental results show that the proposed method is secure, fast, and easy to implement for achieving the security of biometric templates.

1 Introduction

A biometric is measurement of the biological characteristics of a human, which is used for the identification or verification purpose. There are many forms of biometric data for which capture and verification is possible via some device. Fingerprints, iris, voice recognition, retinal, face or hand scanning are feasible with the current technology.

Biometric is of interest in any area where it is important to verify and authenticate the true identity of an individual. Biometric technologies are gaining more attention because of secure authentication methods for user access, e-commerce, and access control; and are becoming the foundation of an extensive array of highly secure identification and personal verification solutions. Biometric has shown itself as an emerging technology, so it can be integrated with a lot of technologies to implement high security [1], [2], [3].

The deployment of a biometric system gives advantages over the traditional personal identification techniques such as PINs, smart cards, and passwords; however, the problem of ensuring the security and integrity of biometric data is critical [4]. Every body leave fingerprints when he touches something and a person's iris image can be captured from a secret camera. So a biometric-based verification system works

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 723 – 732, 2005.

properly only if the verifier system can guarantee that the biometric data came from the legitimate person at the time of enrollment [4].

Indeed, biometric is a cutting edge identification technology and provides uniqueness of its patterns, but it does not provide secrecy of its data. This technology only works well when the communication path via reader to verifier is trusted and reliable. If there is a security hole between reader and verifier, then biometric data could have risk of being hacked, modified, or reused. In order to promote the wide spread utilization of biometric techniques, an increased security of biometric data is necessary.

1.1 Related Work and Motivation

Unfortunately, the security issues of the biometrics data are usually ignored. There are only few papers published in the current era, which adhere to the security pitfalls of the biometric systems.

In the initial stages of the work, Davida et al. [5] published a study on the feasibility of protecting the privacy of a user's biometric data on an insecure storage device. They suggest that providing additional privacy for the biometric data may provide stronger user acceptance. In their second work [6], they utilized the error correction codes and explain their role in the cryptographically secure biometric authentication scheme.

Soutar et al [7], [8] also proposed some encryption scheme of the biometric data. They proposed an optical correlation-based fingerprint system. Their system binds a cryptographic key with the user's fingerprint images at the time of biometric enrollment. The cryptographic key is then retrieved only upon a successful verification.

During recent years, the use of digital watermarking for the biometric data is also explored. The authors of [9], [10], [11], [12], [13] employed the watermarking schemes for the security of the templates. The watermarking of templates could be used to protect the system from the replay attack, as identified by Ratha et al. [14]. Uludag et al. [15] presented a comprehensive analysis on the biometric cryptosystems and compared the different methods published in the current literature. This is a nice study on the security issues of the biometric systems.

Recently, Andy [16] presented a scheme that appears to show vulnerabilities in biometric encryption systems. Andy proposed an approach to attack biometric encryption algorithms in order to extract the secret code with less than brute force effort. He used hill-climbing attack to calculate an estimate of the enrolled image, which is then used to decrypt the code.

Reconstruction of original image is possible even if only part of the template is available. Hill [17] claimed that if only minutiae templates of a fingerprint are available, it is still possible to successfully regenerate artificial fingers of a person.

The problems of biometric template security raise concerns with the wide spread proliferation of biometric systems both in commercial and in government applications. So in this paper, we analyze inherent security pitfalls in a biometric system and propose a novel chaotic encryption method to secure and protect templates for the reliable identity authentication.

1.2 Outline

This paper is divided into five sections. Section 1 delineates the introduction and overview of biometrics with the related work, which has been done for the security of biometric systems. Section 2 gives the detailed descriptions of security holes, which are the vulnerable points to be attacked in a biometric system. Proposed encryption and decryption scheme is reported in Section 3. Section 4 explains the experimental and simulation results. At the end, Section 5 concludes the findings of the paper.

2 Security Holes in Biometric System

According to Ratha et al. [14], there are eight vulnerable points in a generic biometric system. These eight points are very important to be concerned when deploying a reliable and secure biometric system; else the integrity of the system could be at high risk. These attacking points are shown in Fig. 1.

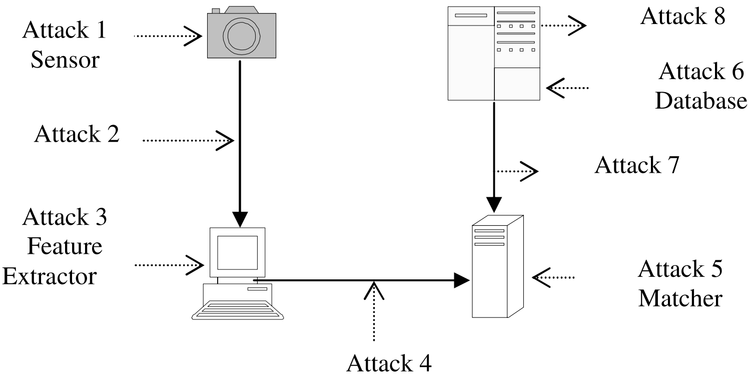


Fig. 1. Security holes in a generic biometric system (Derived from [14])

The first attack on the system is by presenting fake biometric trait at the sensor. Matsumoto et al. [18] spoof the 11 different fingerprint recognition systems with their fake created gummy fingers. They took gelatine to create some gummy fingers. These gummy fingers were easily enrolled by the live systems, which are commercially used for identity verification. The acceptance of the fake gummy fingers was approx. 67% in commercially used fingerprint verification systems.

For the remedy, Derakhshani et al. [19] proposed two methods for liveness detection of fingers at the sensor. They used sweat pores of the finger to detect a live finger. A neural network based classification technique is also employed to distinguish the liveness of fingers. Li et al. [20] also proposed a liveness detection scheme for the facial recognition. They work on the measurement of 3D depth information, and their method is based on the analysis of Fourier spectra of a single face image or face image sequences.

The second attack on the biometric system is called replay attack. In which, a hacker uses previously captured biometric data and by passes the sensor. Ratha et al. [14] proposed data-hiding techniques to secretly embed a telltale mark directly in the compressed fingerprint image.

The third vulnerable point in biometric system is to attack on the feature extractor. In this attack, a hacker can generate feature set by its own algorithm. The forth fragile point on a biometric system is the communication path from feature extractor to the matcher. In this attack, a hacker can change the original feature vector from his own generated set. This method is most vulnerable, if the feature extractor and matcher data flows over the network, so an intruder can easily modify, hack, or change it. The system proposed in [7] could be an ideal candidate to eliminate this attack from a biometric system.

In the fifth attack on the matcher unit, an intruder can change or override the matching results and can generate pre-selected matching scores. The sixth attack on the biometric database is also very critical, in which an intruder can steal or modify the biometric templates from the repository. This unit is one of the most intricate parts of the biometric system due its responsibility for safeguarding the permanent repository of all information collected from the system users [21]. By this attack, an authorized user could be rejected by the system or an imposter could get access of the resources. Smart card also store biometric templates on it and is also susceptible. Encryption or data hiding methods can be deployed before saving biometric templates into the database or smart cards [7].

In the seventh kind of attacks, the templates are at risk of hacking, modification, and reuse if they are sent to the matcher over network. Encryption techniques can be utilized for the communication between biometric repository and the matcher unit. In the eighth attack on the biometric system, the final decisions could be changed and the whole system could be failed. Even, when the whole biometric system works properly, but if the final decision is overridden then the whole system would become un-useful. Encryption of the final decisions could safe the system from this kind of attack.

3 Proposed Method

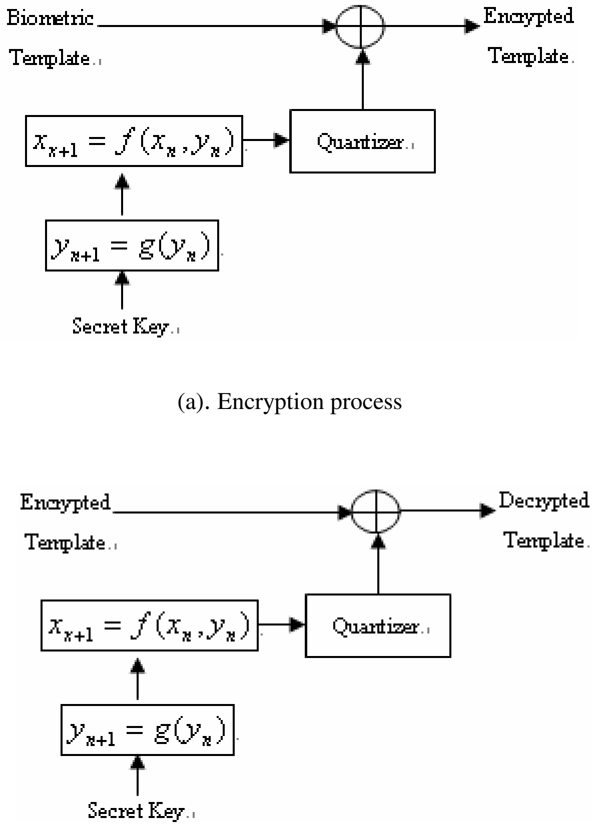
In this paper, we propose a scheme which can be used to secure biometric templates for the reliable identity authentication. For the experimental purpose, we use human iris as a reference biometric. But this system is generalized and can be extended to other biometrics e.g. fingerprint or face. To implement the system, we extract the features by the process and method described by Daugman [22], the inventor of the iris recognition system.

For the presented template encryption method, we propose to use two chaotic maps. The most attractive features of chaos in encryption are its extreme sensitivity to initial conditions and the outspreading of orbits over the entire space [23]. These special characteristics make a chaotic map excellent candidate for encryption, based on the classic Shannon's requirement of confusion and diffusion [23]. In recent years, chaotic maps have been widely used for the security of the signals [9], [24]. Because of these attractive properties, a chaotic sequence generated by two chaotic maps is applied on the iris templates for encryption, which make them more secure.

The proposed approach provides the features such as: (1) it is resistant to the finite word length effect of the chaotic sequence [24], (2) long-term unpredictable, (3) the robust against attacks, and (4) very sensitive to the initial condition, which is used as a secret key. In the following subsections, we would give the details of the proposed method.

(a) Encryption Scheme

The proposed method for the encryption/decryption of biometric templates is depicted in Fig.2. Fig. 2(a) delineates the template encryption model, and Fig. 2(b) shows the template decryption method.



(a). Encryption process

(b). Decryption process

Fig. 2. Encryption and Decryption Model (a) Encryption process (b) Decryption process

In this method, we use two chaotic maps for the encryption process. One map, which is logistic chaotic map, is used to generate 1-D sequence of real numbers that is used as a key. Equation 1 shows the logistic chaotic map.

$$y_{n+1} = g(y_n) = \mu y_n(1 - y_n) \tag{1}$$

where n=1,2,3,.... is the map iteration index and μ is the system parameter. For $3.57 < \mu \leq 4.0$, the sequence is non-periodic, non-convergent, and very sensitive to the initial value, which can be used as a secret key.

Another chaotic map is Henon map [24], which is used to encrypt the templates as shown in equation 2.

$$x_{n+1} = [1 + b(x_{n-1} - c) + 379 y_n^2](\mathrm{mod}1) \tag{2}$$

where *b=0.3* and $1.07 \leq c \leq 1.09$. Modulo operation is performed to restrict the chaotic sequence within limits and it also prevents the chaotic sequence from divergence.

Because the stream generated by Henon map is a sequence of real numbers, the output of the Henon map in Eq. (2) should be quantized into binary stream to perform XOR operation with the biometric template.

The normalized sequence created by Henon map is $S_n \in \{0,1\}$ and the extracted biometric template is $T_n \in \{0,1\}$. The XOR operation is performed between the binary sequence and biometric data, and the encrypted template is $E_n = S_n \oplus T_n$.

Now the biometric template is encrypted and can be used to save into the database or smart card or can be transmitted over a communication channel.

(b) Decryption and Matching Scheme

The decryption model of biometric templates is depicted in fig 2(b). The decryption model is depended on the right key to decrypt the data. The decryption process is also an XOR operation between the encrypted template i.e. E_n and the sequence generated by the chaotic decryption system i.e. S_n

The XOR operation is performed to get the original biometric template for the further matching or verification process, so the decrypted template is $D_n = S_n \oplus E_n$

If the decrypted template is used for matching in the database, then we employ the equation 3 which matches biometric templates for the identification of a person.

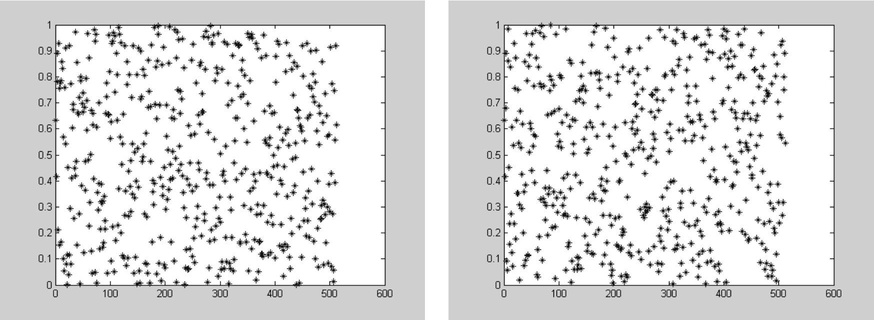
$$M_{Match} = \frac{N_z(xor(T_n, T_n^{'}))}{N} \tag{3}$$

where, N_Z is total number of zeros by an Exclusive-OR (XOR) operation between an original template T_n stored in the database, while $T_n^{'}$ is the decrypted template. N is the size of the template. Here, we may define a matching threshold according to the criticality and usage of the system.

4 Experimental Results

In order to evaluate the performance of the proposed method, portions of the research in this paper use the CASIA iris image database collected by institute of automation, Chinese academy of sciences. CASIA iris image database (ver 1.0) contains 756 iris images from 108 eyes [25]. The size of the extracted feature set from the images is 512 bytes, which is normally the standard size of the commercially used iris based verification system.

The characteristics and performance of the proposed system are demonstrated by the Fig.3, which shows the completely different attractors of the chaotic map at two different values or secret keys.



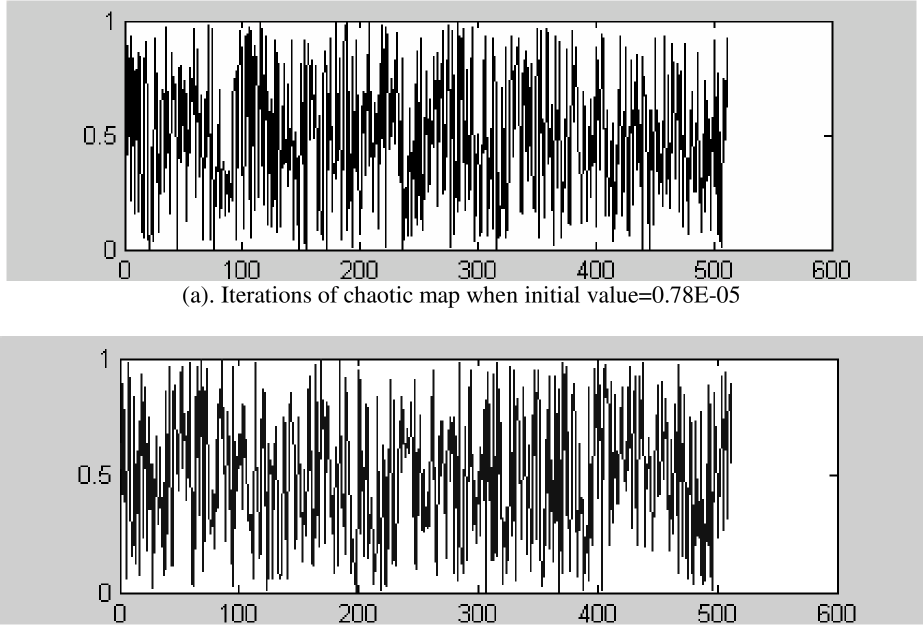
(a). Attractor: when initial value=0.78 (b). Attractor: when initial value=0.78E-10

Fig. 3. Diffusion of chaotic maps

In fig. 3(a), as an example, the initial value (secret key) used is 0.78E-05 and is shown by randomly diffused chaotic attractor. While on the other hand, when the value of secret key is slightly changed i.e. 0.78E-10, then the behavior of the chaotic system is abruptly changed. Fig. 4 also shows the two iteration sets of the chaotic map at the two slightly changed values.

It is also pertinent to say that if the value of the secret key is changed by even a tiny amount, say as tiny as the inverse of Avogadro's number (a small number with an order of 1E-24), checking the attractor at a later time will yield numbers totally different. This is because small differences will propagate themselves recursively until numbers are entirely dissimilar to the original system with the original secret key.

The presented chaotic encryption method also exhibits better security and protection against those, which use only one chaotic map. The important thing in the system to be protected is the secret key, which can be kept at the safe place because if the key is compromised, then the integrity of the whole system is at high risk and an intruder can masquerade a legitimate user.



(a). Iterations of chaotic map when initial value=0.78E-05

(b). Iterations of chaotic map when initial value=0.78E-10

Fig. 4. Two sets of iterations at slightly different initial values

The proposed system is an open ended system, which can be used to encrypt any biometric data e.g. face, fingerprint or hand geometry. This system can also be used for high dimensional biometric templates e.g. multimodal biometrics, whose template size is bigger than the examples used in this paper.

The presented method is very useful to protect and safe the biometric system from the attacks 4, 6, and 7 identified in fig.1. To cure from the attack 4 as shown in fig.1, the feature set can be encrypted by our scheme before sending to the matcher unit. To protect from the attack 6, the biometric template database can also be encrypted by our method. Presented method can also be implemented on smart cards, which contain biometric templates. Furthermore, our procedure can be used to encrypt templates before sending to the matcher unit from the biometric repository unit, as shown in fig.1; which comes under the attack 7 on the biometric system.

5 Conclusion

In this work, we have proposed an innovative encryption scheme to secure the biometric templates for the reliable identity authentication. Due to the excellent randomness of the chaotic systems, we encrypted biometric templates by two chaotic maps. One chaotic map is used as a private key and another for the encryption of the templates. Usage of two chaotic maps makes our system more secured and protected from the attacks and cryptanalysis. Furthermore, our scheme can be applied to protect from different kinds of attacks, which are found in a generic biometric system.

Moreover, we performed a series of experiments to evaluate the performance of the proposed algorithm. The proposed algorithm can be applied for any kind biometric templates for the security e.g. fingerprint, face or palm print etc. So our system is an open ended system for securing the biometric templates.

Acknowledgements

This project is supported by 'Sichuan Youth Funds' under grant number: 03ZQ026-033. Authors are full of gratitude to the Professor Stan Z. Li for inviting their paper at the special session. Authors also pay their thanks to Dr. Bin Zhu for his correspondence and help.

References

1. Muhammad Khurram Khan, Zhang Jiashu: Optimizing WLAN Security by Biometrics. in "Handbook of Wireless Local Area Networks: Applications, Technology, Security, and Standards" [in press], CRC Press Boca Raton, Florida, USA, 2005
2. J. Daugman: High Confidence Visual Recognition of Persons by a Test of Statistical Independence. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, pp.1148-1161, Nov. 1993
3. S. B Pan et al: An Ultra-Low Memory Fingerprint Matching Algorithm and Its Implementation on a 32-bit Smart Card. IEEE Transactions on Consumer Electronics, vol. 49, no. 2, pp. 453-459, May 2003
4. B. Schneier: The uses and abuses of biometrics. Comm. ACM, vol. 42, no. 8, pp. 136, Aug. 1999
5. G.I. Davida, Y.Frankel, and B. J. Matt: On Enabling Secure Applications through Online Biometric Identification. In IEEE Symposium on Security and Privacy, pp. 148-157, 1998
6. G. I. Davida, Y. Frankel, B. J. Matt, and R. Peralta: On the Relation of Error Correction and Cryptography to an Offline Biometric Based Identification Scheme. In Proc. Workshop Coding and Cryptography, pp. 129–138, 1999
7. C. Soutar, D. Roberge, S. A. Stojanov, R. Gilroy, and B. V. K. Vijaya Kumar: Biometric Encryption Using Image Processing. In Proc. SPIE, Optical Security and Counterfeit Deterrence Techniques II, vol. 3314, 1998, pp. 178–188
8. C. Soutar, D. Roberge, S. A. Stojanov, R. Gilroy, and B. V. K. Vijaya Kumar, "Biometric encryption—enrollment and verification procedures," in Proc. SPIE, Optical Pattern Recognition IX, vol. 3386, 1998, pp. 24–35

9. Muhammad Khurram Khan, Zhang Jiashu, Tian Lei: Protecting Biometric Data for Personal Identification. Sinobiometrics'04, Lecture Notes in Computer Science, Springer-Verlag Germany, pp. 629-638, vol. 3383, Dec. 2004
10. Anil K. Jain, Umut Uludag: Hiding Biometric Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp. 1494-1498, November 2003
11. B. Gunsel, U. Uludag, and A.M. Tekalp: Robust Watermarking of Fingerprint Images. Pattern Recognition, Elsevier Science Ltd., vol. 35, no. 12, pp. 2739-2747, Dec. 2002
12. Sonia Jain: Digital Watermarking Techniques: A Case Study in Fingerprints and Faces. Proc. Indian Conference on Computer Vision, Graphics, and Image Processing, pp.139-144, Dec.2000
13. S. Pankanti and M.M. Yeung: Verification Watermarks on Fingerprint Recognition and Retrieval. Proc. SPIE, vol. 3657, pp. 66-78, 1999
14. N. Ratha, J. Connell, and R. Bolle: Enhancing Security and Privacy in Biometrics-based Authentication Systems. IBM System Journal, vol. 40, no. 3, pp. 614–634, 2001
15. Uludag, U., Pankanti, S., Prabhakar, S., Jain, A.K.: Biometric Cryptosystems: Issues and Challenges. Proceedings of IEEE, vol. 92, pp. 948–960, 2004
16. Andy Adler: Vulnerabilities in Biometric Encryption Systems. Accepted at Audio visual Based Biometric Person Authentication (AVBPA), July 2005, USA
17. Hill, C.J.: Risk of Masquerade Arising from the Storage of Biometrics. Bachelor of Science thesis, Dept. of CS, Australian National University (2002)
18. T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino: Impact of Artificial Gummy Fingers on Fingerprint Systems. Proc. of SPIE, Optical Security and Counterfeit Deterrence Techniques IV, vol. 4677, pp. 275-289, 2002
19. R. Derakhshani, S.A.C. Schuckers, L.A. Hornak, and L.O. Gorman: Determination of Vitality from a Non-invasive Biomedical Measurement for Use in Fingerprint Scanners. Pattern Recognition, vol. 36, pp. 383-396, 2003
20. Jiangwei Li, Yunhong Wang, Tieniu Tan, A.K.Jain: Live Face Detection Based on the Analysis of Fourier Spectra. Proc. SPIE on defense and security symposium, vol. 5404, April 2004
21. A.C. Leniski, R.C. Skinner, S.F. McGann, S. J. Elliott: Securing the Biometric Model. IEEE ICCST, Taiwan, October 2003
22. J. Daugman: High Confidence Visual Recognition of Persons by a Test of Statistical Independence. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, pp.1148-1161, Nov. 1993
23. Zhao Dawei, Chen Guanrong, Liu Wenbo: A Chaos-based Robust Wavelet-domain Watermarking Algorithm. Chaos, Solitons and Fractals, Elsevier Science Ltd., vol. 22, pp. 47-54, 2004
24. Jiashu Zhang, Lei Tian and Heng-Ming Tai: A New Watermarking Method Based on Chaotic Maps. Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on, Volume 2, pp. 939 – 942, June 2004
25. CASIA Iris Database, [online] http://www.sinobiometrics.com, March 2005

Intelligent Tracking Persons Through Non-overlapping Cameras

Kyoung-Mi Lee

Department of Computer Science,
College of Information Engineering,
Duksung Women's University,
Seoul 132-714, Korea
kmlee@duksung.ac.kr
http://www.duksung.ac.kr/~kmlee

Abstract. An intelligent surveillance system can judge and handle a situation automatically within a wide monitoring area and unattended environment that has no certain human supervisor. In this paper, we propose a way to track persons through non-overlapping cameras that are connected over a network with a server. To track persons with a camera and send the tracking data to other cameras, the proposed system uses a human model that comprises a head, a torso, and legs. Also, with a trajectory model, the proposed system can predict the probability which an exited person from one camera is incoming to other cameras. The system is updated online during the lifetime of the system. These enable the proposed to keep tracking the recognized person in a wide area, to provide a guide for monitoring multiple cameras, and to adapt changes with time.

1 Introduction

Tracking people based on images using a video camera plays an important role in surveillance systems. The preceding studies on vision-based human tracking and monitoring are limited in that they use a single camera or, even in the case of multiple cameras, they apply three-dimensional human modeling to mainly recognize human body posture. Due to the limited view of each camera, however, the tracking on an individual camera is possible but the image tracked by a camera cannot be transferred to another camera for continued tracking. Since existing monitoring systems require that a human supervisor be present so as to validate with naked eyes the images entering a camera, we need to develop an intelligent surveillance system that allows several cameras to monitor different areas while at the same time predicting a trajectory of a previously tracked person and automating the tracking of the person spotted on camera across networked cameras. Javed *et. al* established correspondence across multiple cameras that determine the FOV line of each camera as viewed in other camera with real-time [2]. Porikli and Divakaran proposed the object-wise semantics from non-overlapping cameras and solved an inter-camera color calibration problem by using color histogram to determine inter-camera radiometric mismatch and correlation matrix [3].

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 733 – 741, 2005.

In this paper, we propose a wide-area surveillance system that tracks persons with non-overlapping cameras and with data transfer between cameras via the network. Section 3 explains the system configuration and the data transfer. Section 4 describes the process of tracking persons with non-overlapping cameras. The results of the experiments and conclusions are explained in Section 5.

2 Related Research

In the last few decades, researchers have used video cameras and proposed different environment and methods to acquire and track a moving object or a person both indoors and outdoors. These methods can be largely classified according to four different criteria: "place (indoors, outdoors)," "range (broad areas, narrow areas)," "number of cameras (a single camera, multiple cameras)," and "color data (in full color, in gray scale)" of the images acquired.

Human tracking methods can be classified into indoor environment and outdoor environment depending on the place of tracking, and broad areas and narrow areas depending on the range of tracking. In case of indoor environment, where mostly airports, the inside of subway stations, and the interior of buildings (the inside of libraries, office, and hallway) have been studied, the biometric data of the tracking target (contours, points, and color data) are very important because the tracking is made in a relatively narrow place. On the other hand, the outdoor environment, where roads, parking lots, the exterior of a building (the front of a house, a library, and a school), golf courses, and playgrounds are studied in most cases, has a broader range of tracking than indoor environment does. Therefore, in tracking persons in a wide area, you may need to consider additional data, like the direction of movement, speed, and position of the tracking target, on top of the biometric data required for narrow-range tracking. Furthermore, outdoor environment is more susceptible to surrounding noise and lighting than indoor environment, so that it becomes very difficult to make a robust tracking of an object or a person in an outdoor environment despite the higher interference from lighting and noise. This paper targets both indoors and outdoors as its experiment conditions, while restricting the range within relatively narrow areas.

In addition, the tracking methods can be grouped into a single camera and multiple cameras depending on the number of cameras. Tracking with a single camera offers only a limited scope of vision, which makes it difficult to track a person's movement or behavioral patterns over a long period of time. That is why more than two cameras are used to expand the limited scope of a single camera and track a person. Multiple-camera tracking is further divided into overlap or non-overlap arrangement, depending on whether or not the cameras share a part of the tracking range. In an overlapping arrangement, more than two cameras share a certain range of view, whereas, in a non-overlapping arrangement, the cameras are placed over a broad area with no overlap with each other. In both cases, however, the cameras need to communicate information on the tracking target with each other. That is why they are all connected to a server that enables the sharing of information among cameras. In this paper, multiple non-overlapping cameras fixed in each position are used to acquire a broad field of vision as in Figure 1. Because the cameras need to exchange information with one another when you use multiple cameras, they have to be able to share and transfer information by connecting to a server.

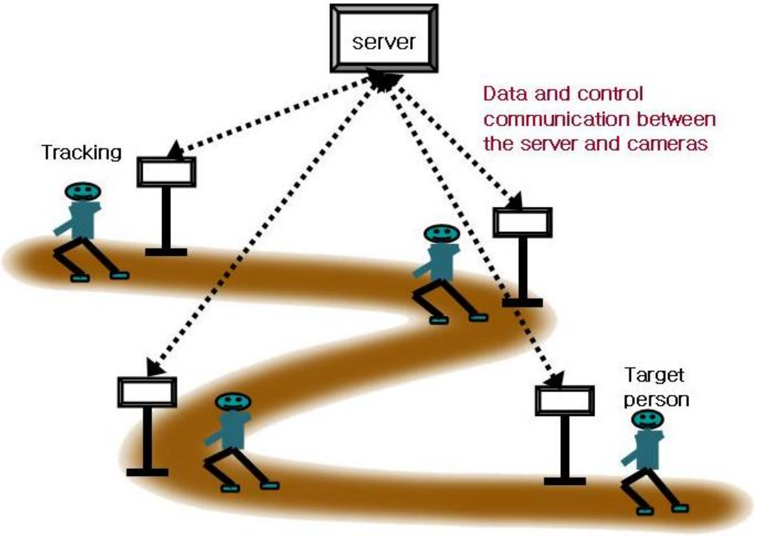


Fig. 1. The proposed networked camera system is networked with non-overlapping cameras for wide area surveillance. The system can track a target person through networked cameras by transferring tracking data between a server and cameras.

3 Non-overlapping Camera System

The purpose of the proposed human tracking system is to use networked cameras with non-overlapping view that facilitate the acquisition of broad view, while letting these cameras transfer the data on a tracking target. The target person is transferred as a form of a human model which contains data on the corresponding including the person's position, color, direction of movement, a central point, and boundaries [5]. The human model information is registered in a personlist that stores information on persons recognized by the server, which can be used in the data communication between the cameras and the server. The server sends out the feature data regarding the person who needs to be tracked to the cameras over the network. With the data from the server, a camera checks if a person who enters its field of vision is the one to be tracked and, if so, initiates the tracking.

Fig. 2 shows how multiple cameras transfer human-model data to one another. When a person enters the view of a camera that is connected to the server, the camera checks if the person is registered in the personlist. If she is the one on the list, the camera keeps on tracking the person. If not, it determines that he or she is a new person entering the network of cameras and register a new human model on the persons list. This way each camera can check if a person entering its field of vision, or video frame, is one registered on the list or if she is a new person, or if the one who has been here and away is coming back. With these continued checks and tracking, the cameras that are distributed over a wide area can be used to make an end-to-end tracking of a person.

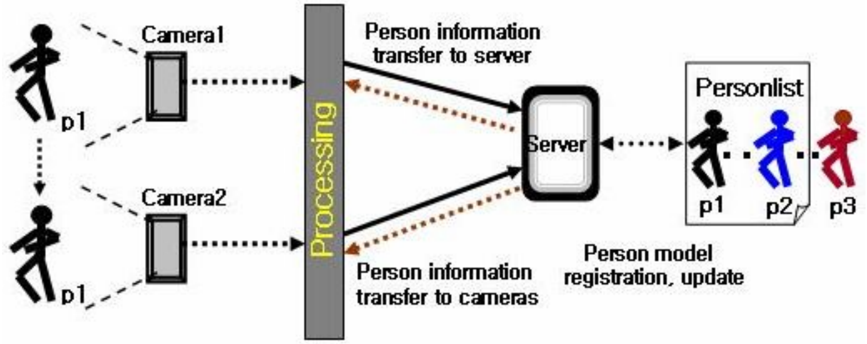

Camera1
p1
Camera2
p1
Processing
Person information transfer to server
Server
Personlist
p1
p2
p3
Person information transfer to cameras
Person model registration, update

Fig. 2. The proposed surveillance system can transfer information among non-overlapping cameras connected over a server

4 Human Tracking Through Non-overlapping Cameras

In this section, we introduce an approach that tracks persons through non-overlapping cameras. During tracking, the proposed approach compensates illumination noises, separates persons, models the separated persons, tracks the modeled persons, and predicts trajectories between cameras (Fig. 3). Also, the proposed system is updated by incoming frames to adapt changes with time.

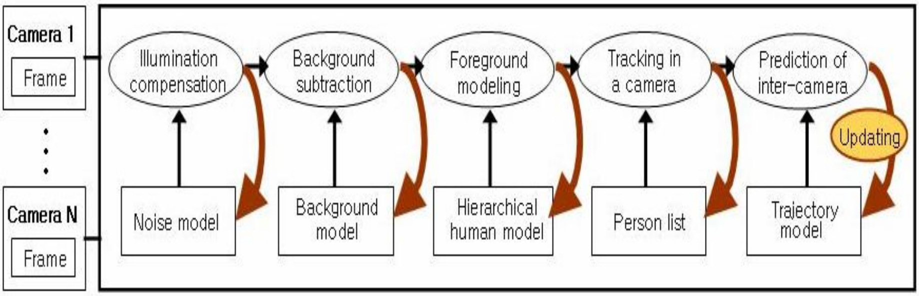


Fig. 3. Flowchart of the proposed tracking system: Ellipses are processes, rectangles data, and curved arrows updating process steps with online learning

4.1 Adaptive Background Subtraction

Video image frames taken by a camera have variation in illumination conditions caused by lighting, time of day, and so on. Since noises by such conditions make tracking difficult, the noise should be removed from the image frame. To separate noises from images, an intrinsic image can be used to get a noise image by subtracting from the image frame. While adding the noise image to the image frame corrects background effectively, it is not sufficient to correct non-background objects. In this paper, we update a noise image frame-by-frame to estimate a time-varying intrinsic image [5]. We first initialize a noise image by subtracting the first image frame from the intrinsic image and then update the noise image frame-by-frame. If a pixel is similar with a noise pixel, the pixel is updated.

To detect moving persons after illumination correction, the proposed system conducts background subtraction. In this paper, we build the adaptive background model, using the mean and standard deviation of the background [5]. Whenever a new frame arrives, a change in pixel intensity is computed using Mahalanobis distance to classify background or foreground (moving persons). The evaluated distance is compared to a difference threshold previously observed from the sequence of images. If a pixel is classified to background, the adaptive background model is updated with the pixel.

4.2 Model-Based Person Tracking

After background subtraction, tracking persons are initialized by a hierarchical person model (Fig. 4). To group segmented foreground pixels into a blob and locate the blob on a body part, we use a connected-component algorithm and then merge small blobs into neighboring blobs that share similar colors to overcome over-segmentation generated by initial grouping. Each blob contains information such as an area, a central position, color, position, a bounding box, and a boundary to form a human body. As a person can be defined as a subset of blobs, which correspond to human body parts, blobs in a frame should be assigned to corresponding individuals to facilitate multiple individual tracking. Each person is divided into three human body parts (head, torso, legs) and separated blobs are allocated to three human body parts using geometrical information and color information according to relative position in the model.

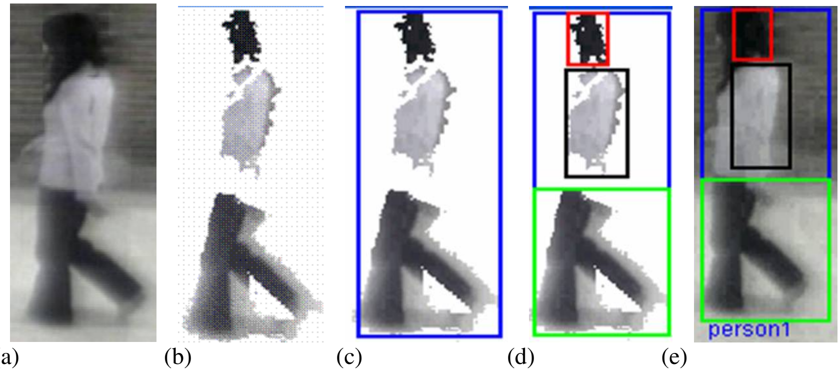


Fig. 4. A hierarchic human model (a) original image (b) background subtraction image (c) person area (d) 3 body parts(head, torso, legs) (e) result

Tracking people poses several difficulties, since the human body is a non-rigid form. After forming blobs, a blob-based person tracking maps blobs from the previous frame to the current frame, by computing the distance between blobs in consecutive frames. However, such a blob-based approach for tracking multiple persons may cause problems due to the different number of blobs in each frame: blobs can be split, merged, or even disappear or be newly created. To overcome this situation, many-to-many blob mapping can be applied [1]. In this paper, the proposed human tracking system treats blobs to a human model's components instead of tracking blobs individually, and could prevent that affect in tracking by error that can happen in image processing. Even though a blob can be missed by an illumination change, person-based tracking can retain individual identity using other existing blobs.

We assume that persons have already been tracked up to frame *t*-1 and new blobs are formed in frame *t*. Multi-persons are then tracked as follows:

Case 1: If a new blob is included in a person model at *t*-1, the corresponding blob in the model at *t*-1 is tracked to the new blob.

Case 2: If a blob in a person model at *t*-1 is separated into several blobs in frame *t*, the blob in the model at t-1 is tracked to one blob in frame *t* and other blobs at time *t* are appended into the model at *t*-1.

Case 3: If several blobs in a person model at *t*-1 are merged into the new blob, one blob in the model at *t*-1 is tracked to the new blob and other blobs are removed from the model at *t*-1.

Case 4: If a new blob is included in a person model at *t*-1 but the corresponding blob does not exist, the new blob is added to the model at *t*-1.

Case 5: If a new blob is not included in a person model at *t*-1, the blob is considered as a newly appearing blob and thus a new person is added to the person list.

where including a region into a person with a bounding box means the region overlaps over 90% to the person. In addition to simplify the handling of lists of persons and blobs, the proposed approach can keep observe existing persons exiting, new persons entering, and previously monitored persons re-entering the scene. One advantage of the proposed approach is to relieve the burden of correctly blobbing. Even though a blob can be missed by an illumination change, model-based tracking can retain individual identity using other existing blobs. After forming persons, the number of blobs in a person is flexibly changeable. The proposed approach can handle over-blobbing (Case 2) and under-blobbing (Case 3) problems.

4.3 Prediction of Inter-camera Trajectories

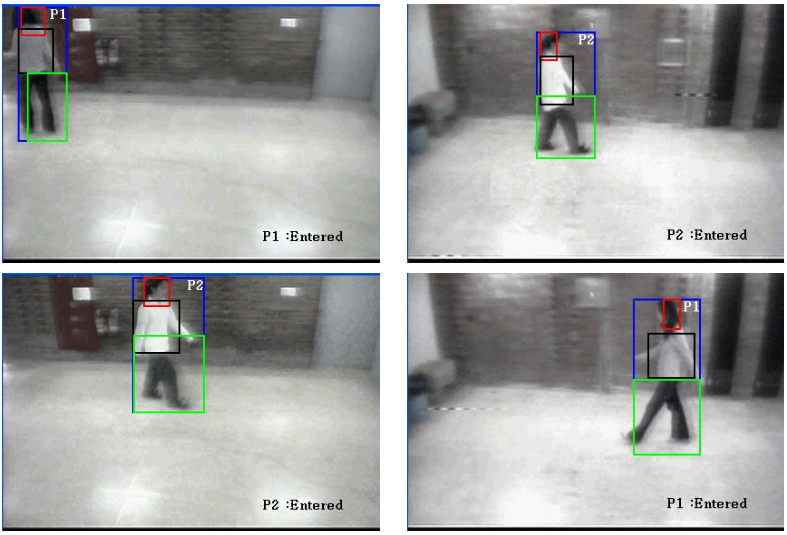
To keep tracking a target person as it moves in a wide area, the proposed system should correspond the track through a region between non-overlapping cameras. The topology of the surveillance area identifies cameras that are spatially adjacent and hence may potentially contain an inter-connecting pathway, directly connecting trajectories from the two cameras. Since the proposed system allows one camera to transfer information on the person to the other cameras, knowledge of the trajectory helps the system predict to anticipate where the target person exited from the camera will reappear.

In order to represent the spatial and temporal information, we can augment the trajectory model to create a tempo-topographical model of the network. By simply calculating statistical information of a dataset of observed trajectories, the proposed system provides a probability and a velocity which the target will reenter to each camera. This trajectory model is automatically learnt from the dataset, using an EAM(Expectation-Allocation-Maximization) algorithm with online update and incremental growth [4]. The EAM algorithm starts with no trajectory clusters in this paper. Whenever a new trajectory between two cameras is detected, EAM first finds a closest trajectory cluster. If the new trajectory is similar to one of existing clusters, the center and the covariance matrix of the similar cluster is updated. If the trajectory is not similar to any existing trajectory clusters and is not currently well represented by existing trajectory clusters, a new trajectory cluster is created with the trajectory and the number of clusters is increased. Thus, after training on the set of trajectories, each cluster represents a possible trajectory in the surveillance area.

5 Results and Conclusions

The proposed method is implemented with JAVA (JMF) on the Microsoft Windows 2000 XP platform. The experiment was carried out on a computer using the images (resolution: 420x316) acquired from two UNIMO CCN-541 security cameras.

The proposed human tracking method with non-overlapping cameras has been experimented in an indoor setting, like a long hallway, and an outdoor setting, like a long path. Fig. 5 and Fig. 6 show the results of the human tracking with the cameras 1 and 2 in indoor and outdoor environments, respectively. The persons initially tracked



Camera 1

Camera 2

Fig. 5. Result images in indoor environments: Upper images are taken at the 21th frame and bottom images at the 96th frame. Left images are achieved from Camera 1 and right images from Camera 2.

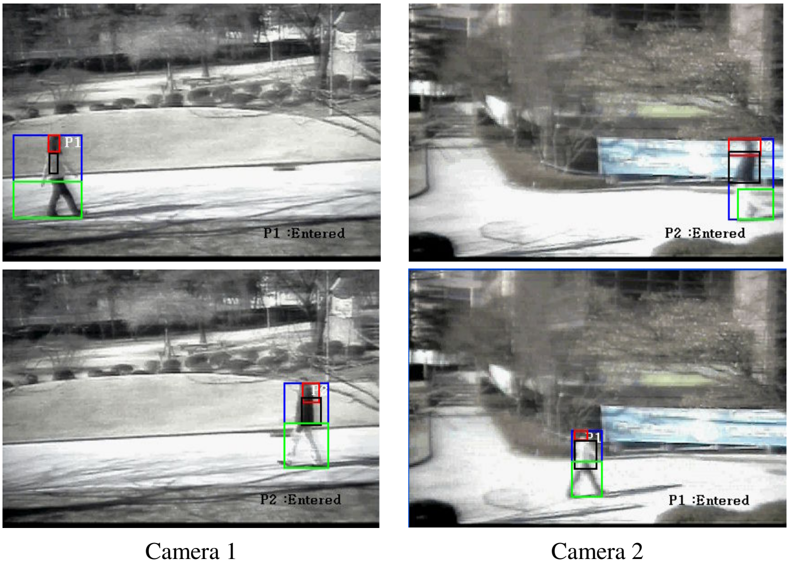


Fig. 6. Result images in outdoor environments: Upper images are taken at the 21th frame and bottom images at the 96th frame. Left images are achieved from Camera 1 and right images from Camera 2.

by camera 1 (person 1) and camera 2 (person 2) have moved over time to the field of vision of the other camera—camera 2 (person 1) and camera 1 (person 2). Each camera sends over to the server the modeling data it created for the features of the person it acquired. Upon receiving the data, the server delivers the feature data regarding the human model to all the cameras connected to the network, so that each camera keeps on tracking the recognized person whose movement enters its FOV.

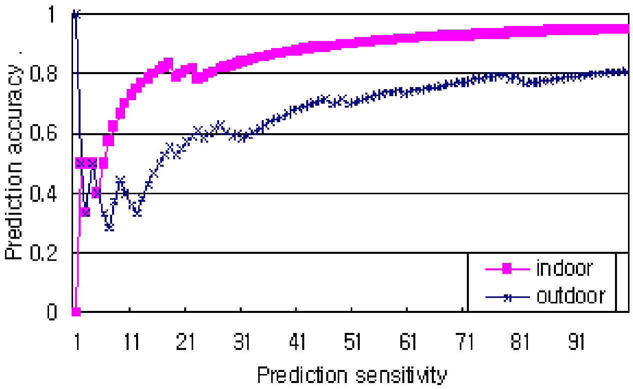


Fig. 7. Prediction accuracy of surveillance areas

To evaluate the tracking performance of the proposed algorithm, we used tracking accuracy which divides the number of correctly tracked persons by the number of tracked persons. The proposed tracking algorithm is achieved 91% of person 1 and 97% of person 2 in the indoor environments, and 82% of person 1 and 80% of person 2 in outdoor environments.

Prediction performance was tested on a dataset consisting of 12,743 trajectories, derived during an 8-hour period. We used two measures to evaluate the performance of the algorithm: sensitivity and accuracy. Sensitivity is the likelihood that a trajectory will be predicted. Accuracy is the likelihood that a trajectory match is actually associated with a trajectory passed by a person. Fig. 7 shows that as the system is trained on more and more trajectories, its ability to correctly predict the trajectory between non-overlapping cameras increases. After a few trajectories, the prediction rate stabilizes at 95% for the indoor environment and 81% for the outdoor environment.

The goal of this paper was to automatically track multi-person in an environment of no human supervisor for wide-area surveillance. By transferring a human model, the proposed system can communicates with cameras connected over a network. Also, to adapt changes with time, the system updates online a noise model for illumination compensation, a background model for background subtraction, person models for tracking, and a trajectory model for prediction of inter-camera trajectories. The future works of this research are to implement a virtual space that will provide an in-depth look into the trails of the tracked persons' movements and to develop an intelligent interface.

Acknowledgements

This work was supported by Korea Research Foundation Grant (KRF-2004-003-D00376).

References

1. S. Park and J. K. Aggarwa: Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing. In Proc. of International workshop on motion and video computing, pp.105-111, 2002
2. O. Javed, Z. Rasheed, O. Alatas and M. Shah: KNIGHT^{M}: A Real-time Surveillance System for Multiple Overlapping and Non-overlapping Cameras. In Proc. of ICME, 2003
3. F.M.Porikli and A. Divakaran: Multi-camera Calibration, Object Tracking and Query Generation. In Proc. of ICME, pp. 653-656, 2003
4. K.-M. Lee: Elliptical Clustering with Incremental Growth and Its Application to Skin Color Region Segmentation. Journal of KISS, 31(9):1161-1170, 2004
5. K. M. Lee and Y. M. Lee. Tracking Multi-person Robust to Illumination Changes and Occlusions. In Proc. of International of Conference on Artificial Reality and Telexistence, pp. 429-432, 2004

A New LUT Watermarking Scheme with Near Minimum Distortion Based on the Statistical Modeling in the Wavelet Domain

Kan Li and Xiao-Ping Zhang

Department of Electrical and Computer Engineering,
Ryerson University,
350 Victoria Street, Toronto, Ontario,
Canada, M5B 2K3
xzhang@ee.ryerson.ca

Abstract. This paper presents a new wavelet domain look-up table (LUT) watermarking algorithm that leads to the sub-optimal embedding of watermarks in the sense of minimizing distortion. The algorithm provides a joint distortion-robustness design of the LUT based watermark. There are two key features in the algorithm: (1) a near minimum-distortion LUT with the maximum run of 2 is designed based on the statistical properties of the wavelet coefficients; (2) an expectation-maximization (EM) algorithm based method is employed to model the statistical distribution of wavelet coefficients and select significant coefficients (coefficients with large magnitude) for watermark embedding. The experimental results show that images watermarked by the proposed algorithm have about 1.5-2.5dB peak-signal-to-noise-ratio (PSNR) gain over the conventional odd-even embedding method, while the system presents great robustness. In the case of 0.2bpp (1/40) JPEG2000 compression, the proposed scheme can ensure a reasonably low bit error rate (BER).

1 Introduction

Digital watermarking techniques have made considerable progress in recent years. Various watermarking schemes have been proposed to fit the design requirements of various applications. For example, watermarks for ownership protection need to survive common processing and intentional attacks and thus need to be robust. For tampering detection applications, the embedded secondary data are used to determine whether the host media is tampered or not, so only fragile (fragile to any changes) or semi-fragile (robust to common processing, but fragile to intentional attacks) watermarking is expected [1].

Based on embedding mechanisms, the existing watermarking schemes can be classified into two categories: host-independent watermarking and host-dependent watermarking. In host-independent watermarking, the embedded watermark is independent of the host data. In these watermarking systems, a weak signal, which is not related to the host data, is added to the cover media [2], [3], [4]. The additive spread spectrum algorithm [2] is a representative of this category. In host-dependent water-

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 742 – 750, 2005.

marking, the embedded watermark and the watermarking scheme exploit the information about the cover media. It is observed that these adaptive watermarking methods can achieve better performance [2]. Quantization based watermarking schemes [5] are in this category. Quantization based watermarking algorithm embeds a watermark, which is often a binary sequence, into a host image by quantizing the image pixels or transform coefficients. Look-up table (LUT) embedding is a type of quantization based watermarking. A LUT is a set of quantized values. Each quantized value in a LUT carries side information “1” or “0”. A general LUT watermark embedding algorithm maps original coefficients to the closest quantized value associated with the desired watermark information bit in a LUT [6]. The LUT is generated beforehand. The embedded watermark can then be extracted from the received image by looking up the LUT.

The well known odd-even embedding is the simplest case of LUT embedding. The table entries for embedding "1" and "0" are arranged in an interleaving order, which is also formulated [7] as a LUT with run of 1. Here run means the largest number of the consecutive 0’s or 1’s. Apparently, the larger the run is, the better the robustness is. It is also indicated in [7] that LUT embedding with larger run constraints introduces larger distortion with enhanced robustness and security. This conclusion is based on the assumption that the source follows a uniform distribution. However, if the original coefficients do not follow a uniform distribution, the probabilities that the coefficients fall into each quantization cell will not be exactly the same. It is then possible that LUT embedding with maximum run of 2 may achieve less distortion than LUT embedding with run of 1.

Many wavelet-based watermarking schemes have been proposed for applications such as copyright protection or image authentication [8], [9]. It is well known that wavelet coefficients do not follow the uniform distribution. Therefore, by carefully designing the LUT according to the statistical distribution of wavelet coefficients, smaller distortion may be achieved without loss of robustness. Previous research also indicates that, to embed a robust watermark into multimedia, only significant portions of the host image have to be modified [8], [9]. Appropriate significant coefficients selection method has to be used for watermark embedding.

In this paper, we present a new LUT based watermarking technique in the wavelet transform domain, which can achieve near minimum distortion under certain run (robustness) constraint. Knowledge about the host image and the watermark is exploited in our watermarking system. The new scheme includes two main features: (1) the LUT is designed to minimize the distortion according to the statistical distribution of wavelet coefficients; (2) Expectation-Maximization (EM) algorithm is used to model the wavelet coefficients. Only the significant coefficients are selected to embed the watermark.

2 New LUT Based Watermark Embedding

2.1 Significant Coefficient Probability Based on a Gaussian Mixture Model in the Wavelet Domain

In our scheme, only wavelet coefficients with large magnitude are selected to bear watermarks. In general, these coefficients do not change much after image processing and compression attack. A statistical method is presented to pick the embeddable coefficients based on a Gaussian mixture model in a wavelet subspace by an expectation-maximization (EM) algorithm [10].

In our scheme, only wavelet coefficients with large magnitude are selected to bear watermarks. In general, these coefficients do not change much after image processing and compression attack. A statistical method is presented to pick the embeddable coefficients based on a Gaussian mixture model in a wavelet subspace by an expectation-maximization (EM) algorithm [10].

The wavelet coefficients have a peaky, heavy-tailed marginal distribution [11]. Only a few significant coefficients take large values at the positions where edges occur, while most others take small values. This statistical characteristic can be expressed by using a two component Gaussian mixture:

$$p(w_i) = p_s \cdot g(w_i, 0, \sigma_s^2) + p_l \cdot g(w_i, 0, \sigma_l^2) \tag{1}$$

$$p_s + p_l = 1, \tag{2}$$

where the class of small coefficients is represented by subscript "*s*" and the class of large coefficients by subscript "*l*". The *a priori* probabilities of the two classes are represented by p_s and p_l respectively. The Gaussian component corresponding to the small coefficients has a relatively small variance σ_s^2 , capturing the peakiness around zero, while the component corresponding to the large state has a relatively large variance σ_l^2 , capturing the heavy tails. An EM algorithm as in [10] can then be applied to find out the Gaussian mixture model by obtaining the model parameters $[p_s, p_l, \sigma_s^2, \sigma_l^2]$.

The Gaussian mixture model is then used to find large coefficients. We select the significant coefficients by examining the probability of the coefficient belonging to the large coefficient class and its probability of belonging to the small coefficient class.

2.2 Near-Minimum-Distortion LUT Design Algorithm

First the wavelet coefficients to be embedded watermark are uniformly quantized. The quantization interval is *q* and the number of quantization cells is *M*. Then a LUT can be generated to embed the secondary data, i.e. each cell in the LUT carries either side information bit "0" or "1" which is denoted as *T*(*k*) for the *k-th* cell. Because the number of coefficients to be embedded "0" and "1" fall into each quantization cell is different from each other, different look-up tables will introduce different distortions. Let $P_{0,k}$ denote the probability of the coefficients to be embedded "0" fall into the *k-th* cell; and $P_{1,k}$ denote the probability of the coefficients to be embedded "1" fall into the *k-th* cell. Note that $\sum_{k=1}^{M} \left(P_{0,k} + P_{1,k}\right) = 1$. We may obtain the near-minimum-distortion *M*-level LUT by examining the probabilities $P_{0,k}$ and $P_{1,k}$, k=1,2,…,M. It can be seen that any coefficient will not be mapped to its corresponding cell if the side information bit is different from the bit that is carried by the cell. It is also noted that if a given coefficient is mapped into its corresponding cell, the quantization distortion is minimum. The difference $PD_k = | P_{0,k} - P_{1,k} |$ between the pair of probabilities: $P_{0,k}$ and $P_{1,k}$ determines the distortion cost we have to pay for LUT embedding. That is, the more PD_k is, the more distortion is there if we set the entry to the bit with less probability. Then each entry should be set to the bit corresponds to the higher probability in priority with regard to the probability difference.

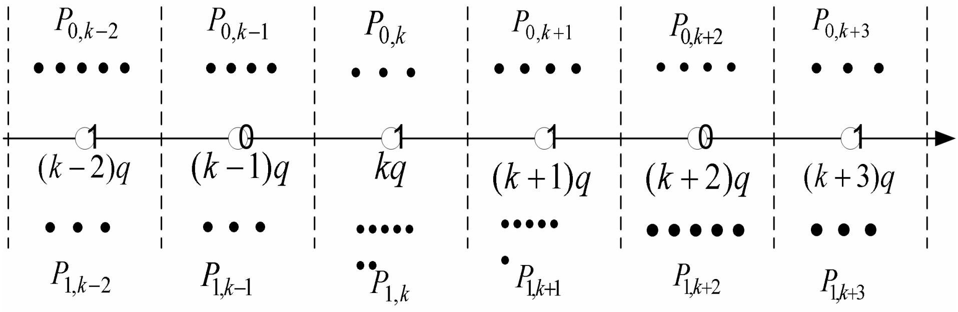


Fig. 1. Example of a run=2 near-minimum-distortion LUT

First sort PD_k in descending order. Then the look-up table is built in the way that the entry corresponding to the largest probability difference are set to the bit corresponds to the larger probability in priority. For example, if PDk is the largest in the queue currently and the look-up table entry Lookup(kq) is still not determined yet, we set Lookup(kq) = 0 if P0,k>P1,k, otherwise Lookup(kq) = 1. After each operation, we remove the largest probability difference PDk from the queue and move on the next. There is a rule we must keep in mind due to the maximum run constraint (r=2): the maximum run for "0" or "1" must be equal or less than 2. The algorithm for run of 2 can be summarized in the following three steps:

STEP 1: Calculate the difference PD_k between each pair of probabilities: $P_{0,k}$ and $P_{1,k}$.

STEP 2: Sort the absolute value of PD_k in descending order.

STEP 3: The largest absolute value of PD_k is found from the above queue (assume it is the *k-th* cell without loss of generality). If *Lookup*(*kq*) has been determined, go to step 4. Otherwise, among the determined lookup table entries, if any of the following situations occurs, $Lookup((k\text{-}2)q) = Lookup((k\text{-}1)q) = l$, $Lookup((k\ \text{-}1)q) = Lookup((k{+}1)q) = l$, $Lookup((k{+}1)q){=}Lookup((k{+}2)q) = l$, $l \in \{0,1\}$, *Lookup*(*kq*) is set to the complement of *l*: *Lookup*(*kq*) = *mod*(*l*-1, 2), no matter *j* value; otherwise *Lookup*(*kq*) = *j*, the original feature corresponds to $P_{j,k}$ is not shifted to other quantization points.

STEP 4: PD_k is removed from the queue. If the queue is not empty, go back to **STEP 3**.

To illustrate the above near-minimum-distortion LUT generation algorithm, an example is provided in Fig. 1. Note that the numbers of dots represent the probabilities.

3 Experimental Results

The proposed watermarking scheme is tested on seven 512×512 images of various types. We evaluate the quality of watermarked and attacked image by peak-signal-to-noise-ratio (PSNR), and the robustness under attacks is denoted by bit error rate

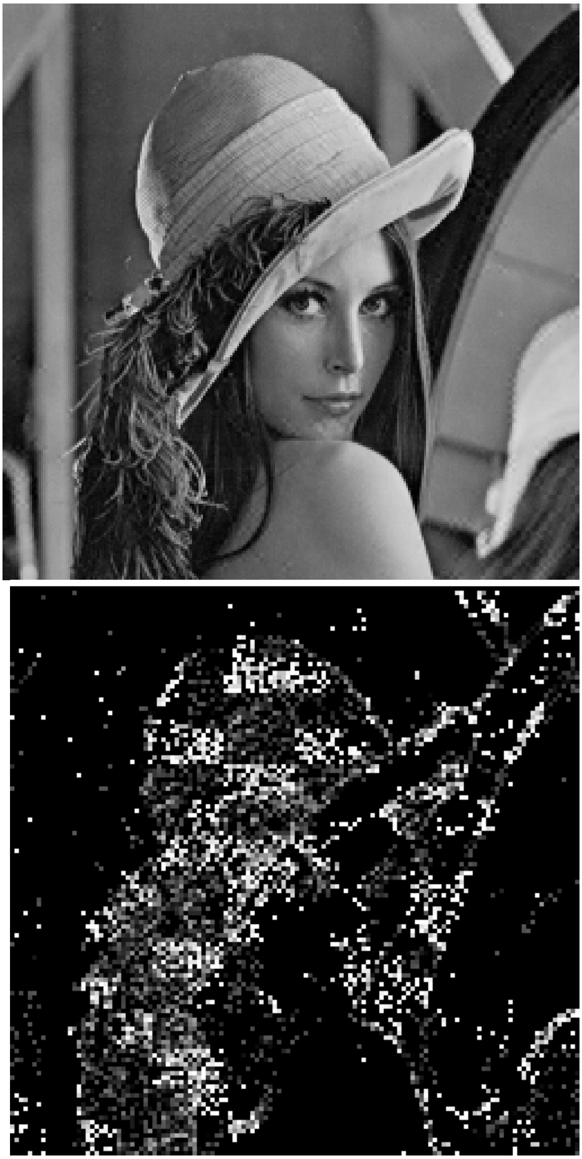


Fig. 2. The watermarked image and the difference from the original image with black denoting zero difference

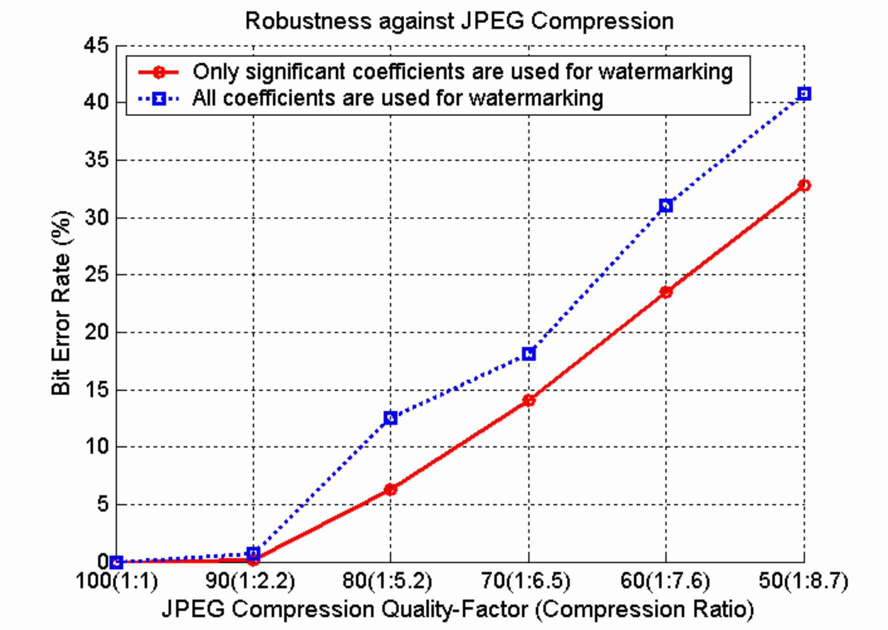


Fig. 3. Robustness against JPEG compression

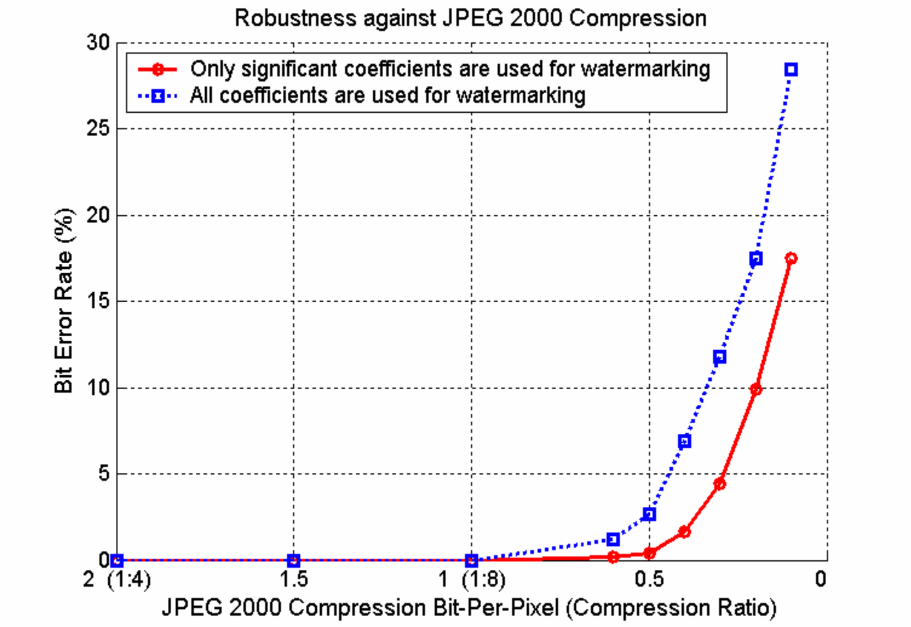
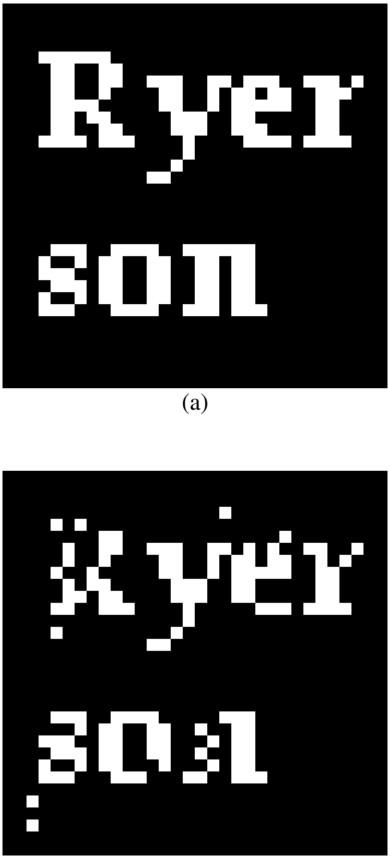


Fig. 4. Robustness against JPEG 2000 compression



(a)

(b)

Fig. 5. (a) The embedded watermark and (b) the watermark extracted from the JPEG 2000 severely compressed (1:40) image

(BER). First we inserted binary watermark into the images by applying the new method. Fig. 2 shows one example (Lena). The modified significant coefficients are mainly at the edges of the image. The watermark robustness to common operations such as image compression is tested. The discrete cosine transform (DCT) based coding system, JPEG baseline, and the discrete wavelet transform (DWT) based coding system, JPEG 2000, are the two compression attacks in our tests. We evaluate the robustness by the average BER for all test images. As shown in Fig. 3, the BER of the extracted watermark is less than 25% until the compression quality factor is smaller than 60. Embedding watermark in the wavelet domain, the presented algorithm has perfect robustness against DWT based JPEG 2000 compression attack when the compression is relatively less. The results are shown in Fig. 4. The decoded watermark can be 100% reconstructed after JPEG 2000 compression of 1bpp and is reliable until the compression bit-rate is smaller than 0.2bpp (1:40). The comparison of watermark embedding with (solid line) and without (dashed line) the statistical model based significant coefficient selection is shown in Figs. 3 and 4, where same embedding strategy is applied in the same wavelet subband. The advantages of the new method with the coefficient selection are apparent. Fig. 5 is an example of the watermark extraction with JPEG 2000 severe compression (1:40) and shows that the new scheme can survive JPEG 2000 compression very well.

Finally, we compared the distortion performances of the interleaving LUT (odd-even embedding, run=1), our near-minimum-distortion run of 2 LUT, and the average distortion of all run of 2 LUTs in terms of various quantization levels. According to Fig. 6, the fidelity of the new method is the best.

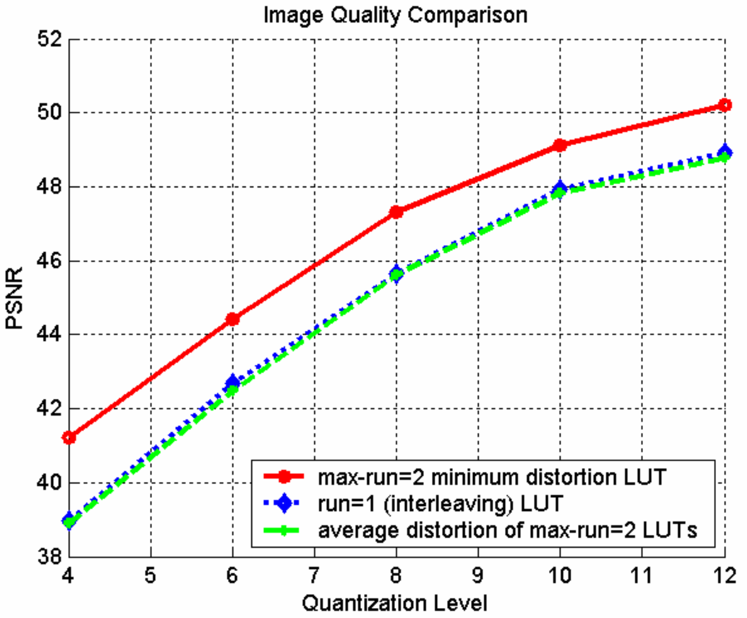


Fig. 6. The image quality comparison among run of 1 LUT embedding, run of 2 near-minimum-distortion LUT embedding and the average distortion of all run of 2 LUTs

4 Conclusions

We present a LUT based robust watermarking scheme with near minimum distortion, which can be integrated with the DWT based image coding standards such as JPEG 2000. A statistical model based coefficient selection method is proposed. Based on a Gaussian mixture model in the wavelet domain, significant and embeddable wavelet coefficients can be found at different wavelet scales. A novel near-minimum-distortion LUT design algorithm is presented, which can improve the quality of watermarked image. Experimental results show that our scheme is superior to the odd-even embedding in terms of image quality. Since the watermark is embedded on selected significant wavelet coefficients, it shows great robustness against general signal processing attacks.

References

1. B. Zhu, M. Swanson, and A. Tewfik: When Seeing Isn't Believing. IEEE Signal Proc. Mag., vol. 21, no. 2, pp. 40-49, Mar. 2004
2. J. Cox, J. Kilian, T. Leighton, and T. Shamoon: Secure Spread Spectrum Watermarking for Multimedia. In Proc. of the ICIP'97, pp. 1673-1687, October 1997
3. J. Cox, J. Kilian, T. Leighton, T. Shamoon: Secure Spread Spectrum Watermarking for Multimedia. IEEE Transaction on Image Processing, vol.6, no.12, pp.1673-1687, 1997
4. C. Podilchuk, W. Zeng: Image Adaptive Watermarking Using Visual Models. IEEE Journal Selected Areas of Communications, vol. 16, no.4, May, 1998
5. B. Chen and G. W. Wornell: Dither Modulation: A New Approach to Digital Watermarking and Information Embedding. Proc. of the SPIE: Security and Watermarking of Multimedia Contents, vol. 3657, pp. 342-353, Jan 1999
6. M. Yeung and F. Mintzer: An Invisible Watermarking Technique for Image Verification. Proc. ICIP'97, Santa Barbara, CA, pp. 680-683
7. M. Wu: Joint Security and Robustness Enhancement for Quantization Based Data Embedding. IEEE Trans. Circuits and Systems for Video Technology, vol. 13, no. 8, August 2003
8. X. G. Xia, C. G. Boncelet, and G. R. Arce: A Multiresolution Watermark for Digital Images. In Proc. of IEEE ICIP'98, pp. 548-551, Nov. 1998
9. H. J. Wang and C.-C. Jay Kuo: An Integrated Approach to Embedded Image Coding and Watermarking. In Proc. of ICASSP'98, Seattle, WA, USA, May 1998
10. H. Yuan and X. P. Zhang: Fragile Watermark Based on the Gaussian Mixture Model in Wavelet Domain for Image Authentication. In Proc. of ICIP2003, Barcelona, Spain, September 14-17, 2003
11. J. Romberg, H. Choi and R. Baraniuk: Bayesian Tree-structured Image Modeling Using Wavelet-domain Hidden Markov Models. IEEE Trans. on Image Proc., vol. 10, no. 7, July 2001

A Secure Image-Based Authentication Scheme for Mobile Devices

Zhi Li[1,2], Qibin Sun[1], Yong Lian[2], and D.D. Giusto[3]

[1] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore 119613
[2] Department of ECE, National University of Singapore, Singapore 119260
[3] Department of EEE, University of Cagliari, Cagliari 09123, Italy
{stuzl, qibin}@i2r.a-star.edu.sg; eleliany@nus.edu.sg; ddgiusto@unica.it

Abstract. Motivated by the need for designing secure and user-friendly authentication method for mobile devices, we present a novel image-based authentication (IBA) scheme in this paper. Its mnemonics efficacy rests on the human cognitive ability of association-based memorization. To tackle the shoulder-surfing attack issue, an interactive authentication process is presented. System performance analysis and comparisons with other schemes are presented to support our proposals.

1 Introduction

Today, the prosperity of e-business based on mobile terminals (e.g. PDA, smartphone and etc.) has boosted the development of secure and convenient user authentication solutions for touch screen devices. Traditional textual password or PIN, however, rely on keyboard as the input device. Many researchers thereby look at an alternative approach - graphical password, or image-based authentication (IBA) in a broader sense. Besides the convenience of password input, it is deemed more user-friendly in terms of memorability and recallability. The basic hypothesis is that human brain is more capable of storing graphical information than numbers or alphabets; in addition, IBA utilizes an easier and more human-friendly memorization strategy - *recognition-based* memory, instead of *recall-based* memory for textual password. In view of its potentials, JPEG (i.e. ISO/IEC JTC 1/SC 29/WG1) is considering to standardize such technologies [1].

We identify two mainstreams of state-of-the-art IBA approach up-to-date: *i)* click-based approach [2] [3] and *ii)* image-selection-based approach [4] [5] [6]. The former is based on sequential clicks of some points on an image, in which the location and order of the clicks are used as the password. In the latter approach, the user selects some "recognizable" secret images from a given image list. The whole authentication process consists of several rounds of such selections. One common problem with both approaches is that the password entropy is relatively small (see the evaluation in Section 5). This motivates us to design a new IBA scheme, which has large password entropy, and in the mean time, still preserve the user-friendliness.

In this paper, inspired by a classic mnemonics - *Method of Loci*, we present a novel *association-based* IBA scheme. The principal idea rests on the human

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 751–760, 2005.

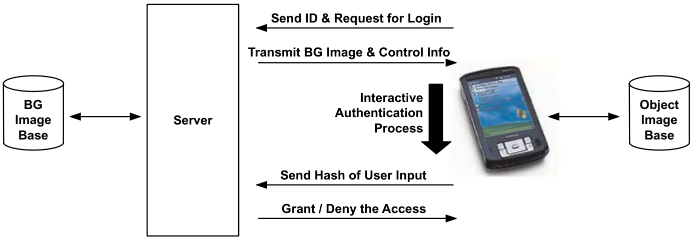

BG Image Base
Server
Send ID & Request for Login
Transmit BG Image & Control Info
Interactive Authentication Process
Send Hash of User Input
Grant / Deny the Access
Object Image Base

Fig. 1. Overview of the proposed IBA framework for mobile devices. The object images (see Subsection 3.1) are stored locally whereas the background (BG) images are transmitted.

cognitive ability of association-based memory. The mnemonic efficacy is enforced by creating "bounds" between the password elements, which is analogous to splitting a telephone number into chunks to aid memorization. Note that as will be addressed in Section 3, the password entropy is not necessarily reduced.

This scheme also tackles another issue which is commonly engaged in many user authentication processes – shoulder-surfing (SS) attack, namely, the person behind your shoulder can observe and remember your input, and impersonate you afterwards. We realize that this problem is similar to the problems solved by the zero-knowledge proof protocols in cryptography [7]. An interactive authentication process with mechanism to prevent full secret unveiling during the input process is thereafter presented.

Also note that the transmission data size is an important factor to consider for mobile devices. In the proposed IBA scheme, some reusable images are kept in local image base to avoid unnecessary transmission and speed up the authentication process. On the other hand, the images to be transmitted are coded in JPEG2000 format to facilitate progressive decoding on the mobile devices, namely, the image with coarser resolution will appear first, followed by the emergence of finer and finer details. In this way, experienced users are able to complete the authentication process in shorter time. In case of network traffic jam, the efficacy of the authentication algorithm will not be affected, since the interactive authentication process is locally conducted and there is no data exchange during this process.

This paper is organized as follows. Section 2 gives an overview of the proposed authentication framework and introduces a typical application scenario. In Section 3 and 4, we discuss the interactive authentication process in detail. We propose a primary authentication scheme and a modified authentication scheme which is resistant to SS attack in each section respectively. Section 5 compares our designs with some prior related work. Section 6 concludes this paper.

2 Authentication Framework Overview

Refer to Fig. 1, the whole authentication process consists of two rounds of data exchange between the server and the mobile device. Initially, a request to login together with the user's ID is sent to the server. The server retrieves the ID's corresponding registered background (BG) image, generates some control information such as the object image appearance order, and send them to the mobile device. Thereafter the interactive authentication process takes place locally at the mobile device end. The user's input sequence is then hashed and transmitted to the server. Lastly, the server verifies the hash and grant or deny the access.

The typical application scenario involves two more parties - Alice and Bob. Alice's objective is to authenticate herself to the server via the mobile device. The server's job is to verify whether the person trying to authenticate herself/himself is Alice or another impersonator. Bob - the shoulder-surfer - is to obtain the password shared between Alice and the server such that he could impersonate Alice. Assume that Bob is capable of observing the full interactions between Alice and the mobile device. For example, he has set up a hidden camera such that he can capture all the details of the mobile device's display and Alice's input.

3 Primary Authentication Scheme

3.1 Description

In the user registration phase, Alice is required to pick a desirable BG image. The image is partitioned into some small regions, each partition being a locus. Define the locus alphabet as the set of all the loci $\mathbf{L} = \{l_1, l_2, \ldots l_{|\mathbf{L}|}\}$. Also define an object alphabet $\mathbf{O} = \{o_1, o_2, \ldots o_{|\mathbf{O}|}\}$ and a color alphabet $\mathbf{C} = \{c_1, c_2, \ldots c_{|\mathbf{C}|}\}$. The object alphabet consists of clip-arts images of objects, such as images of a cup, a bike, a cat and etc. The color alphabet consists of colors like red, blue, green, cyan etc. To create her unique password profile, Alice is then required to create N triplets, each triplet with one element chosen from each alphabet $\phi_n = \{l'_n, o'_n, c'_n\}$, for $1 \leq n \leq N$. Note that Alice tends to choose some "salient points" as the pass loci, therefore, in practice, l'_n is selected from a subset $\mathbf{L}' \subset \mathbf{L}$.

Note that in implementation, the object images (i.e. the object alphabet $\mathbf{O}$) are stored locally in the mobile device. The object images are coded in bi-level format in order to save storage space. In addition, they can be rendered by the colors defined in the color alphabet $\mathbf{C}$.

Fig. 2 schematically illustrates the authentication process (Step 1 → 2 → 3A). The authentication phase consists of N rounds. Triplet ϕ_n serves as the "pass triplet" for round n, with l'_n, o'_n and c'_n being the pass locus, pass object and pass color, respectively. In round n, Alice needs to click on the region of the BG image associated with the pass locus l'_n (Step 1). After the click, a window pops up, showing a list of object elements $\mathbf{O}_1 \subset \mathbf{O}$, including the pass object $o'_n \in \mathbf{O}_1$. The remaining subset $\mathbf{O}_2 = \mathbf{O}_1 \setminus \{o'_n\}$ is called the decoy object set. Alice needs to select the pass object o'_n from the list (Step 2). After the selection, another window pops up, showing a list of color elements $\mathbf{C}_1 \subset \mathbf{C}$, including the pass color $c'_n \in \mathbf{C}_1$. Similarly, the remaining subset $\mathbf{C}_2 = \mathbf{C}_1 \setminus \{c'_n\}$ is the decoy color set. Alice needs to correctly select the pass color c'_n (Step 3A). This procedure repeats for N rounds. Alice is verified as authentic only when all

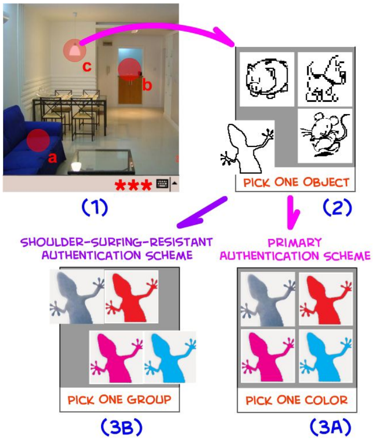

c
b
a
***
PICK ONE OBJECT
(1)
(2)
SHOULDER-SURFING-RESISTANT AUTHENTICATION SCHEME
PRIMARY AUTHENTICATION SCHEME
PICK ONE GROUP
PICK ONE COLOR
(3B)
(3A)

Fig. 2. Schematic illustration of the interactive authentication process

the pass loci are correctly clicked, and all the pass objects and pass colors are correctly selected.

3.2 Analysis

We consider user-friendliness and security as two mutually contradictory design goals for any user authentication system. We now analyze the primary authentication procedure based on these two criteria.

1) User-friendliness – Due to its nature, graphical password is considered advantageous over traditional textual password in terms of memorability. However, if the authentication process is too tedious (e.g. too many rounds of selection), it may still create memorization difficulties and annoy Alice. Our goal is to simplify the authentication process and create solid mnemonic effect, while still maintaining the password entropy large enough.

A classic mnemonic strategy called *Method of Loci* has attracted our attention. This method is described as follows:

First of all, choose a familiar place such as your own house. Take a mental walk through the rooms, and pay particular attention to the details that makes your mental images more vivid. Along the route create a list of loci, i.e. well defined parts of the room that you can use later to remember things, such as a door, a bed, an oven etc.

Now, when you are faced with a list of items to be memorized, you must form visual images of them and place them, in order, on the loci in your route. A loaf of bread sticking out of the letterbox; a giant apple in place of the door, etc. More striking the created image, more easily you will remember the thing.

This mnemonics can be dated back to the ancient Greeks, and has been proven to be surprisingly useful. For example, in an experiment targeted at college students [8], the subjects using Method of Loci frequently recall two to seven times as much as the controlled subjects. In [8], Bower systematically studied the Method of Loci from a psychological point of view. He identified that the most essential part of this method is "the formation of imaginal associations between known cues and previously unknown list items at input, and use of these cues for recall."

In our proposed primary authentication scheme, two levels of association are created – the association between the locus and the object, and the association between the object and its color. By using mnemonics technique similar to the Method of Loci, Alice could remember the associated locus, object and color as a "bundle", rather than in separation. In [8], Bower gives some very useful tips for establishing such associations – *i)* visualization must be conducted, no matter whether the user has witnessed the scene in real life. *ii)* The objects must be depicted in some kind of "interacting unity". For example, "a doll waving a red flag" is easier to be remembered than "a doll sits beside a flag that is red". Note that arbitrary associations may create some "bizarre scenes" (e.g. a blue banana in the bath), but remember that as addressed in the Method of Loci, more striking the created image, more easily you will remember the thing. Therefore, Alice is encouraged to create "bizarre scenes" to enhance the mnemonic effect. Another great advantage of this strategy is that the password can be unbiasedly distributed among users and thus the password entropy can be maintained.

We argue that this association-based approach is superior compared to the recall-based approach, since associative memory is what the human is better at. On the other hand, it is superior compared to the recognition-based approach, because it leaves Alice much more choices of action to take, leading to much larger password entropy.

2) Security – We address two types of security measurement here: *i)* the password entropy, which measures the probability that Bob obtains the correct password based on random guessing; *ii)* resistance to SS attack, in terms of the number of observations Bob needs, in order to interpret the correct password.

The password entropy can be calculated as follows: for simplicity, assume all passwords are evenly distributed. Then the entropy is:

$$H(X_{std}) = Nlog_2(|\mathbf{L}'||\mathbf{O}_1||\mathbf{C}_1|) \tag{1}$$

For a typical application, suppose the size of the salient point set of an image $|\mathbf{L}'|$ is 30, $|\mathbf{O}_1|$ and $|\mathbf{C}_1|$ are both 4, and the number of rounds N is 4, the entropy is computed as 35.6 bits, which is equivalent to the entropy of a 6-digit textual password. However, note that the above analysis is valid only under the following rule. In our scheme, the subsequent display of the object list $\mathbf{O}_1$ and the color list $\mathbf{C}_1$ after the locus selection may probably leak some information to Bob. For example, during Bob's random trials of the password, if Bob observes two different lists of objects displayed after the same input in two different trials, he may be able to interpret the pass object by intersecting the two lists. One useful rule to work against this attack is to make sure that the display of the object and color list must be "invariant", i.e., in Round n, $\mathbf{O}_1$ and $\mathbf{C}_1$ are deterministic functions of n and Bob's input (i.e. click or selection) in that current round only. More precisely, the object list can be determined by:

$$\mathbf{O}_1 = \begin{cases} \{o'_n\} \cup \mathbf{O}_2 \sim h_1(n, l_B), & \text{if } l_B = l'_n \\ \mathbf{O}_2 \sim h_2(n, l_B), & \text{otherwise} \end{cases} \tag{2}$$

where l_B is the locus Bob clicks on, and $h_1(\cdot)$ and $h_2(\cdot)$ are two one-way hash functions returning $(|\mathbf{O}_1| - 1)$ and $|\mathbf{O}_1|$ elements, respectively. The color list can be derived in a similar way.

The above analysis presents the primary scheme's security level against random-guessing attack. Now for the SS attack, in this scheme, after Bob observes the authentication procedure once, the password is fully revealed, thus this scheme is susceptible to the SS attack as all other schemes previously mentioned. One solution to this problem will be presented in the next section.

4 Authentication Scheme Resistant to Shoulder-Surfing Attack

The SS attack urges us to look for a new approach to work against it, and in the mean time, we still want to preserve the primary authentication's user-friendliness. In [9], a challenge-response-based graphical password scheme is proposed to counter the SS attack. However, the proposed procedure involves several rounds of jigsaw-puzzle-like challenges, which is practically not very feasible. Another shoulder-surfing-resistant method based on PIN entry [10] is proposed recently. This scheme's feasibility is based on human's cognitive limitation on short term memory. However, we notice that to challenge human's memorability, the authentication procedure has to be intentionally tedious. Besides, they also proposed a probabilistic cognitive trapdoor game approach, which still suffers from the same tedious input procedure. In this section, we propose a method that is only a slight variant of the method proposed in Section 3, but the SS resistance property is nicely achieved.

4.1 Principle

We realize that the SS problem is similar to the problems solved by the zero-knowledge proof protocols in cryptography [7]. The principle of this protocol is: if Alice wants to prove to another verifier her knowledge of some secret information, but without revealing the secret's detail to the verifier, she can prove it by solving a "hard problem" - the "hard problem" is a special question that is easy to solve if the secret is known, and extremely hard if unknown. Therefore, by solving the problem, Alice can thereby prove her knowledge of that secret. Note that the "hard problem" must be carefully designed such that the verifier cannot get any information about the secret by observing Alice's solution. The SS resistant authentication involves a slightly different situation of the zero-knowledge proof protocol (see Fig. 3). The basic idea is that if Alice can prove to the server that

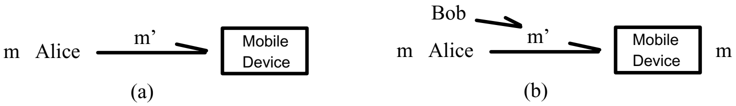


Fig. 3. Comparison of application scenarios of zero-knowledge proof protocol *(a)* and SS resistant authentication *(b)*, where m is the secret information (i.e. the password) and m' is the solution to the "hard problem" (i.e. the authentication input)

she knows some secret information but without revealing it during the process of proof, she can authenticate herself to the server, and in the mean time, avoid revealing this information to the shoulder-surfer Bob.

4.2 Description

The next job is to design the "hard problem" which is secure and does not complicate the authentication process. We propose the following solution: randomly cluster the colors in the color list $\mathbf{C}_1$ into size-K subsets $\mathbf{C}_{1,i}$, for $1 \leq i \leq |\mathbf{C}_1|/K$, where $|\mathbf{C}_1|$ is the size of set $\mathbf{C}_1$. For convenience, choose $|\mathbf{C}_1|$ and K such that K divides $|\mathbf{C}_1|$. The "hard problem" is to select the subset that contains the pass color c'_n. Consider that Alice knows the pass color, so choosing the right subset is an easy job; however, since Bob does not know the pass color, he has no clue which subset to choose, but only can take his chance to guess. Note that this "hard problem" is not "perfect" because it is not fully secret-concealable – Bob still can get some information about the right pass color during the observation (i.e. narrowing down the possible pass colors to the subset selected by Alice). In the next subsection, we shall analyze the security level of this scheme. The new authentication procedure is illustrated as Step 1 → 2 → $3B$ in Fig. 2. In step $3B$, instead of asking Alice to select the correct pass color, now we ask Alice to select the correct subset that contains the pass color.

4.3 Analysis

Since this SS resistant authentication is only a slight variant of the primary scheme, we assume they have the same level of user-friendliness. We focus on analyzing the system's security level in terms of *i)* password entropy *ii)* resistance to SS attack. The password entropy can be calculated as:

$$H(X_{ssr}) = Nlog_2(|\mathbf{L}'||\mathbf{O}_1||\mathbf{C}_1|/K) \tag{3}$$

Compare to Eq. 1, we notice that the password entropy has reduced and thereby facilitated Bob's opportunity to interpret the password by random guessing. Nevertheless, this modified approach provides resistance to SS attack, so we shall consider that this approach trades-off random-guessing security with SS security.

We then measure the resistance to SS attack in terms of how many times Bob needs to observe Alice's authentication process, in order to interpret the correct password. The best case happens to Bob when in the second observation, for every round, the random clustering puts the pass color in a totally different subset without any overlapping decoy colors as in the first observation. In this case, Bob needs to observe twice to interpret the right password; in the worst case, however, when there are always overlapping decoy colors, it takes infinite observations for Bob to discover the password.

We also want to know the average number of observations needed. Define $P_{rnd}(M)$ as the probability that Bob reveals the pass color for a single round after M observations. Then the probability of revealing all the pass colors in less or equal to M observations is:

$$P_{all}(m \leq M) = [\sum_{m=1}^{M} P_{rnd}(m)]^N \tag{4}$$

Therefore the probability of revealing all the pass colors in M observations is:

$$P_{all}(M) = P_{all}(m \leq M) - P_{all}(m \leq M - 1) \tag{5}$$

The average number of observations needed is:

$$\bar{M}_{all} = \sum_{m=1}^{\infty} m P_{all}(m) \tag{6}$$

To find a general $P_{rnd}(M)$ is difficult. Consider the case $|\mathbf{C}_1| = 4$ and $K = 2$, then $P_{rnd}(M)$ can be found as:

$$P_{rnd}(M) = (2/3) \times (1/3)^{M-2} \tag{7}$$

In the proposed system, setting $N = 4$, the average number of observations needed is computed as 3.3913. That is, on average Bob needs to observe more than three times in order to interpret the right password. In practice, the chance is rare for Bob. Therefore, we consider our system as secure against SS attack. Moreover, higher security level can be easily achieved by increasing the color list size (e.g. by setting $|\mathbf{C}_1| = 6$ and $K = 3$) or increasing the number of rounds N (e.g. by setting $N = 6$).

5 Comparisons with Prior Work

In this section, we compare our proposed schemes with some prior related work in literature.

The calculation of password entropy for various methods is in Table 1. For [9], due to its ambiguous nature, we will not give quantitative analysis and comparison here. Fig. 4 presents a comparison of various methods in a 3D plot. The evaluation is based on *i)* user-friendliness, *ii)* security against random guessing *iii)* security against SS. Since the user-friendliness is subjective to users, we only present some qualitative analysis (by giving score 1 to 5, 5 being the most user-friendly) based on our understanding.

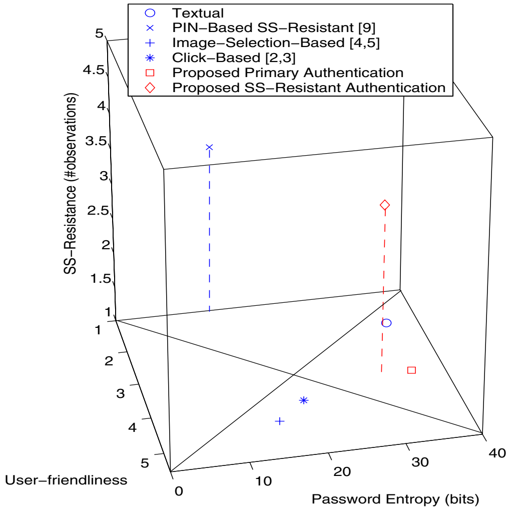


Fig. 4. Comparison of various authentication schemes in terms of user-friendliness, password entropy and SS resistance

Table 1. Comparison of Password Entropy

Password Scheme & Descriptions	Password Entropy (bits)
Textual. 6 numbers or alphabets.	$6 \times log_2 62 = 35.7$
PIN-based SS-resistant. 4 digits. [10]	$4 \times log_2 10 = 13.3$
Image-selection-based. 5 runs, in each run select 1 from 9 images. [4] [5]	$5 \times log_2 9 = 15.8$
Click-based. 4 loci (30 salient points). [2] [3]	$4 \times log_2 30 = 20.0$
Proposed primary authentication. 4 loci, 4 objects, 4 colors (30 salient points).	$4 \times log_2(30 \times 4 \times 4) = 35.6$
Proposed SS resistant authentication. 4 loci, 4 objects, 4 colors, K=2 (30 salient points).	$4 \times log_2(30 \times 4 \times 2) = 31.6$

6 Conclusion

In this paper, we proposed a novel IBA scheme for mobile devices. Its mnemonics efficacy rests on the *association-based* memorization strategy. We also presented an interactive authentication process which successfully tackled the SS attack, but without adding extra complexity to the whole procedure. Our future work includes conducting user studies and experiments to examine the effectiveness of our methods.

References

1. G. Ginesu, D. Giusto, T. Onali: Image Based Authentication (IBA): A Review. N3461, ISO/IEC JTC 1/SC 29/WG1, Nov, 2004
2. G. Blonder: Graphical Passwords. United States Patent 5559961 (1996)
3. http://www.viskey.com
4. The science behind Passfaces. Real User Corporation (Sept. 2001) http://www.realuser.com
5. R. Dhamija, A. Perrig, Déjà Vu: User study using images for authentication. 9th USENIX Security Symposium, 2000
6. W. Jansen, S. Gavrila, V. Korolev, R. Ayers, R. Swanstrom: Picture Password: A Visual Login Technique for Mobile Devices. NISTIR 7030
7. B. Schneier, Applied Cryptography. New York: Wiley 1996
8. G. H. Bower: Analysis of a Mnemonic Device. 496-510, American Scientist, Sep,Oct 1970
9. L. Sobrado and J.C. Birget: Graphical passwords. The Rutgers Scholar, vol. 4, 2002
10. V. Roth, K. Richter, R. Freidinger: A PIN-Entry Method Resilient Against Shoulder Surfing. 11th ACM Conference on Computer and Communications Security (CCS'04), Washington DC, USA, Oct, 2004

A Print-Scan Resistable Digital Seal System

Yan Wang and Ruizhen Liu

Assuredigit Technology Co., LTD.,
Room 717, No.477, Wensan Road, Hangzhou Zhejiang 310013, P.R.China
{wangy, liurz}@assuredigit.com

Abstract. Digital seal can be used to assure the authenticity, integrity and undeniability of electronic documents. In this paper, we implement a new digital seal system called as ASS (AssureSeal System) based on digital watermarking and digital signature technologies. ASS can protect not only the electronic documents, but also their paper copies through digital signature as well as robust and reliable digital watermarking technology. First we introduce the framework and the function modules of ASS, then propose a new technology that adds seal image into electronic documents, which combines the patented binary digital watermarking algorithm with normal digital signature technology, integrated PKI digital certificate. At last, we introduce the specifications and applications of ASS.

1 Introduction

With the maturity of Internet and the development of multimedia technology, network becomes an indispensable platform of people's work and life. EC (electronic commerce), as a kind of simple, convenient, low-cost electronic trading mode, is entering into people's daily life. Online trading is widely used in e-government, online-bank, stock market, enterprise's logistics, and even personal business activities. In these cases, the contracts or the documents are often submitted or transferred in electronic form rather than paper form. Thus, traditional handwritten sign and stamp are not suitable to them, which should be replaced by a novel tool. Therefore, digital seal appears to be the suitable one to solve these problems. Digital seal is the visual representation of electronic signature that is done by asymmetric encryption technology. It appears to be an elaborate electronical form, which has logical relationship to its host document and can be used to identify the document signer, guarantee the digital data's integrity and illustrate the agreement to its content. It performs the same function as traditional handwritten sign and stamp. Digital seal has extremely important function in protecting the authenticity, integrity and undeniability of the electronic documents.

However, for important and secret documents, the trading partners still tend to use paper documents to keep evidence instead of electronic ones. Even for normally used electronic data protected by digital seal, the traditional digital seal system can only protect electronic documents, not paper documents.

We will introduce a new digital seal system called as ASS (AssureSeal System), based on digital watermarking [1] and digital signature technology [2], [3]. ASS has some new advantages over other digital seal systems. ASS can protect not only the

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 761 – 770, 2005.

electronic documents, but also their paper copies through robust and reliable digital watermarking technology. The digital signature of the document is considered as digital watermark, which is embedded into the seal image. Different document has different digital watermarks to be embedded. The seal image, digital certificate and the private key are stored in E-Key (the hardware medium, known as electronic key). Only the legal owner with the right password can operate the key and add watermarked seal image into host documents. The digital certificate in E-Key can identify the document signer and the document source. Thus it can verify the integrality of the document, prevent various kinds of forgery and tamper.

The paper is organized as follows: Section 2 gives the framework of ASS; Section 3 introduces the function modules ; Section 4 describes three key technologies used in ASS; Section 5 describes the specifications, characteristics and applications of ASS; Section 6 concludes this paper.

2 Framework of ASS

In order to guarantee the integrality, usability and security of the whole system, many aspects were considered when designing it, such as the security of the system, the structure security of the platform, the convenience and the controllability for the users, etc. Fig. 1 shows the framework of ASS.

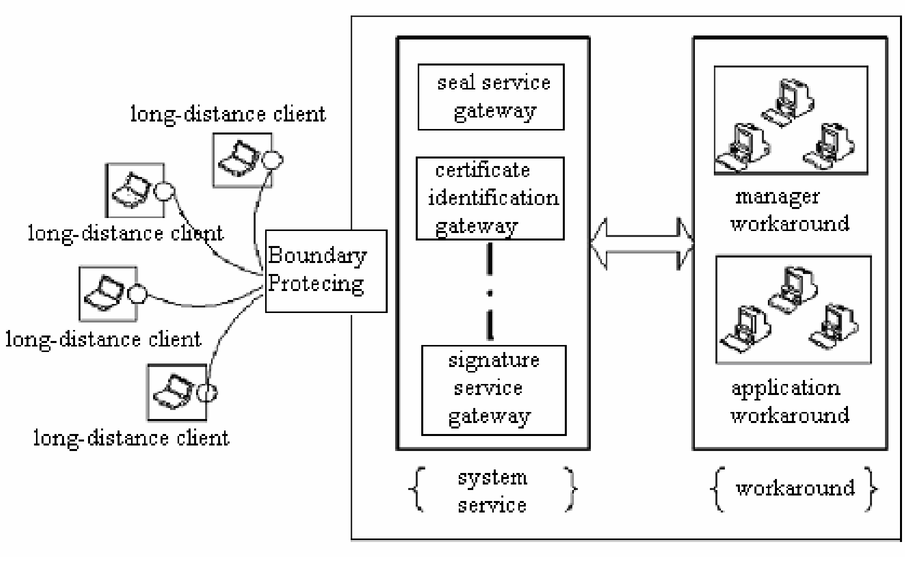


Fig. 1. The framework of ASS

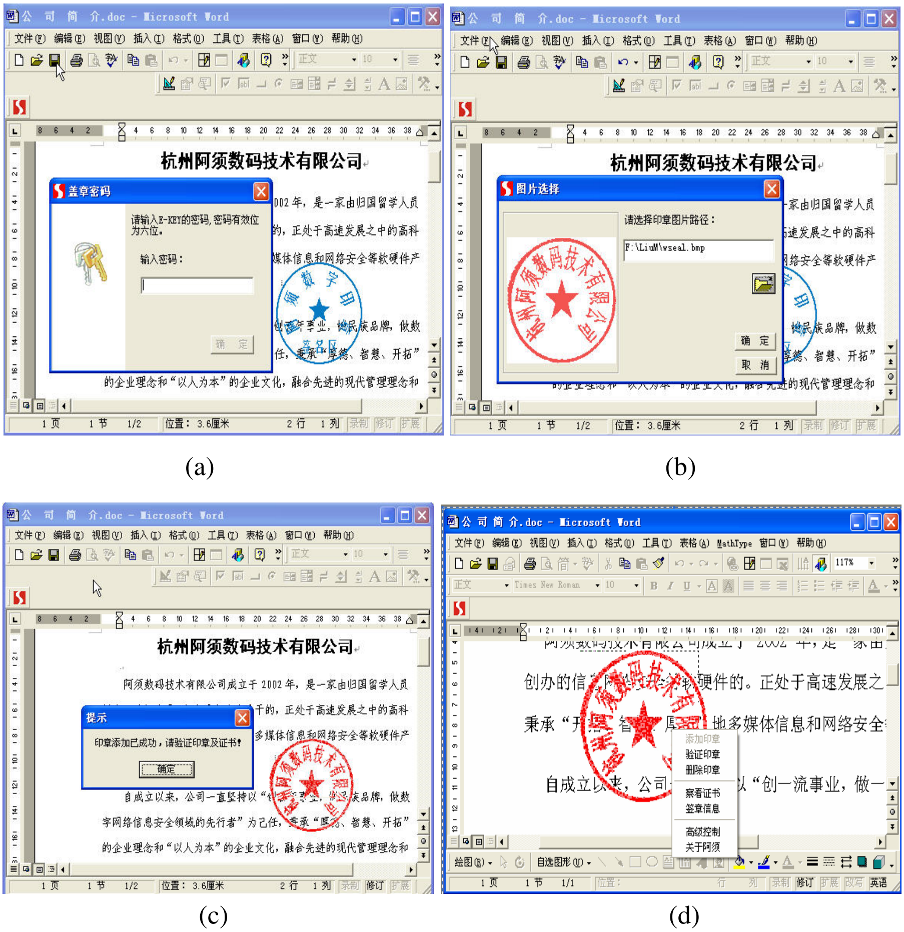
The whole system includes four parts:

- ASS Management Server: it is used to perform the work that is related with digital seal management, such as producing, storing and removing the seal image;
- ASS Manager Terminal: used to maintain ASS by administrator;

- ASS Seal-Maker Terminal: in order to guarantee the security of the whole system, the operator is considered as one unit to finish the work, like producing and removing digital seal;
- ASS Client Terminal: download the seal and add the seal to the document, etc.

3 Function Modules

After installing the ASS software, the seal button appears in the toolbar of OFFICE software (such as MS-OFFICE, WPS etc.) or the Internet browser. Its form is plug-in package or controlling component. Through plug-in, the signer can add seal image to the document after E-Key is plug into the computer, select the seal position, control



请输入E-KEY的密码，密码有效位为六位。

印章添加已成功，请验证印章及证书!

阿须数码技术有限公司成立于 2002 年，是一家由归国留学人员创办的信息网络安全软硬件的，正处于高速发展之中的高科技企业。……多媒体信息和网络安全等软硬件产……

自成立以来，公司一直坚持以“创一流事业，做民族品牌，做数字网络信息安全领域的先行者”为己任，秉承“厚德、智慧、开拓”的企业理念和“以人为本”的企业文化，融合先进的现代管理理念和

Fig. 2. (a) input E-key password; (b) choose path of the seal image; (c) the seal has been added successfully;(d)verification and view the seal information

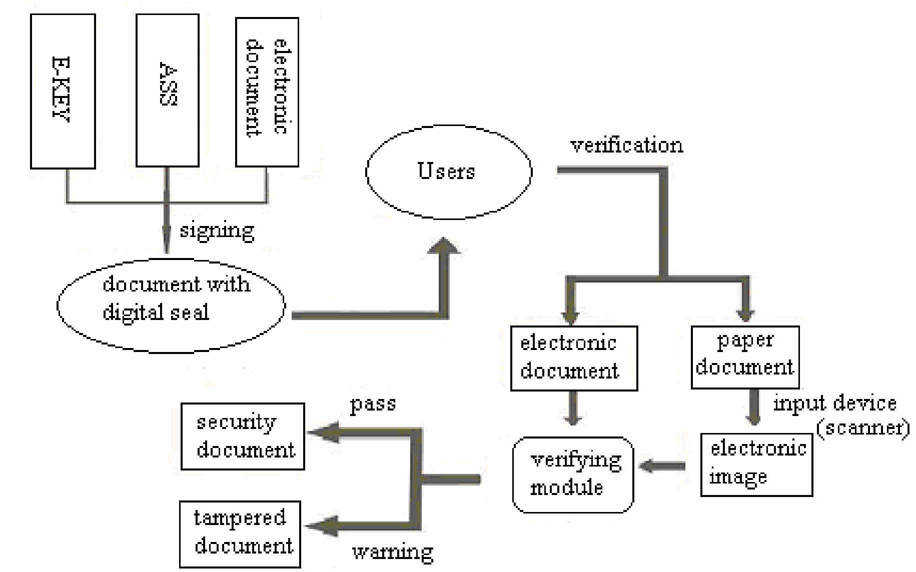


Fig. 3. Users sign the electronic document using ASS with E-Key to get the signed document, and they can verify the signed electronic document or its paper copy by scanning it into electronic image, to check if it is tampered

the printing permission and verify the integrity of the signed documents. The system supports two kinds of verification: electronic verification and paper verification. For verification, plugging E-Key into the computer is an option. One document can be signed more than once, which means that more than one seal image can be added into one document. Fig.2 shows the process of adding the seal to a document in MS-Office.

3.1 System Function Modules

Document-signing Module: the signer signs the electronic documents with E-Key. It is similar to traditional stamping on paper documents. Signature operation is performed within E-Key, in order to ensure the security of the signing process;
Document-verifying Module: it supports two kinds of verification: electronic and paper verification. The former verifies the electronic document directly; the latter needs to convert the paper document to electronic version, and then verify it;
Signer-identifying Module: verifies the digital certificate and signature information (signing time, signer's information, information about the digital seal, etc.), so as to preventing from faking identity;
Seal-setting Module: the signer can set the position of the seal and remove the seal , as well as lock the document before signing it, but when removing the seal, PIN code is needed in order to check the permission.

Print-controlling Module: it is mainly used to control the print permission. Each signer can set the print permission, but once one signer get the permission, the others lose this print control automatically. Before using this module, it needs to identify the operator's identity, only the person who signs the document can set the authority.

The structure and function of ASS is shown in Figure 3.

4 Key Technologies

There are three key technologies used in ASS to implement its function: digital watermarking, digital signature and digital certificate technologies. Digital watermarking ensures the digital seal can't be forged, which is a kind of popular digital anti-forgery technology [4], [5], [6]; digital signature is a secure technology that uses digital certificate to check tamper, integrity and document authentication [7].

4.1 Digital Watermarking

Digital watermarking has been proposed as a solution to the problem of copyright protection of multimedia documents in networked environments [1], [8]. It makes possible to tightly associate a code (called a watermark) to a host data allowing the identification of the data creator, owner and so on. The host data could be images, audio or video, etc [9], [10]. In ASS, host data is the seal image, while watermark could be electronic signature of the documents, digital certificate or other important information.

4.1.1 Principle

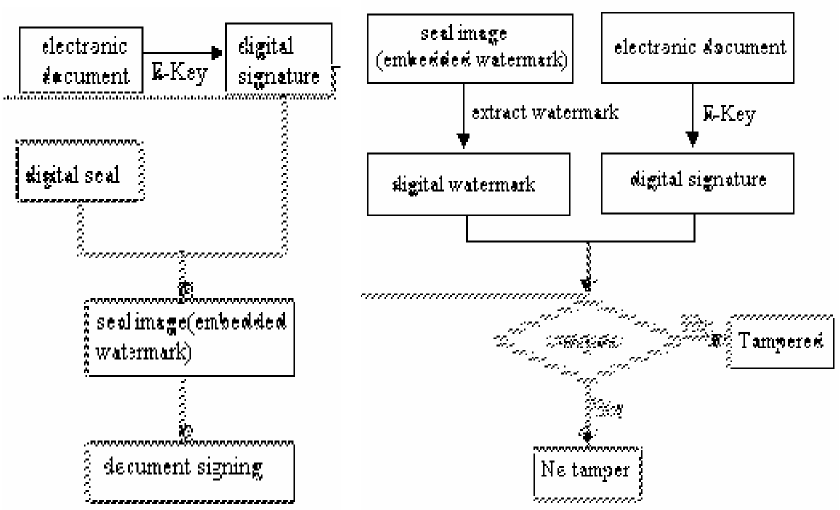
The digital watermarking algorithm used in ASS is the patented binary digital watermarking algorithm [11]. In order to extract the embedded watermark correctly after the document has been printed, error-correcting code technology (such as BCH [12] or RS[13] algorithms) is used to strengthen error-correcting ability and robustness of watermark. We also have considered the privacy of coding way and embedded position of the digital watermark, combining with cryptography; it can resist various kinds of common attacks.

4.1.2 Algorithm Descriptions

ASS is independent to network operation; its procedure mainly includes document signing, document verifying and permission controlling.

Document Signing Procedure: first obtain the content of the document, and calculate the digital signature using the digital certificate in E-Key, then embed it into the digital seal image with binary digital watermarking method.

Document Verifying Procedure: first extract the watermark from the seal image, then check if the document has been tampered by comparing the extracted watermark with the digital signature of the document. Seeing Fig. 4.



(a) (b)

Fig. 4. (a) document signing procedure, (b)document verifying procedure

4.2 Digital Signature

Digital signature is used to resolve undeniability and to prevent forging, tampering and signature imitation.

It has the following specifications:

- The signer can't deny his signature after signing to the information;
- The receiver can verify the authenticity of the signed information from the sender;
- Because of the security of PKI, so long as it is secret to the signer's private key, that will prevent others forging the signature;
- No one can tamper the original information, which will result in the invalidation of the signature.

In ASS, SHA-1 [14] is selected as the hash function, and RSA [14] as public key algorithm. These two algorithms are option. Others can also be used in the case of according with the standard of digital signature, such as PKCS7 or X-509 etc [14]. The signature operation is performed in the computing unit in E-Key, and the private key also is stored in E-Key, which guarantees the security of digital signature.

4.3 Digital Certificate

The digital certificate [15] is a kind of key management media of PKI system. It is an authoritative electronic file, used for proving the signer's identity and legitimacy of the public key. The certificate connects the signer's identity with the public key, thus offers an authoritative assurance for signer's identity. Its format is decided by CA (certificate agent), which is a credible organization and used to promulgate the digital certificate.

We use the digital certificate promulgated by Zhejiang CA in ASS. The main content included in the digital certificate is as follows: certificate version number, certificate serial number, signature algorithm, certificate promulgator, the validity period of the key, subject information of the certificate, the information about the public key, etc. When verifying the signer's identity by checking the certificate, it doesn't need E-Key, because ASS has added the content of the certificate to the software component when signing for the document.

5 Specifications and Applications

5.1 Specifications

1. Large capacity: it can embed digital watermarks more than 4000bits in one seal image;
2. Strong robustness: watermark information can also be extracted completely after printing, scanning or JPEG compression, 100% extracting success rate, 100% detecting success rate, the accuracy of measuring can reach one byte;
3. The digital watermarking algorithm is simple and more efficient, combining with the mature encryption algorithm, it can resist various attacks;
4. Easy expansibility: it can offer SDK products or modules which contain expansible API interfaces, adapting different application fields, like e-government, e-commerce and different enterprises;
5. The format of the seal image can be BMP□TIF□JPG;
6. ASS Management Server supports Linux/Unix operation system, Web Server uses Apache, Tomcat, and based on JSP programming.

5.2 Charateristics

ASS has some new advantages and better results and efficiency compared with other systems existing in the same field. The most important is that it supports paper verification, and its efficiency is better. The table and seal images below give the test results.

Table 1. Different parameters were set and the test result is shown in following table. 2320 bits watermarks were embedded in one seal image. " Yes " means the watermarks embedded in the seal can be extracted correctly, contrarily, " No " means the test is not passed.

Seal Size(cm)	**BenQ3300v scaner(dpi)**	**HP LaserJet 6L printer**	**Epson c63 printer**
3.0	100	No	No
3.5	200	Yes	No
3.75	300	Yes	Yes
3.0~3.75	>300	Yes	Yes

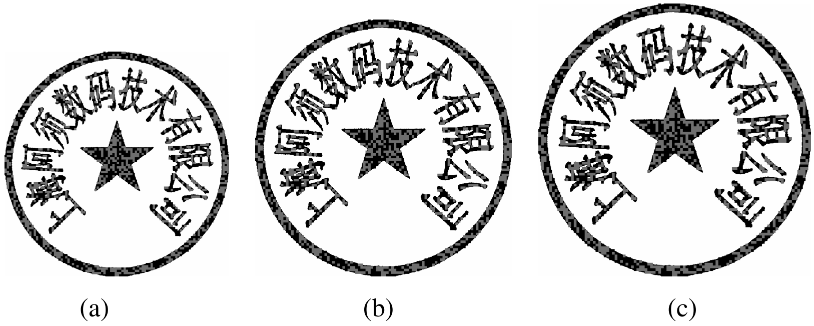


Fig. 5. three watermarked seal images. (a) seal image of size 3.0cm, (b) seal image of size 3.5cm, (c) seal image of size 3.75cm.

The experimental result shows:

1. Printable: supporting electronic and paper verification. It doesn't need original seal image when extracting digital watermark;
2. Secure: ASS use popular PKI system. Private key and digital certificate are stored in E-Key, the information in E-Key can only be read and can't be re-write, which improves the system's security;
3. Accurate: any little change to the electronic document can be detected by the verification module, and can ensure the accurate information transmission by checking the digital certificate to confirm the signer's identity;
4. Convenient: the plug-in package which is used to add the seal appears in the toolbar of OFFICE software and the browser directly, it is convenient for users;
5. Popular: common printers and scanners can be used to ASS, it meets the demand of different operating environments;

5.3 Applications

ASS is a secure seal management system, which is designed to satisfy e-government, e-commerce requirement.

ASS can be applied in many fields, for example:

- It can improve enterprise's profit: enterprises transfer electronic contracts, agreements through the network, that can shorten trading time, reduce the transaction cost greatly, improve working efficiency; reduce personnel errands to economize the travel charge;
- It can reduce the administrative cost: electronic official documents replace paper files. To some fields, in which an agreement needs to be signed by several persons scattering in different places, digital seal has more advantages over traditional paper. Because of the high security of the digital seal, the partners will not need to be afraid that the documents and the tables' contents are modified. It also can save much time and cost losing on transmitting the documents;

– It is compatible between electronic and paper media. This characteristic reduces the initial difficulty for people to use digital signature. With ASS, digital signature is manifested in a visual, elaborate form called digital seal, in which digital signature can be printed in paper media and can be rightly extracted as well.

The digital seal service platform [16], [17], based on AssureSeal System, integrates seal usage, verification and management into a secure platform. This platform will undertake important role in e-commerce and e- government, as well as in circulation of official documents, office automation, electronic contracts, electronic documents and electronic notes, etc.

6 Conclusion

In this paper, we introduce a novel system -- AssureSeal System. An introduction is made to the system from the framework, function modules, key technologies to specifications, characteristics and applications, etc. ASS can protect not only the electronic documents, but also their paper copies, that's unique in China. The implementation of "Electronic Signature of the People's Republic of China"[18] on April 1, 2005, indicates that the country confirms the validity of " electronic signature " in law, which means electronic signature has equal force adeffect to traditional signature. ASS will play important role in electronic commerce and electronic government.

References

1. Cox I J, Miller M. L.: The First 50 Years of Electronic Watermarking. EURASIP J. of Applied Signal Processing (2002) 126-132
2. HSU C T,WU J L.: Hidden Signatures in Image[A].Proc of ICIP'96[C].1996, 3: 223-226
3. Mohan Atreya: Digital Signatures [M]. McGraw-Hill Pub.2002
4. J.Fridrich: Image Watermarking for Tamper Detection. Proc.ICIP'98, Chicago, Oct 1998
5. Lu,C.S:LiaoH-Y.M: Multipurpose Watermarking for Image Authentication and Protetion. IEEE Transactions on Image Processing [J],2001,10(10): 1579-1592
6. Ching-Yung Lin: Watermarking and Digital Signature Techniques for Multimedia Authentication and Copyright Protection. Columbia University 2000
7. Pitas I.: A Method for Signature Casting on Digital Images. IEEE International Conference on Image Processing, 1996, 3: 215-218
8. Fabin A.P. Petitcolas,Ross J.Anderson: Attacks on Copyright Marking Systems. Second Workshop on Information Hidding, Portland, Oregon, USA, April,1998: 218-238
9. Stefan Katzenbeisser□Fabien A.P.Petitcolas: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Boston, London 1999
10. Kunder D,Hatziakos D.: Digital Watermarking for Telltale Tamper-proofing and Authentication. Proceedings of the IEEE, 1999, 87(7):1167-1180
11. "A DIGITAL ANTI-FORGING METHOD", US Patent 10/496, 926, 30/11/2001
12. Rughooputh, H.C.S., Bangaleea, R.: Effect of Channel Coding on the Performance of Spatial Watermarking for Copyright Protection. IEEE AFRICON. 6th,Volume 1, 2-4 Oct. 2002 Page(s): 149 - 153 vol.1

13. Hailing Zhu; Clarke, W.A.; Ferreira, H.C.: Watermarking for JPEG Image Using Error Correction Coding. AFRICON, 2004. 7th AFRICON Conference in Africa Volume 1, 15-17 Sept. 2004 Page(s): 191 - 196 Vol.1
14. Bruce Schneier: Applied Cryptography. Second Edition: Protocols, Algorthms, and Source Code in C (cloth)
15. http://www.cnw.com.cn/issues/2000/35/3510.asp
16. http://hz.axu.cn
17. http://www.assuredigit.com/
18. http://news.xinhuanet.com/newscenter/2004-08/28/content_1908927.htm

ThresPassport – A Distributed Single Sign-On Service*

Tierui Chen[1], Bin B. Zhu[2], Shipeng Li[2], and Xueqi Cheng[1]

[1] Inst. of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
chentierui@software.ict.ac.cn, cqx@ict.ac.cn
[2] Microsoft Research Asia, Beijing 100080, China
{binzhu, spli}@microsoft.com

Abstract. In this paper, we present ThresPassport (Threshold scheme-based Passport), a web-based, distributed Single Sign-On (SSO) system which utilizes a threshold-based secret sharing scheme to split a service provider's authentication key into partial shares distributed to authentication servers. Each authentication server generates a partial authentication token upon request by a legitimate user after proper authentication. Those partial authentication tokens are combined to compute an authentication token to sign the user on to a service provider. ThresPassport depends on neither Public Key Infrastructure (PKI) nor existence of a trustworthy authority. The sign-on process is as transparent to users as Microsoft's .NET Passport. ThresPassport offers many significant advantages over .NET Passport and other SSOs on security, portability, intrusion and fault tolerance, scalability, reliability, and availability.

1 Introduction

As computer networks and systems proliferate to support more online accesses and business, a user is typically required to maintain a set of authentication credentials such as username and password for each service provider he or she is entitled to access. A user is facing a dilemma between using different authentication credentials for each individual service provider for the sake of security, resulting in escalating difficulty in memorizing all those credentials, and using the same credentials for many service providers for easy memorization at the cost of lowered security. Forcing a user to enter authentication credentials frequently when the user accesses different service providers or the same service provider multiple times is also an awkward user experience. It is desirable to have an authentication service to manage a user's sign-on credentials and allow the user to authenticate him or her conveniently to a variety of service providers.

Single Sign-On (SSO) has been proposed as a potential solution to the implications of security, credentials management, and usability for the aforementioned applications. SSO utilizes a centralized credentials management to provide authentication services for users to access participating service providers. With SSO, a user needs to authenticate him or her to an authentication service only once, which in turn enables him or her to automatically log into participating service providers he or she has access permission when needed without any further user interactions. Such a system

* This work was done when Tierui Chen was an intern at Microsoft Research Asia.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 771 – 780, 2005.

makes the complexity to log into an increasing number of service providers completely transparent to a user. From a user's point of view, there is no difference between logging into one service provider and into multiple service providers. The complexity is handled by the SSO system behind scene. In other words, SSO enhances usability in logging into multiple service providers dramatically with a centralized authentication service.

Several different SSO systems have been proposed. Kerberos [1] is an SSO system which is widely used when users, authentication servers, and service providers are under a centralized control such as in the same company. In Kerberos, a user authenticates to an authentication server and obtains a valid Ticket Granting Ticket (TGT) which is used to authenticate the user to a Ticket Granting Server (TGS) when requesting a Service Granting Ticket (SGT). To access a service, a user requests an SGT from a TGS and presents it to the service provider which checks validity of the ticket and makes a decision if access is granted or not. Kerberos is not suitable for use in an untrusted environment such as the Internet [2].

The Liberty Alliance [3], a consortium of over 150 member companies, recently developed a set of open specifications for web-based SSO. Security Assertions Markup Language (SAML) [4], a standard, XML-based framework for creating and exchanging security information between online partners, is used in the specifications. The most popular and widely deployed web-based SSO should be Microsoft's .NET Passport [5] which has provided services since 1999. The core of Passport's architecture is a centralized database which contains all the registered users and associated data and credentials. Every account is uniquely identified by a 64-bit number called the Passport User ID (PUID). Each participating service provider is also assigned a unique ID, and needs to implement a special component in its web server software and to share with the Passport server a secret key which is delivered out of band. To log into a participating service provider, a user's browser is redirected to the Passport server which tries to retrieve and verify validity of a Ticket Granting Cookie (TGC) from the web browser's cookie cache. If such a cookie is not found, then the user needs to enter account name and password to authenticate to Passport, which saves a fresh TGC in the browser's cookie cache. A TGC is encrypted by a master key known only to Passport. If everything goes all right, Passport saves in the browser's cookie cache a set of cookies encrypted with the secret key shared between Passport and the specific participating service provider. The set of cookies acts like Kerberos' SGT and is used to authenticate the user to the participating service provider. More details of different SSO architectures can be found in [2].

There are a few major concerns on security and availability of .NET Passport that prevent users and service providers from widely adopting .NET Passport as a web-based login service, esp. for accessing web services such as a bank account which require higher security and contain sensitive private data. These issues are analyzed and discussed in detail in [6], [7]. In .NET Passport, a user's authentication information is centrally managed by the Passport server. Every user has to be identified and authenticated with the help of the data stored in the central database. Every participating service provider depends on the response of the Passport server and its security. .NET Passport is not scalable. The Passport server is a single point of failure and a central point of attacks for the system. It is an attractive target for hackers to paralyze the whole system through distributed denial-of-service attacks. A single compromise of the Passport server may endanger the whole system. Passport cookies are the only authentication proofs in .NET Passport. Unless a user chooses the automatic sign-in mode which uses persistent cookies, a cookie's lifetime in .NET Passport is determined only by the browser's lifetime and the encrypted cookie's expiration time. A user who forgets to log off the Passport account on a public computer could leave valid authentication tokens for anyone to recover and reuse, which is particularly dangerous for persistent cookies that are strongly discouraged to use.

Threshold-based secret sharing [8], [9] has been extensively studied in cryptography. A (*k, m*) threshold scheme splits a secret into *m* shares and distributes each share to an entity. Any *k* shares can be used to fully recover the secret while any number of shares less than *k* will not be able to recover the secret. Threshold-based secret sharing has recently been proposed to use in CorSSO, a distributed SSO service by Josephson et al. [10]. CorSSO is used to authenticate users, programs, and services, which are referred to as principals. In CorSSO, each party has a pair of public and private keys. A set i of authentication servers create a pair of public and private keys $\{K_i, k_i\}$ and uses a threshold scheme with a threshold t to split the private key k_i and stores a distinct share at each authentication server of the set. The public key K_i is sent to and stored by an application server A which uses the set of authentication servers for authentication service. The private key k_i speaks for the set of the authentication servers. A principal C also has a pair of public and private keys $\{K_c, k_c\}$ where the private key k_c speaks for the principal. When a principal C wants to access an application server A , the principal C uses its private key k_c to encrypt a fresh challenge from the application server A , and requests authentication servers to certify its public key K_c . Each authentication server, after proper identity checking, generates for the principal C a partial certificate which is an encrypted version of the content including the principal C , its public key K_c , valid time of the certificate, etc. with its partial share of k_i . The principal C combines the t partial certificates received from t authentication servers to compute a certificate signed with the authentication private key k_i , which is then sent together with the challenge encrypted with the principal's private key k_c to the application server A . The application server A uses the authentication servers' public key K_i to verify the received certificate, and then extracts the principal's public key K_c to decrypt the encrypted challenge and compare with the original challenge it sends to C to decide if the principal is allowed to access the application server. It is clear that the threshold scheme and authentication servers are used to replace the conventional Certificate Authority (CA) to certify the public key for each principal in CorSSO. The requirement of a pair of public and private keys for each principal renders CorSSO inappropriate for web-based single sign-on authentication service for users, i.e. the application arena targeted by .NET Passport and the Liberty Alliance, since CorSSO does not provide any portability in its authentication service. A user cannot easily use different computers to access a web service the user has permission to access since it is very inconvenient and insecure to carry his or her private key around.

In this paper, we present a distributed, user-friendly SSO system based on threshold-based secret sharing. Our SSO system is called *ThresPassport* – a threshold scheme-based Passport. In ThresPassport, a participating service provider S selects a secret key K_s and utilizes a threshold scheme to split K_s into partial shares, each partial share is sent to an authentication server out of band during registration of the service provider. ThresPassport's client module utilizes a user's account name and password to generate a distinct login credential for the user to authenticate to each authentication server. An authentication server uses its partial share of the secret key K_s to encrypt a challenge from the service provider S passed to it from a user's client module. The client module combines t encrypted challengers from t authentication servers, computes a challenge encrypted by the service provider's secret key K_s, and passes the result to the service provider, which decrypts the received encrypted challenge and compares with the original challenge to decide if the user is granted access permission. ThresPassport shows many significant advantages over .NET Passport and CorSSO, which are discussed in detail later in this paper.

The paper is organized as follows. In Section 2 we describe in detail the architecture and protocols of our distributed SSO system, ThresPassport. Security and comparison with .NET Passport and CorSSO are then presented in Section 3. The paper concludes in Section 4.

2 ThresPassport

A ThresPassport SSO system consists of three parties: users who want to access service providers, service providers who provide services to users, and authentication servers which offer single sign-on services for participating users to access participating service providers. In ThresPassport, a server module is installed in the participating service provider's server, and a downloadable web browser's plug-in is installed to a user's client machine. Before going to ThresPassport details, the notation used in this paper is introduced first.

2.1 Notation

S	A participating service provider.
U	A participating user.
A_i	The *i*-th authentication server.
UID	A unique ID for a participating user U .
SID	A unique ID for a participating service provider S .
AID_i	An unique ID for the *i*-th authentication server A_i.
K_S	A secret key generated by and known only to S .
K_S^i	The *i*-th partial share of K_s generated by a threshold scheme.
K_U^i	A secret key for U to authenticate to the *i*-th authentication server A_i.
p_1, p_2	Two properly selected prime integers, $p_2 > p_1$.

g	A generator in $Z_{p_1}^{*}$, $2 \le g \le p_1 - 2$.
SK_{U,A_i}	A session key between a user U and the i-th authentication server A_i.
$< m >_k$	A message m encrypted by a symmetric cipher with a key k .
$< m >^{k,p}$	It means $m^k \bmod p$ where $m \in Z_p$.
n_X	Nonce generated by entity X .
r_X	A random number generated by entity X .
$[\, x \,]$	x is optional in describing a protocol.

2.2 ThresPassport Protocols

ThresPassport is divided into two phases: the setup phase and the authentication phase. In the setup phase, participating service providers and users register to authentication servers, and generate and send secret keys securely to authentication servers out of band. Those keys will be used in the authentication phase to authenticate a user to authentication servers and to a service provider. In the following, we assume that there are *n* authentication servers in total and a *(t, n)* threshold scheme is used to share a service provider's secret key K_s .

2.2.1 Setup Protocols for Participating Service Providers and Users

During the setup phase, both participating service providers and users are required to register with the authentication servers and install a server module on service providers' servers and a client web browser plug-in on users' machines. A participating service provider *S* utilizes the following protocol to register securely to authentication servers.

1. S : Generates a secret key K_S , $1 \le K_S \le p_2 - 2$, and calculate K_S^{-1} such that $K_S^{-1} K_S = K_S K_S^{-1} = 1 \bmod (p_2 - 1)$.
2. S : Uses a *(t, n)* threshold scheme to split K_S into n shares $K_S^i, 1 \le i \le n$.
3. $S \rightarrow A_i, 1 \le i \le n$: SID , K_S^i .
4. $A_i, 1 \le i \le n \rightarrow S$: Success. A_i stores SID and K_S^i for later usage.

A user *U* also needs to register with the authentication servers before he or she can enjoy the authentication service provided by ThresPassport. The following protocol is used to register a user U to the authentication servers. The registration process must be secure.

1. U : Generates a unique user name and a good password. The client program generates a unique *UID* from the user name.
2. U : Computes $K_U^i = hash\,(UserName, Password, A_i), 1 \le i \le n$.

3. $U \rightarrow A_i, 1 \le i \le n$: UID , K_U^i .

4. $A_i, 1 \le i \le n \rightarrow U$: Success. A_i stores UID and K_U^i for later usage.

2.2.2 User Authentication Protocol to an Authentication Server

If a user U has not authenticated to an authentication server A_i yet during a single sign-on process of ThresPassport, the user is required to authenticate to A_i before A_i can help authenticate the user to a service provider S . A challenge-response protocol such as the following one using the shared key K_U^i derived from the user's password can serve the purpose and generate a session key for subsequent confidential communications between the user and the authentication server.

1. $U \rightarrow A_i$: Authentication request.
2. $A_i \rightarrow U$: n_{A_i} .
3. $U \rightarrow A_i$: UID , $< r_U, n_U, n_{A_i} >_{K_U^i}$.
4. $A_i \rightarrow U$: $< r_{A_i}, n_{A_i}, n_U >_{K_U^i}$ or failure.

In Step 3, U generates the authentication key K_U^i from U 's password with the equation $K_U^i = hash\,(UserName, Password, A_i)$. In Step 4, A_i uses the received UID to extract the corresponding key K_U^i to decrypt the received message and encrypt the message to be sent. The decrypted nonce n_{A_i} is compared against that sent in Step 2 to decide what to send in Step 4. If the protocol ends successfully, a session SK_{U,A_i} is generated at both ends by hashing the communicated random numbers r_U and r_{A_i} : $SK_{U,A_i} = hash(r_U, r_{A_i})$. This session key is used for subsequent confidential communications between U and A_i for the session. Once the session ends, SK_{U,A_i} is destroyed and a user has to authenticate to A_i again through the above protocol. A session can be terminated by a user or when the lifetime set by the security policy expires.

2.2.3 Single Sign-On Protocol

The following protocol is used for a user's client module to acquire an authentication token from authentication servers and to gain access to a service provider.

1. $U \rightarrow S$: Request access to a service.
2. $S \rightarrow U$: SID , n_S , $[< g >^{r_S, p_1}]$, [a list of t authentication servers $\{A_{d_f}, 1 \le f \le t\}$] .

3. For $1 \leq f \leq t$

3.1: $U \to A_{d_f}$: $SID,\ n_S,\ [<g>^{r_U, p_1}]$, $[\ UID\ ]$

3.2: $A_{d_f} \to U$: $<UID, U, n_S, [<g>^{r_U, p_1}]>^{K_S^{d_f}, p_2}$

4. $U \to S$: UID , $<UID, U, n_S, [<g>^{r_U, p_1}]>^{K_S, p_2}$, $[<N_S>_k]$,
where $k = <g>^{r_S \cdot r_U, p_1}$.

5. $S \to U$: access is granted or denied.

In Step 2, the service provider picks up t live authentication servers from all available authentication servers based on workloads, bandwidths, processing power, reliability, etc. and sends to the user's module. This means that a service provider may need to monitor status of authentication servers. An alternative solution is that the client's module tries to find t live authentication servers from the list of n authentication servers received from the service provider. If the list of authentication servers is already known to clients, there is no need to send the list to a client.

In Step 3, if the user has not authenticated to the t authentication servers yet or the preceding sessions have expired, the user authentication protocol described in Section 2.2.2 is used to authenticate the user to each authentication server A_{d_f} and set up a secure communication channel between U and A_{d_f} with a session key $SK_{U, A_{d_f}}$ before going to Step 3.1. Note that the communications between the user and an authentication server in Steps 3.1 and 3.2 are confidential by using the session key $SK_{U, A_{d_f}}$ obtained when the user is authenticated to the server, although the message sent in Step 3.2 is not necessary to be confidential since it is already encrypted. The client in Step 4 computes an authentication token $<UID, U, n_S, [<g>^{r_U, p_1}]>^{K_S, p_2}$ from the received t partial authentication token $<UID, U, n_S, [<g>^{r_U, p_1}]>^{K_S^{d_f}, p_2}$. In Step 5, the service provider uses the secret key K_S^{-1} known only to itself to decrypt the received token: $((UID, U, n_S, [<g>^{r_U, p_1}])^{K_S})^{K_S^{-1}} = (UID, U, n_S, [<g>^{r_U, p_1}]) \bmod p_2$, and makes a decision if access is granted or denied. If secure communication is desired after U is signed to S , the optional items related to the generator g are also communicated in the protocol. The session key for subsequent confidential communications between U and S is set to be $<g>^{r_S \cdot r_U, p_1}$, which is k in Step 4. This session key is in fact generated with the Diffie-Hellman key agreement [11].

Both the authentication token $<UID, U, n_S, [<g>^{r_U, p_1}]>^{K_S, p_2}$ and the partial authentication token $<UID, U, n_S, [<g>^{r_U, p_1}]>^{K_S^{d_f}, p_2}$ contain U which is an unique network ID of the user U 's client machine such as the network address. Note that nonce and random numbers in different protocols have no relationship even though we use the same notation in describing the protocols.

3 Security and Comparison with Other SSOs

3.1 Security of ThresPassport

In ThresPassport, a service provider's key K_S is generated by and known only to the provider. Authentication servers do not know and cannot deduce this secret key unless t or more authentication servers collude. This secret key never transfers over a network and is under full control by its rightful owner. Such a design guarantees the security of the secret key. On the client side, a user's password is never used directly in authentication. Instead it is used with a one-way function to derive the authentication keys used to authenticate the user to authentication servers. An authentication server A_i cannot use the authentication key K_U^i it knows to recover the password or the user's authentication keys to other authentication servers without a brute force attack. Note that the authentication key K_U^i is never transferred over a network except during the setup stage. That said, a user's password should be complex enough to avoid weak keys since the authentication keys K_U^i are generated from the password, and hence contain no more entropy than the password.

Since passwords are entered at the client side, certain security and tamper resistance are required for the client module. Such a requirement is typical in most security software at the client side. For example, there should be no malicious module between the user and the client module to launch a man-in-the-middle attack to impersonate the user in communicating with the client module. The session keys stored by the client module during the life of the session should not be examined by untrustworthy programs. Our design also minimizes such a risk. In ThresPassport, a user's password is live in memory in a very short time. It is overwritten once the authentication keys $\{K_U^i\}$ are generated. Once the authentication process to authenticate a user to servers is over, the authentication keys $\{K_U^i\}$ are overwritten. Only the temporal, one-time session keys are stored in memory and used in subsequent communications between the client and authentication servers during the life of the session.

3.2 Comparison with Other SSOs

In this subsection, we would like to compare ThresPassport with .NET Passport [5] and CorSSO [10]. To an end user, ThresPassport appears the same and as easy to use as .NET Passport. The complexity to authenticate a user to multiple authentication servers in ThresPassport is completely hidden inside the protocols and software. On the other hand, ThresPassport shows several important advantages over .NET Passport. On the security side, there is no single central point containing all the secret credentials in ThresPassport. All secret credentials are completely controlled by each rightful owner: a service provider's key is controlled by and known only to the provider. A user's password is controlled by and known only to the user (and to the client's module in a very short time). Hackers have to compromise up to t authentication servers to incur security damage to ThresPassport, thanks to the *(t, n)* threshold scheme used in the system. Since .NET Passport requires SSL/TLS channels to communicate between the user and the Passport server, an appropriate Public Key Infrastructure (PKI) must be in place. Like Kerberos, ThresPassport does not depend on any PKI. In ThresPassport, session keys replace authentication cookies in .NET Passport for authentication, and therefore mitigate the risk that a subsequent user recovers the preceding user's authentication cookies in .NET Passport to impersonate the preceding user to illegally access service providers. A user's privacy is also better protected in ThresPassport, thanks to the notorious privacy track record of cookies.

On the reliability side, ThresPassport is no longer a system of a single point of failure like .NET Passport due to its distributed authentication servers. Any t out of the total n authentication servers can provide authentication services to users in the system. It is much more difficult to launch a distributed denial-of-service attack to disable all but $t-1$ or less authentication servers. On the contrary, a successful denial-of-service attack to the Passport server would disrupt authentication services completely in .NET Passport. ThresPassport is also scalable, dealing well with both small and large systems with a large variety of users and service providers.

ThresPassport also shows several significant advantages over CorSSO. ThresPassport enables portability that CorSSO lacks. A user can use any computer (as long as the ThresPassport's client module is downloaded and installed) to sign on and access a service provider in ThresPassport. In CorSSO, a trustworthy authority is assumed, whose role is to generate a pair of public and private keys $\{K_i, k_i\}$ for a set of authentication servers and to use a threshold scheme to split the private key k_i into partial shares distributed to and stored by individual authentication servers. In ThresPassport, each party controls its own secrets, and there is no dependency on the existence of such a trustworthy authority. This advantage is extremely attractive when authentication servers are controlled and administrated by different companies since in this case federation is needed to achieve a virtual trustworthy authority. A third advantage is that appropriate PKI is required in CorSSO, recall that each of the three parties in CorSSO, a principal, a service provider, or a set of authentication servers, has a pair of public and private keys speaking for itself. As we have just mentioned above, ThresPassport does not depend on any PKI which dramatically increases its chance to be widely adopted and employed.

4 Conclusion

In this paper, we have presented ThresPassport, a web-based, distributed single sign-on system using passwords, threshold-based secret sharing, and encryption-based authentication tokens. In ThresPassport, critical secrets such as a service provider's sign-on key and a user's password are always controlled by and known only to the original owner. Every authentication server owns partial authentication information of a client or a service provider. A threshold number of authentication servers are required to accomplish an authentication service. ThresPassport depends on neither PKI nor existence of a trustworthy authority. It is as transparent and easy to use as .NET Passport. ThresPassport offers many significant advantages over .NET Passport and other proposed SSOs on security, portability, intrusion and fault tolerance, scalability, reliability, and availability.

References

1. Internet Engineering Task Force, RFC 1510: The Kerberos Network Authentication Service. (V5), Sept. 1993
2. A. Pashalidis and C. J. Mitchell: A Taxonomy of Single Sign-On Systems. 8th Australasian Conf. Info. Security and Privacy (ACISP) 2003, Wollongong, Australia, July 9-11, 2003, Safavi-Naini and J. Seberry (eds.), LNCS vol. 2727, pp. 249–264, Springer-Verlag, July 2003
3. http://www.projectliberty.org
4. http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security
5. http://www.passport.com
6. D. P. Kormann and A. D. Rubin: Risks of the Passport Single Signon Protocol. IEEE Computer Networks, vol. 33, pp. 51--58, 2000
7. Rolf Oppliger: Microsoft.NET Passport: A Security Analysis. IEEE Computer Magazine, vol. 36, no. 7, pp. 29-35, July 2003
8. A. Shamir: How to Share a Secret. Communications of ACM, vol. 24, no. 11, pp. 612 – 613, 1979
9. V. Shoup: Practical Threshold Signatures. Proc. EUROCRPT'00, Springer Verlag, LNCS vol. 1807, pp. 207 – 220, 2000
10. W. K. Josephson, E. G. Sirer, and F. B. Schneider: Peer-to-Peer Authentication with a Distributed Single Sign-On Service. 3rd Int. Workshop on Peer-to-Peer Systems (IPTPS'04), San Diego, USA, February 2004
11. A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone: Handbook of Applied Cryptography. CRC Press, London, New York, 1997

Functional Architecture of Mobile Gateway and Home Server for Virtual Home Services

Hyuncheol Kim[1] and Seongjin Ahn[2,*]

[1] Dept. of Electrical and Computer Engineering, Sungkyunkwan University,
300 Chunchun-Dong, Jangan-Gu,Suwon, Korea, 440-746
hckim@songgang.skku.ac.kr

[2] Dept. of Computer Education, Sungkyunkwan University,
53 Myungryun-Dong, Jongro-Gu, Seoul, Korea, 110-745
sjahn@comedu.skku.ac.kr

Abstract. With the progress of portable appliances such as cell phone and handheld PC, it can be foreseen the popularization of Personal Area Network (PAN) and the diversification of services that are based on Personal Mobile Gateway (PMG). Although inter-operability among home appliances reached a service stage, researches about virtual home services and middlewares for a new small scale network such as PAN are at an early stage. Network service/connection methods, terminal control schemes, and middlewares which are used in the traditional home networks must be enhanced to accommodate PMG-based PAN. In this paper, we propose an integrated virtual home network platform that guarantees seamless connections between home network and PAN. We also analyze the proposed indispensable functions and presents functions that should be added to the existent home gateway or the home server.

1 Introduction

The tremendous growth in the Internet has created a new paradigm in politics, economics, society, and culture. Now, Internet is the most essential element of national competition and extensive research on high-speed Internet has been achieved. The high-speed Internet has start from the backbone network. With the development of wired/wireless communication techniques as shown in Fig. 1, high-speed Internet has made through the access network and is currently spreading rapidly to the home network. Home network is a small-sized network that inter-connects various household appliances and systems together in order to transmit data among the household machineries. Home network can be connected to the external network devices in order to provide remote access and control. Although home network technology began for connecting home PC with printer, scanner, and other peripheral devices, with the development of network technology and the change in residential environment, entertainment services for multimedia communication has been consisting the main current [1][2].

[*] Corresponding Author.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 781–790, 2005.

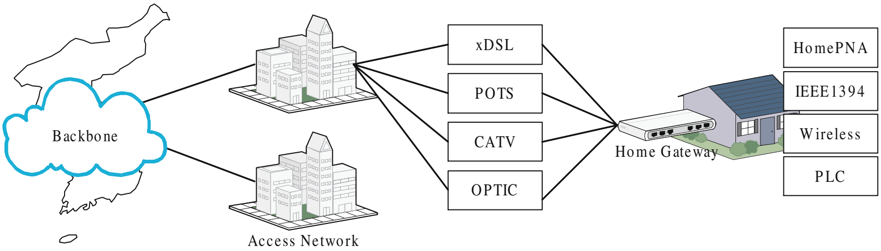


Fig. 1. Backbone/Access/Home Network

On the other side, cellular phone which is the most typical mobile communication does not only provide simple voice communication service but also can offer many additional functions such as digital camera, MP3/Video player, and high capacity storage. In illustrating these current trends, as shown in Fig. 2, PAN (Personal Area Network) services which use terminal such as cellular phone by PMG (Personal Mobile Gateway) reached the commercial stage. PMG contains all devices which use a personal mobile terminal such as cellular phone as a gateway. PMG can exchange contents in a wireless mode with digital equipment such as digital camera, PDA (Personal Digital Assistant), voice recorder, and MP3 player that are in close distance.

Therefore, a new inter-networking model and new protocol is needed between PMG and home network platform in order to enhance the home network coverage to the public network. Through the new model, the devices in the PAN environment and the devices in the home network can share and access information such as contents, data, and event, and provide service interworking.

This paper proposes a new platform structure for virtual home network that converge home network and public network by connecting home network with PAN. Device interworking inside the home network is currently being standardize in the DLNA (Digital Living Network Alliance) and is currently going into the service commercial phase. But, the new form of mobile gateway based PAN, a small-sized network, and its service interworking and middleware does not have much progression. The existing middleware can not fundamentally provide function of terminal exchange with the circumstances of the network and of prestorage of data that is needed by the user [3][4][5]. In order to provide virtual home service between PAN and home network, this paper proposes techniques to provide non-stop new services independent of the user's location and network circumstances through virtual home connection management technique, virtual home distributed information sharing technique, virtual service control technique, virtual home automatic and distributed configuration technique.

The content of this paper is as follows. First, chapter 2 examines elemental technique in providing home network service based on home gateway and servers. Chapter 3 specifically illustrates the proposed new virtual home network platform, and application of the proposed technique and its far-reaching effects and the new research area will be concluded in chapter 4.

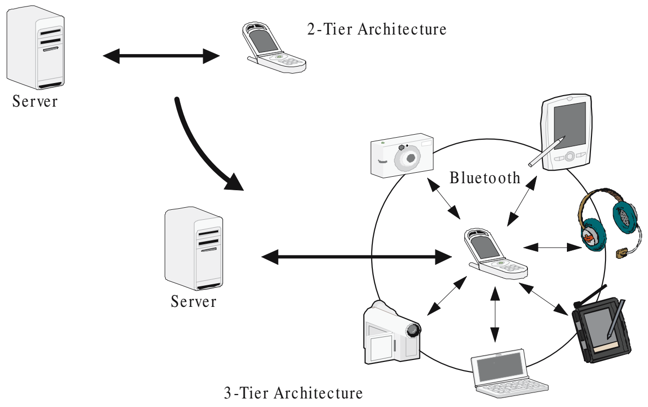


Fig. 2. 3-Tier PMG Architecture

2 Home Network Components

In the beginning, home network technologies started with PC being the center sharing files and peripheral devices such as printer and using the data oriented Internet services. But, with the change in the living environment such as cyber-apartment and home automation, home automation services have spread actively through the wired and wireless media. Very high-speed Internet service will be developed based on high quality entertainment service using the audio and video devices hereafter. Home network infrastructure keep changing in the direction of wireless technologies such as wireless LAN, UWB (Ultra Wideband) from the conventional Ethernet, telephone line, and power line. Many different home appliances has a censor that can cense and control itself and is evolving into the censor-based ubiquitous home network.

Home network consists of broadband network access technologies of user network and home network, and the interconnection of these networks through the home gateway system. Along with the software systems, middleware technology is one of the most important elements of the home network.

First, broadband network access technology connects backbone network with the home network and is called a subscriber network or access network. Home gateway is a equipment in a form of household devices and interconnects and mediates home network wired-wireless technology to the various access network (WAN) technology such as xDSL (x Digital Subscriber Line), cable, fiber, and satellite. Home gateway uses middleware technology to perform network and self-service to the home network users [6].

2.1 Home Gateway

Home gateway or residential gateway is used to interconnect and mediate home network and outer network with two major purposes. One is to connect and

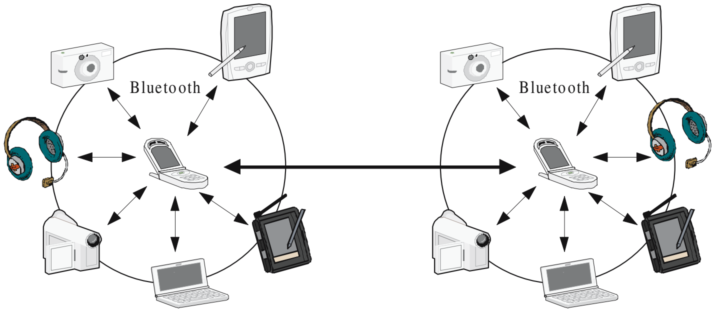


Fig. 3. PAN-to-PAN Service

manage PCs and household appliances in the home acting as a hub, and the other is to act as a gateway to interconnect the home and the outside. Acting as a management hub, the home gateway connects home devices through wired and wireless connection and control these home devices and provide multimedia services.

Early home gateway was developed for the purpose of remote meter reading, but with the advent of various home network services, the home gateway started to accommodate many different functions. Home gateway has developed to provide Internet sharing, NAT (Network Address Translation), security service such as firewall, protocol convergence function, broadband access gateway, voice traffic transmission, home PBX (Private Branch Exchange), and streaming video transmission function. Further than that, home gateway has evolved to converged gateway that combines voice and data.

2.2 Home Server

Home server is a system that store, manage, and distribute multimedia data and control, manage, and provide interworking function of various appliances that are connected to the home network in order to provide integrated digital home services. Home server is a logical system, thus it can be a stand-alone device or coexists with other device. Home server is in charge of remote service management, digital broadcasting server for home audio, video, and game, energy management, home automation, security.

The techniques that configure home server are hardware platform, software such as middleware and real-time OS, and application technique. The services that can be provided through the home server are not fixed, and actual structure and function of the home server may be various according the country and company. Home server hardware platform consists of wired-wireless devices to communicate with the home terminals, codec and storage device for broadcasting and streaming. Software inside the home server platform contains control middleware such as real-time OS for home server, JVM (Java Virtual Machine), and open service framework and multimedia middleware. Early home server

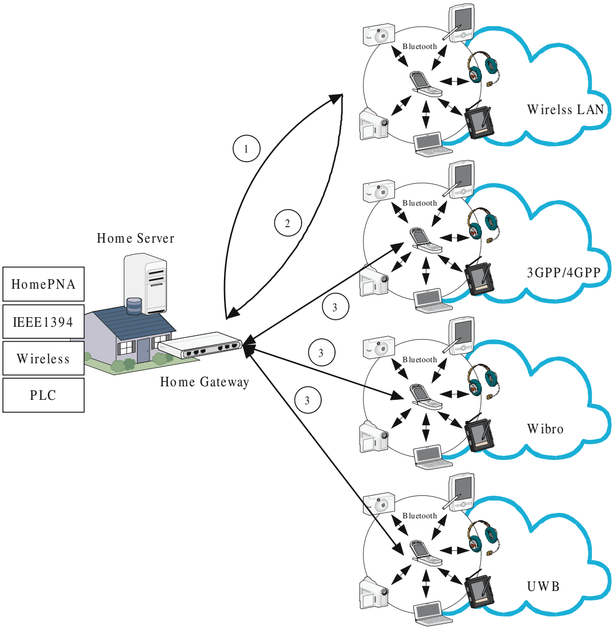


Fig. 4. Virtual Home Service

were based on control and management function. Nowadays, it has developed to include home gateway function.

2.3 Home Network Middleware

Middleware can be defined as software that provides easy and secure computer to computer communication and provides direct and indirect management. That is, it is a software the support communication between different protocol and system OS and communication between database and application. It provides support to the application to operate in any environment. In order to provide interoperability between various systems, middleware of the home network is one of the very important element and major home networking middleware includes OSGi (Open Service Gateway initiative), UPnP (Universal Plug and Play), Jini, VESA (Video Electronics Standards Association), HAVi (Home Audio Video Interoperability), OpenCable, and DVB (Digital Video Broadcasting) [7][8] [9][10] [11].

3 Virtual Home Service Architecture

Home network is a user oriented customized infrastructure and is based on many core technologies such as home services, home network equipments, intelligent terminals and appliances, ubiquitous network. That is, home network is connected to the information electronic appliances and is configured with infrastructure that provides service environment independent of time and place. The concept of home network is extended to provide home service that does not put limitation on the service area to only internal to home but also to outside to the home through the introduction of new form of PMG and service integration as shown in Fig. 2, Fig. 3. The virtual home platform or virtual home service supports seamless connection between personal mobile terminal and home electronics inside the PAN based on mobile gateway and provide mobile platform and service that is irrelevant of place as shown in Fig. 4.

3.1 PMG-Based PAN

Mobile PAN technology which is a personal mobile networking solution for short distance of around 10m in the normal home service is currently being becoming noteworthy. It would be quite a burden in connecting all the devices inside the home when using the conventional wired home networking techniques such as Ethernet, PLC (Power Line Communication), and HomePNA (Home Phoneline Networking Alliance). Therefore, the demand is increasing for cheaper short-distance mobile networking technique that can connect devices without using cable.

The PAN which is limited only inside the home can be extended using the PMG which can also configure service provision environment such as home network to outside the home and extends the limited home network services to the public network. With the active progress of telematics, service interworking technique that connects terminal in vehicles center on mobile gateway and home network is one of an important item that should be considered. Corporation such as IXI Mobile Company in US develop is own cell phone and various digital devices with its own developed PMG software and will initiate mobile telco based PAN service. Also, PMG interworking services is another new type of service on the rise as shown in Fig. 3. The typical example of service scenario that is provided by virtual Home service is as follows.

- In case of an emergency, all the voice/video data in the home is transmitted to all the terminal in the PAN or store all the data in the PAN terminal to the home server.
- Control all the voice/video data inside and outside the home and transmit only voice or video as according to the situation. Also, add/delete new media channel such as still image.
- Urgent message or abnormal situation data can be transmitted to pertaining terminal by understanding user and manager presence information. If the transmitted media does not answer, it automatically transmits data to other terminal.
- Data can be transmitted to the selected terminal as selected time with the selective wake-up function in the PAN.

- Add/delete function of desired sleek terminal (for example, exchanging 2-inch monitor sleek terminal to 3-inch monitor terminal or exchanging stereo headset to 3D stereo headset. Also, in the case when network in changed to 3GPP to 4GPP during mobile, provide customized service according to network bandwidth without the exchange of the sleek terminal).

Conventional middlewares support only automatic software processing for peripheral device changes, but the method proposed by this paper provides selection and negotiation functions of adequate peripheral device according to the environment for PMG based PAN. That is, virtual home network that uses PMG based PAN requires capability to negotiate various parameter with the home gateway and server in real-time in order to reflect the current PAN and network status and needs to perform in the opposite according to the users selection as shown in Fig. 4.

3.2 Virtual Home Network Platform

Fig. 5 illustrates the proposed extended home network platform to provide efficient virtual home service. The proposed extended platform is to extend the middleware for the home gateway/server to add proxy function and dynamic control of various terminal and services according to the network/user in order to provide PMG based PAN. First, terminal recognition of the conventional middleware recognizes new devices during booting or in the form of artificial subscribing method. Since the off-premises devices such as PAN is configured with more than one terminal group, it can add/delete devices freely even during service. Therefore, in order to provide contiguous services, it needs basically different recognition technique and service insertion/withdrawal technique. This means that it requires extensive improvement in middleware and a new form of Open API [12].

Fig. 5 ① shows extended metadata auto creation, integration, and retrieval part to share information between PAN terminal and home network devices based on PMG. Conventional metadata was designed to efficiently use one media, but with establishment of various standardizations in providing service, its use has diminished to very limited scope. Also, in case where PAN consists of many terminals providing same service but with different quality (for example, stereo headset and mono headset), it would be difficult to efficiently provide services using the conventional metadata creation and retrieval method. Therefore, in reflecting these situations, the proposed a solution to create metadata that is bundled from many conventional metadata for negotiation and a method to select terminals that transmit intelligently retrieved contents.

One the other hand, conventional middleware has limited service scope, thus it was difficult to process in real-time of the home network devices moving to off-premises (Fig. 4 ①) or the external device moving into the home(Fig. 4 ②). Therefore, SIP (Session Initiation Protocol) based call control was needed to construct virtual home, thus present middleware needs to extend its function to call control interworking.

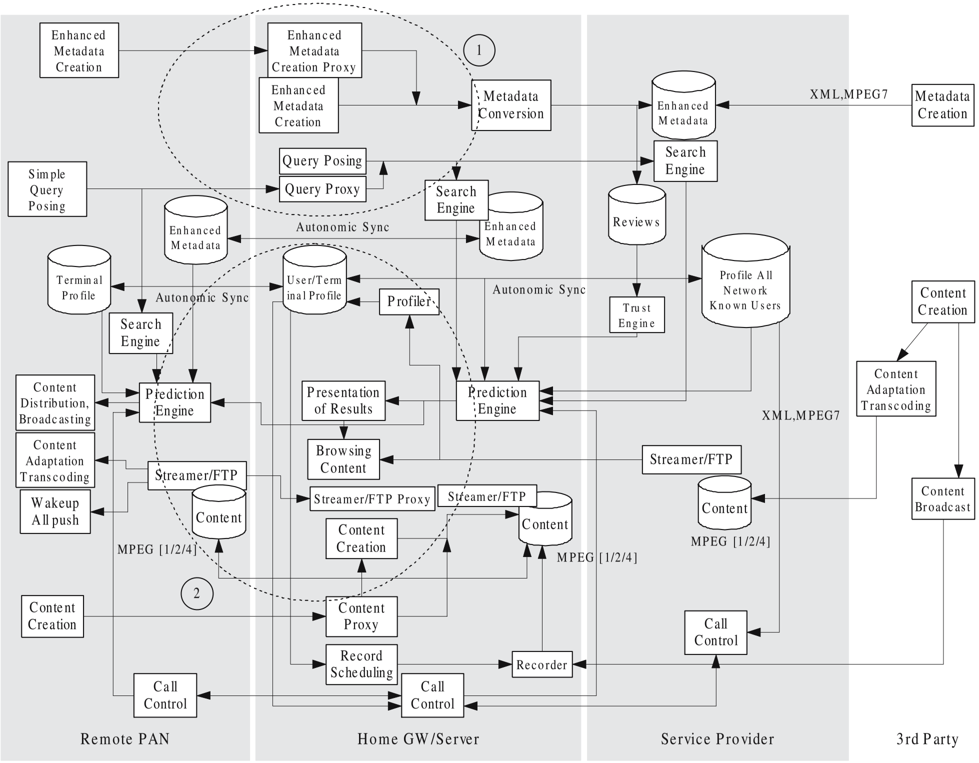


Fig. 5. Extended Home Network Platform, Meta Data, and Information Synchronization for Virtual Home Service

As shown in Fig. 5, this paper proposes SIP protocol to be used to perform these functions. Especially, external network has very different network environment from the fixed home network. The external network provides different service quality accordingly to the user's location as shown in Fig. 4 ②, ③. Thus, it has extended middleware to reflect the network status of servicing the devices in real-time. Also, the proposal extended the protocol and interworking part to the home server based on SIP to recognize the network status change in real-time and select adequate PAN terminal for non-stop service in order to efficiently use various terminals configured in the PAN accordingly to the quality of service.

Fig. 5 ② shows real-time information synchronization between personal mobile terminal and home network devices. Fall in prices of large capacity storage allows data distribution to different devices from the concentrated server method. Data synchronization and real-time transmission is essential in executing metadata and contents distribution. This paper uses autonomic synchronization function that considers various parameters such as characteristics of the terminal in the PAN and the time of execution and limited bandwidth between home server and off-premises devices for data synchronization. Also, efficient streaming between home network and PAN is different from the conventional TCP/UDP based streaming because it needs to control traffic accordingly to the network and user situation. This paper proposed method in using SIP to use negotiation result to the TCP/UDP parameter.

Also, diversity in contents and enforcement of intellectual property is not solved using a simple access control and authentication technique, home network should also consider contents protection technique such as DRM (Digital Rights Management) and IPMP (Intellectual Property Management and Protection). Thus, this paper proposed using SIP Presence function in the home network environment to provide services irrelevant of the users position.

4 Conclusion

In spite of many advantages which home network services bring, a compatibility problem is becoming the obstacle at the activation of the market. Although a identical medium can be used, there are too many home networking equipments with different specifications competing with one another in the home network. That is guaranteeing the interoperability of home digital machinery and tool is most big and urgent problem. To solve the problem, a worldwide integration of middleware solutions and home platform architecture must be preceded.

This paper proposes platform structure that is needed to use conveniently the home services anywhere, anytime through configuration of virtual home network that converge home network and public network by connecting home network with PAN. The virtual home network architecture proposed in this paper offer an information share service and a service integration service basically between home network and PAN. Through the proposed architecture simplification and normalization of service interconnection possible.

Through the virtual home service defined as a set of PAN in this paper the distribution of the service and information is possible. Strengthening the competitiveness of home mobile devices including cellular phone and an activation of the home network services and equipment development also can be expected.

References

1. Chen, W. Y.: Emerging Home Digital Networking Needs. Community Networking Proceedings, Sep. (1997) 7–12
2. Saito, T, Tomoda, I., Takabatake, Y.: Home Gateway Architecture and its Implementation. IEEE Transactionson Consumer Electronics, Vol. 46, Issue 4, Nov. (2000) 1161–1166
3. van, Wolfswinkel R., van, Smirren D.: The Xhome Functional Architecture. XHome Draft 2.1
4. Home Networked Device Interoperability Guidelines V1.0. DLNA Forum
5. Kolberg, M, Magill, E.H,Wilson, M.: Compatibiltiy Issues Between Services Supporting Networked Appliances. IEEE Communications Magazine, Vol. 41, Issue 11, Nov. (2003) 136–147
6. Corcoran, P. M.: Mapping Home-Network Appliances to TCP/IP Sockets Using a Three-tiered Home Gateway Architecture. IEEE Transactions on Consumer Electronics, Vol. 44, Issue 3, Aug. (1998) 729–736
7. Kyeongdeok Moon, Younghee Lee, Youngsung Son, Chaekyu Kim: Universal Home Network Middleware Guaranteeing Seamless Interoperability Among the Heterogeneous Home Network Middleware. IEEE Transactions on Consumer Electronics, Vol. 49, Issue 3, Aug. (2003) 546–553
8. Dongsung Kim, Jaemin Lee, Wookhyun Kwon: Design and Implementation of Home Network Systems using UPnP Middleware for Networked Appliances. IEEE Transactions on Consumer Electronics, Vol. 48, Issue 4, Nov. (2002) 963–972
9. JooYong Oh, JunHo Park, Gi-Hoon Jung, SoonJu Kang: CORBA based Core Middleware Architecture Supporting Seamless Interoperability between Standard Home Network Middleware. IEEE Transactions on Consumer Electronics, Vol. 49, Issue 3. Aug. (2003) 581–586
10. Dobrev, P., Famolari, D.,Kurzke, C., Miller, B.A.: Device and Service Discovery in Home Networks with OSGi. IEEE Communications Magazine, Vol. 40, Issue 8, Aug. (2002) 86–92
11. Lee Koonseok,Lee Kyungchang ,Lee Suk, Oh Kitae, Baek Seungmyun: Network Configuration Techniques for Home Appliances based on LnCP. IEEE Transactions on Consumer Electronics, Vol. 49, Issue 2. May (2003) 367–374
12. Chiu Ngo: A Service-oriented Wireless Home Network. Consumer Communications and Networking Conference 2004, Jan. (2004) 199–203

Validation of Real-Time Traffic Information Based on Personal Communication Service Network

Young-Jun Moon[1], Sangkeon Lee[2], and Sangwoon Lee[3]

[1] The Korea Transport Institute, Goyang, Korea
[2] Korea Research Institute of Human Settlements, Anyang, Korea
[3] Munhwa Broadcasting Corporation, Seoul, Korea
sklee@krihs.re.kr

Abstract. This research demonstrates an efficient traffic information provision system for mitigating traffic congestion in the street networks based on multi-type traffic detectors which include inductive loop, image recognition, beacon, and personal communication service (PCS). A methodology is demonstrated in this study, which combines and processes data including vehicle location and speed from the wireless communication network, i.e. PCS in order to validate real time traffic information through PCS and internet.

1 Introduction

Cellular phones have become widespread throughout the world. Over 30 million people use cellular phones in Korea. This situation in Korea gives people good opportunities to use many wireless phone-based applications. Since the cellular phone network can provide many methods to identify the location of mobile phone user, many applications called Location Based Services (LBS) will be available in the future with this location information. One of these applications can be used in the field of transportation that has focused on the movement of people and goods. There have been many attempts to collect information on travelers in order to explain transportation phenommena, to forecast future transportation condition, and to plan transportation networks. Therefore, identifying and detecting the routes and the location of travelers is one predominant issue of transportation research. There have been many ways to get information on travelers and traffic conditions. However, most of them need high maintenance or installation costs even though there is the possibility of unreliable reports and no guarantee of accuracy.

Cellular phones should be considered as alternative data collection devices. The advantages of using Personal Communication Services (PCS) data are abundance of data and a procedure that can determine the location of users regardless of inaccuracies in the system.

The purpose of this study is to propose a method of generating transportation information such as speed and travel time by using multi-type detectors including loop detector, PCS, etc.. The most significant issue in this paper is how to correct PCS data location errors with data from other detectors and whether it is too difficult to correct them, and what the possible quality of information services based on the corrected data.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 791 – 800, 2005.

This research demonstrates an efficient traffic information provision system for mitigating traffic congestion in the street networks based on multi-type traffic detectors which include inductive loop, image recognition, beacon, and PCS. A methodology is demonstrated in this study, which combines and processes data including vehicle location and speed from the wireless communication network, i.e. PCS in order to validate real time traffic information through PCS and Internet.

2 Wireless Communication Technologies

2.1 Dedicated Short Range Communications (DSRC)

The short-range communications capabilities of 5.8 - 5.9 GHz DSRC are uniquely targeted toward supporting location-based mobile services. This type has the bandwidth and other performance characteristics which would be the most reliable, effective and efficient way to support the communication of localized information between the roadside and vehicles. In addition, DSRC has the potential to provide low latency communications, which have been identified as a necessary capability to support the transportation applications including traffic information collection and provision, vehicle safety, and traffic management. In order to realize this potential, however, the correct technological choices must be made for the rules in each country for this spectrum, and for the standards to be used in this band.

2.2 Digital Cellular and PCS

The packet data capabilities of always-on from cellular and PCS technologies will virtually eliminate the call set-up delays of data connections over current cellular systems. However, end-to-end latency is likely to remain in the range of at least several seconds, due to the server processing required in the mobile location registers, and the multiple packet forwarding necessary to deliver data to/from dynamically changing cellular sites. Moreover, data communications over the networks tend to be lower priority than voice communications, so data packets can be expected to encounter buffer-based latency if the networks are busy with voice traffic. These latency limitations will likely preclude the use of cellular communications for the vehicle safety and/or traffic management applications but not influence the utilization for the majority of traffic information collection and provisions.

2.3 Wireless LAN

The wireless local area network (LAN) system could provide extensive data downloads to garaged vehicles, as well as allowing the vehicles to download non-time-critical information to wider area networks. These developments offer the opportunity for on-board equipped vehicles to upload and download data through these wireless LANs while the vehicles are within range of the communications. Most of the traffic information collection and provision applications require two-way communications, as is generally the case in many aspects of wireless LAN for the mobile environment. However, the modifications necessary include reducing the data rate for more reliable communications at highway speeds, and reducing/eliminating the LAN-based “handshaking” required in order to reduce the system latency from seconds to milliseconds.

2.4 Differential Global Positioning System (DGPS)

The DGPS system will use differential GPS reference station transmitter systems located throughout the country to broadcast additional information that can be used by GPS receivers to generate more accurate location estimates. These DGPS transmitters broadcast in the 300 KHz frequency range. Using DGPS to augment the GPS system currently provides accuracy in the one-meter range for the better quality receivers in the vehicles that are moving. Once this system is fully deployed, 1-meter positioning accuracy can be expected to be widely available from the combination of GPS / DGPS signal reception and more processing power in receivers. This should allow the geopositioning requirements of many of the transportation applications including traffic managements and safety, travellers information systems, etc.

2.5 Ultrawideband (UWB)

The use of UWB at very low power outputs, and within a limited range of spectrum has been granted in the few countries. Consequently, the commercial development of this technology has been focused on very low-power applications for sensing and communications. The largest limitation of UWB for vehicle communications is the limited range expected with the initial systems that are likely to become available. One of the positive aspects of UWB for vehicles is the all-digital implementation, which potentially allows low cost, light weight and small size to be realized. In addition, UWB appears to be fairly immune to multipath interference, a significant benefit for moving vehicles. The low-cost, small size characteristic of UWB devices, coupled with their potential use as an integrated communications, positioning and radar solution, makes UWB a reasonable candidate to monitor for further developments that may allow its use for vehicle safety and management applications.

3 Estimating Travel Time Information from PCS

3.1 Characteristics of PCS Data

PCSs adopt a cellular system, which divides an area into cells. A *cell* is the geographical area in which a mobile user can communicate with a particular base station that connects a wire line network with wireless network. Each cell uses assigned frequencies, which allow for efficient use of limited frequency resources. A *base station* offers interfaces between wired and wireless networks, and it sends and receives signals with mobile devices, i.e. PCS. Also a base station controls network resources within a cell.

When a mobile user turns on his or her cellular phone, the device sets a link with the channel with strongest signal out of 33 set-up channels of 333 available channels for a short interval, which means that the mobile terminal *mostly* (it does not mean all the time) selects the nearest base station. The device delivers service requests through the channel that is already set. Then the base station is linked with mobile phone and selects a directive antenna. When the user gets out of the coverage of the cell where he or she is calling, the cellular phone is assigned new channel from the other cell in a process called *hand-off*. When the user terminates his or her call, the link between the base station and the mobile device is removed and start paging-monitoring that detects the mobile device in the cell.

The data used in this study are base station location data that contain coordinates, polling times, addresses of the base stations, status of mobile phone, and the number of the mobile phones. That means these data do not indicate the real location of cellular phone user, but the location of the base station connected to the user. This is a very important point in this study because most of the processes mentioned in the paper are efforts to overcome the limitation. Figure 1 shows an example of the PCS base station data.

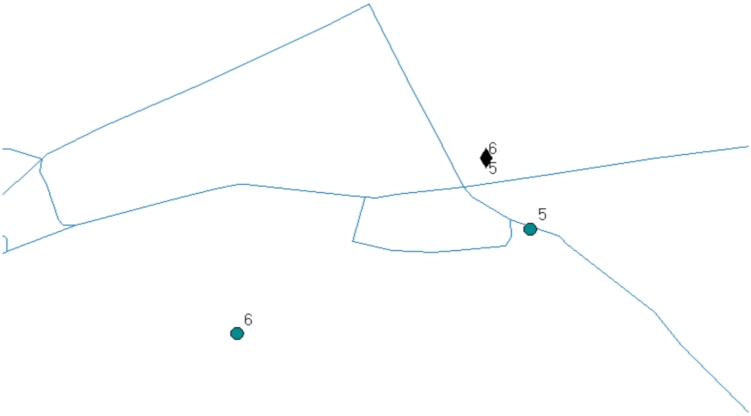


Fig. 1. PCS Base Station Location and Actual User Location

As shown in Figure 1, the numbers indicate successive points of time (a 5-minute interval) through with mobile user drives. The diamond-shaped point is the location of the base station that is connected with the mobile user that is symbolized in shape of dot. Even though the location of the cellular phone changes with time, the base station does not change because the user stays in the coverage area of the same base station. The two different locations of the user between these time frames (i.e., point 5 and point 6) cannot be distinguished from each other via base station data. This situation is an obstacle to detecting the position of mobile devices.

Another limitation is the collecting time interval. The location information of a user is normally collected at five-minute intervals. Under this time frame the location of the user would not be changed because PCS still might be in the coverage area of the cell. As a result, the estimation of the path on which the mobile phone user travels is needed. If a shorter collection interval time is provided, more accurate location estimates can be made because the time when the hand-off occurs can be estimated.

3.2 Creating Travel Time Information from PCS Data

The cellular network that is not supported by any network method for detecting the location of users cannot give the location of mobile phone by itself but only of the base station that belongs to the cell in which user is located. However, sequences of PCS data can give clues to enable one to figure out the path along which a cellular phone user moves. This section explains the methodology of collecting traffic information including average speed and travel time from PCS data.

Before creating algorithms to get information from PCS data, an analysis of the data is required. Figure 2 shows the comparisons of the data that contain the location of a vehicle travelling in the Seoul area, by PCS and GPS.

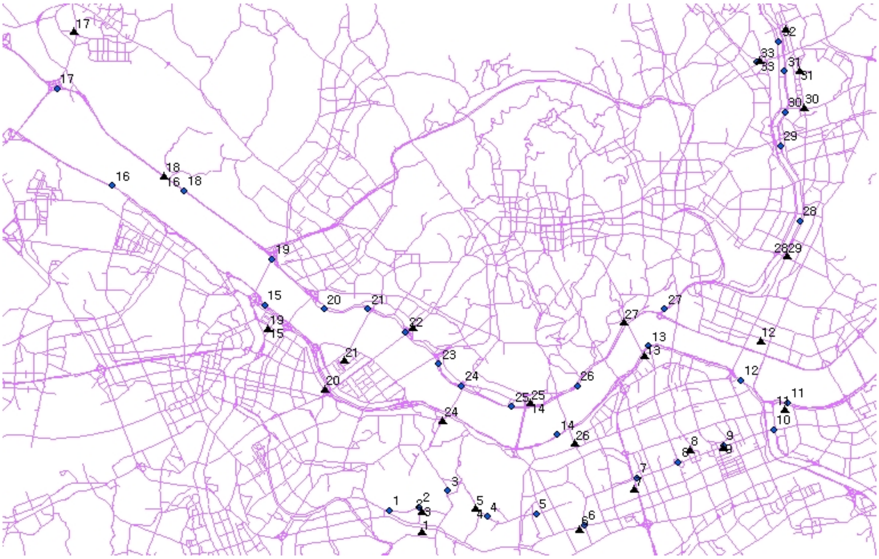


Fig. 2. Estimated and Actual Mobile Phone User Routes

As shown in Figure 2, the triangles stand for the user locations gathered from PCS data, and the round dots symbolize the locations obtained from GPS. These locations are regarded as actual positions of the vehicle. The same numbers on the map indicates the same time frames.

Figure 3 shows the locations of PCS base stations in the Seoul metropolitan area with the Han River flowing through middle of the map. Areas where data were collected for this study include the metropolitan freeways along the both sides of the river. There are no obstacles between these freeways to block signals between a given base station and a mobile unit. As a result of this absence of interference, a mobile device in this area can be connected to the base station that is located on the other side of river. This seems to cause fluctuation of the reported locations along the river according to the PCS data. This variation of data makes it hard to determine which route the user takes. Other routes in the downtown area seem to have little fluctuation.

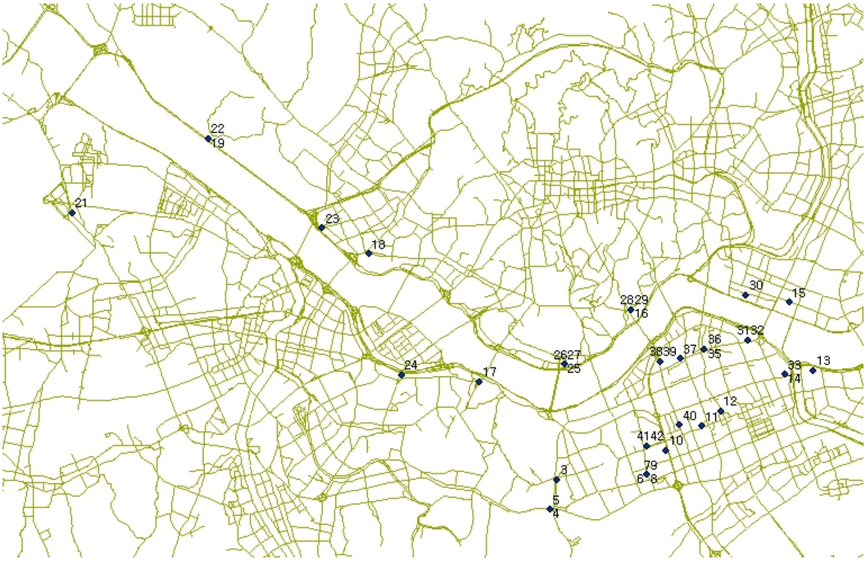


Fig. 3. Estimated Routes by PCS Base Station Data

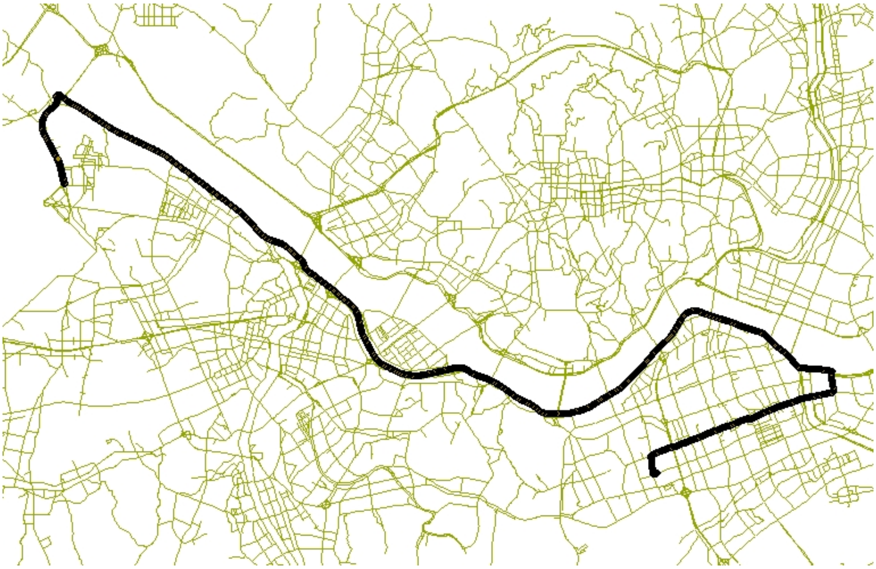


Fig. 4. Actual Routes Along Which the Vehicle Travelled

This route is located in an area that has a high density of base stations; therefore, base stations are not required to cover a large area. Figure 4 shows the real route that the vehicle travelled.

Some process to correct the erroneously reported route is needed. In order to correct the errors between the actual and the estimated routes, the road along both sides of the river should be taken into account. For example, if data indicate a point on one of those two roads, the other road should be checked to see whether the data belong to it or not. Also monitoring the trend of data movement is helpful in determining where the vehicle is.

The following steps are proposed to help distinguish one location from the other and detect the location of the vehicle.

1) Set the points on the road map corresponding to the base station location data.
2) Check the position before entering the freeway beside the river.
3) Determine possible alternative routes and locations relative to the pre-checked location.
4) If current location is the same with the location at previous time frame, wait until next the data are available and go to step 3. If not, go to step 5.
5) Calculate the distance for each route and set the criteria in terms of measurement such as speed, distance, and travel time while considering traffic condition.
6) Select the available location in step 3 according to the criterion that is set in step 5.
7) Calculate the speed between the locations.
8) Calculate travel time between specific locations if necessary.

In step 1, it is necessary to set the nearest point on the road from the base station as the location of mobile user.

4 Validating Travel Time Information Based on GPS Data

The average speed of each segment corresponding to time frame along the predefined route was calculated and the results are as follows.

The value of the t-test indicates that the average of the average speeds from the two different kinds of data sources are the same, but the speed from the PCS data moves up and down on a line that shows the actual speed. A series of data that comes from the same cell constitutes the same base station location data, which results in a reported speed of 0 during that time. Also there are losses and gains of distance between continuous data. Therefore, there are discrepancies between GPS reported speeds and PCS reported speeds, as shown Figure 5.

The graphs above show that the longer the time period based on the average speed, the more similar it is to the graph of the actual speed of mobile user. As the average error decreases, the interval between time frames gets longer. The 20-minute average speed data would be closer to the actual speeds. However though more accurate data can be obtained with longer time intervals, the value of information decreases. The average speed data for an hour with very high accuracy in this matter would not represent to be useful for a driver in downtown area in the case of traffic jam. Therefore, there are some trade-offs between accuracy and length of data collecting time intervals of data. The 10 to 15-minute time periods, in which over 80% accuracy is achieved, are proposed in this study.

Table 1. Comparison of Speed from 5-minute Time Interval

Time frame	GPS distance (Meters)	PCS distance (Meters)	Remark	Speed (GPS) (Km/hr)	Speed (PCS) (Km/hr)	Relative error
8	1389	2056		16.668	24.672	0.48020158
9	1524	949		18.288	11.388	0.37729659
10	1628	2113		19.536	25.356	0.29791155
11	1607	1213		19.284	14.556	0.24517735
12	1674	2095		20.088	25.14	0.25149343
13	3168	2968		38.016	35.616	0.06313131
14	4019	3921		48.228	47.052	0.02438418
15	11213	10411	10 min	67.278	62.466	0.07152412
16	6075	5576		72.9	66.912	0.08213992
18	5209	4457		62.508	53.484	0.14436552
19	3522	5514		42.264	66.168	0.56558773
20	2255	1328		27.06	15.936	0.41108647
21	1435	274		17.22	3.288	0.80905923
22	1455	2288		17.46	27.456	0.57250859
23	1482	0		17.784	0	1
24	1057	2550		12.684	30.6	1.41248817
25	1717	2266		20.604	27.192	0.31974374
26	2194	875		26.328	10.5	0.60118505
27	3882	3260		46.584	39.12	0.16022669
					Average error	0.41523743
				t-test value	0.63281587	

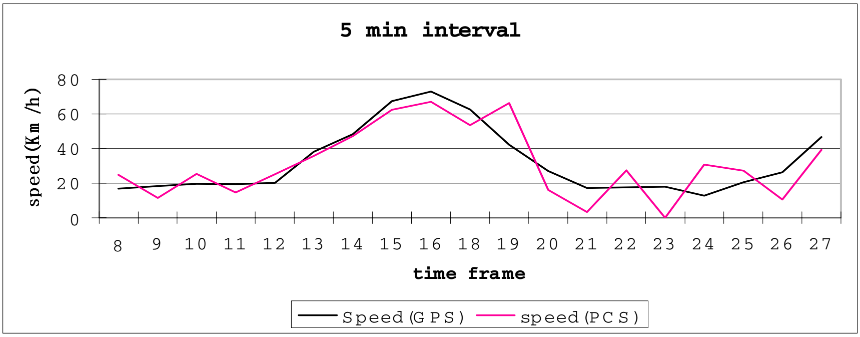


Fig. 5. Comparison of Speed Calculated on 5-minute Time Interval

To reduce this effect, the average speed over a longer time period is calculated for two successive route segments. Average speeds calculated based on 10-minute and 15-minute time periods are shown in Figures 6 and 7, respectively.

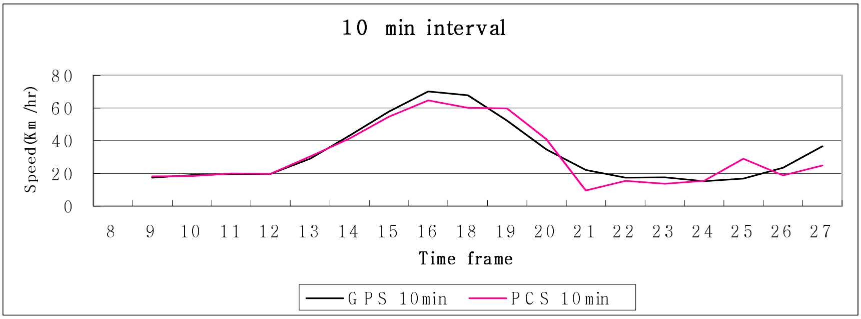


Fig. 6. Comparison of 10-minute Average Speeds

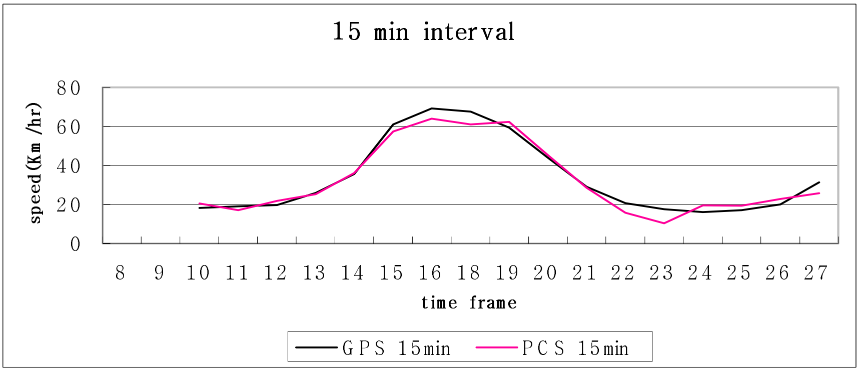


Fig. 7. Comparison of 15-minute Average Speeds

5 Conclusions

This paper has attempted to determine the characteristics of a PCS system that uses cellular systems, and has made an effort to correct errors made while collecting transportation information from PCS data.

A prototype was developed in this study as an integrated system which provides real time information through PCS. This would be able to be implemented on the provision of travel time based on the demand of the PCS users as one of services.

Although it is impossible to find the exact location of a mobile phone user with single base station location data, the route of a mobile user can be closely approximated. Assigning specific locations to corresponding base stations, determining the route of mobile phones and calculating differences in route distance are the steps needed to obtain average travel times for segments along the route. This technique includes a certain amount of error, but adjusting the time frame period is a solution to that problem. The 10- to 15-minute time periods are recommended in this study as a way to reduce the errors. However, the methodology and the algorithm developed in this study needs to be improved by utilizing the location based services (LBS) for a commercial system and/or service, which provides real time traffic information on demand such as congestion, link travel time, incidents or accidents, emergency conditions, toll collections, etc.

The limitation of data restricts the application of PCS data and makes it difficult to improve the accuracy of data. It seems difficult for PCS-based transportation information services to provide services to the user that require high accuracy, but the travel time between locations that are considerably far apart from each other can be roughly determine. PCS data may be used on Inter-city highways and urban freeways at low cost. If there is sufficient data, any statistical analysis can be applied. Also it can be possible to abstract specific information from many trends in traffic flow.

References

1. Olariu, S., Pintti, M. C., and Wilson, L.: Greedy Algorithms for Tracking Mobile Users in Special Mobility Graph, Discrete Applied Mathematics 121 (2002) 215-227
2. Turner, S. M., Eisele, W. L., Benz, R. J., and Holdener, D. J.: Travel Time Data Collection Handbook. Federal Highway Administration (1998)
3. Park, T., Lee, S., and Moon, Y. J.: Real Time Estimation of Bus Arrival Time under Mobile Environment. Lecture Notes in Computer Science 3043 (2004) 1088-1096
4. Quiroga, C.A. and Bullock, D.: Travel Time Studies with Global Positioning and Geographic Information Systems: an Integrated Methodology. Trans. Research 6C (1998) 101-127

A PKI Based Digital Rights Management System for Safe Playback

Jae-Pyo Park[1], Hong-jin Kim[2], Keun-Wang Lee[3], and Keun-Soo Lee[4]

[1] School of Computing, Soongsil University, Korea
pjerry@dreamwiz.com
[2] Dept. of Computer Information, KyungWon College, Korea
[3] Dept. of Multimedia Science Chungwoon University, Korea
[4] Dept. of Computer Engineering Hankyong National Univ., Korea

Abstract. In this paper we first propose an I-frame encryption scheme for encryption of moving image video data. Second, we propose a licensing agent which provides automatic user authentication and data decoding when multimedia data encrypted in the system server is executed in the client system by the user. The licensing agent performs user authentication based on Public Key Infrastructure (PKI) using a shared key pool and encryption/decryption of multimedia data. After designing and implementing the proposed system, performance tests were then performed using video data files of various sizes for performance evaluation. We verified that the proposed system significantly reduces delay time, including decryption time, when playing back video data files in the client system compared with existing systems.

1 Introduction

Existing DRM implementations do not take privacy protection into consideration for the reason that user privacy protection is not directly necessary for copyright protection. Therefore, user information leaked during user authentication for license issuing, and usage details reporting the process of monitoring against illegal content usage has caused user privacy infringements [1-4].

Methods for implementing security for digital contents can be divided in two categories: upper-level security and lower-level security [5]. Upper-level security schemes relate to user authentication, while lower-level security relates to the protection of the data itself. User authentication means that an authenticated user is not subject to limitations in content usage [6]. Therefore, a function used for monitoring the amount of content usage in order to maintain information on the number of users accessing a specific multimedia content is required [7]. However, since in this scheme an authenticated user can obtain illegitimate copies of data that can be distributed, it cannot provide perfect protection of distributed data. As such, protection in the data itself performs encryption on the content to restrict user's access to the content. Therefore, in DRM, security is implemented by performing encryption on the content data itself.

The algorithms used for encryption are the private key algorithm and public key algorithm. The secret key algorithm is an algorithm which performs encryption at

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 801 – 810, 2005.

high speed by using a single key for encryption. However, in this method, there is the problem of key distribution; that is, the sender and recipient have to exchange their private keys beforehand. In addition, if large capacity moving images are transferred, a large amount of processing time is required if a transfer is carried out simultaneously with encryption. The public key algorithm offers the advantage of using separate keys for encryption and decryption so that the sender and recipient can safely exchange keys. However, its drawback is slow execution speed.

In this paper we propose a shared key pool scheme which encrypts the secret key using each user's private information in order to prevent exposure of the secret key by the user while performing authentication for digital content users. This proposed scheme prevents the exposure of the secret key by the user while encrypting contents beforehand to improve transfer speed. In addition, by using a licensing agent, a license is downloaded from the licensing server which manages the licenses within a database when executing content so that offline execution is possible. For security of the transmitted key pool information, the key pool information is encrypted using a secret key and the relevant secret key is encrypted using the user's public key in order to improve the encryption speed and security of the encrypted key pool which is being transmitted.

2 DRM System

Using DRM technology, international companies such as InterTrust and Microsoft, as well as Korean companies such as Digicap, offer various types of DRM solutions [8]. However, since existing DRM solutions perform encryption using secret keys when the user downloads files, a large amount of encryption time is required. In addition, decryption must be performed first for large capacity contents so that the user cannot perform the playback of the file in real-time. Besides, if the key used for encryption and decryption is exposed by the user, the copyrighted content can no longer be protected.

Existing DRM solutions perform static copyright management by inserting information such as data protection conditions or copyright management into moving image data. As such, dynamic copyright control is difficult and data needed to prove illegal activities in case of copyright infringement is difficult to obtain. Therefore, existing DRM solutions use software agents to monitor a user's data usage in order to solve this problem, but this solution is subject to the functional constraints present in off-line usage environments. Hence, a digital copyright management technology which is applicable to all types of contents in both online and offline environments, and is capable of dynamic copyright management and real-time management and tracing, should be developed.

Microsoft's WMRM (Windows Media Rights Manager) is an end-to-end DRM system which provides safe distribution of digital media files to content providers and consumers [9, 10]. WMRM delivers media such as music, video, etc., that is protected in encrypted file format, to content providers through the Internet. In WMRM, each server or client instance receives a key pair through an individualization process, and instances that are determined to be cracked or unsafe are excluded from service through the certificate cancellation list. WMRM is widely used in embedded form with the Windows Media Player, but it shows limited adaptability to dynamic environment changes, is only applicable to the Windows Media Player, and only supports a limited range of file types. In addition, it has the disadvantage of potential leakage of user information such as user ID or e-mail address since no particular protection technology is applied on the certification stage for issuing licenses.

3 The DRM System for Safe Playback

For data protection and authentication of original content, data should be protected not only by simple access restriction or password authentication, but also through user authentication and data encryption implemented by PKI technology. The proposed system is a client/server configuration and its overall layout is illustrated in Figure 1.

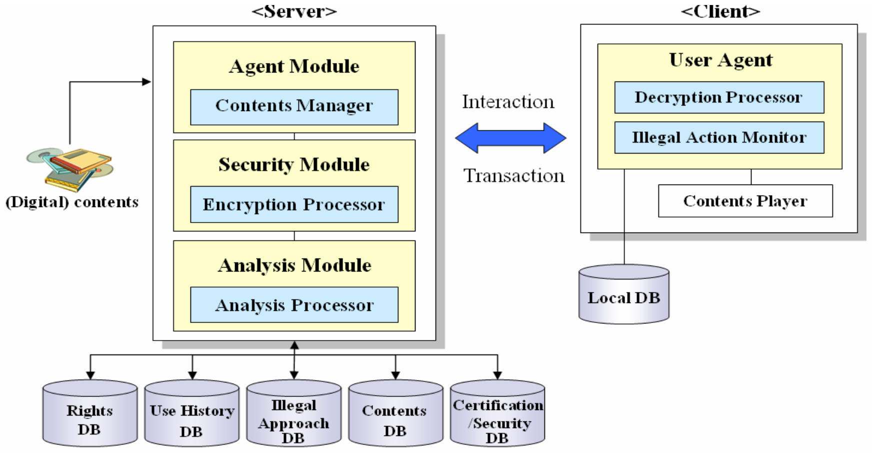


Fig. 1. System Architecture

When content is registered on the system server through the external interface, content monitoring processing is performed by the agent module and an encryption is placed on the content. When the user accesses the content, user authentication is performed by the licensing agent that is dispatched by the server. If the user is authenticated, the content is executed by an application program; otherwise, a warning message is displayed. Monitoring against illegal usage is performed on the content by the licensing agent, and all illegal user activities are stored on the server interface through the monitoring interface. Even in the case of authenticated users, content is protected by encryption of the content itself to restrict content usage according to access privilege levels.

3.1 Encryption and Decryption Using the Shared Key Pool

The content's author sends the generated content to the server. The server then encrypts that content using an arbitrary secret key (Ks) and stores the encrypted content C together with the secret key (Ks) on the server's content database.

$$C = EKs\ [data\ ] \tag{1}$$

The user can download a desired content from the server through the authentication process or copy it from another user. However, downloaded content is encrypted and therefore has to be executed through the agent. The server generates a shared key pool for encrypting the secret keys in order to prevent leakage of the secret key through the user. The server encrypts the I-frame of the content's GOP using a secret key through either AES or SEED algorithm and stores that content's ID and secret key on the server's database. An arbitrary shared key pool applicable in the encrypted content is then generated and is also stored on the database. If a user is registered, the server performs user authentication using the user's certificate, and then extracts the user's information from the private key pool using the user's certificate in order to generate the user's key. Private information and the encrypted shared key pool are stored on the server's database and also on the user's database using the user agent.

To generate a shared key pool, the content producer can encrypt the content to be distributed using a secret key (Ks) and divide it into k bit columns as in Equation 2.

$$Ks = Ks_1 \mid Ks_2 \mid \ldots \mid Ks_k \tag{2}$$

Generally, in secret key encryption, a secret key (Ks) with a length of 128 bits is used. The shared key pool consists of $k*2^{\frac{n}{k}}$ bits and is generated according to Equ. (3).

$$\left\{ a_1^0, a_1^1, a_1^2, \ldots, a_1^{2^{\frac{n}{k}-1}}, a_2^0, a_2^1, a_2^2, \ldots, a_2^{2^{\frac{n}{k}-1}}, \ldots, a_k^0, a_k^1, a_k^2, \ldots, a_k^{2^{\frac{n}{k}-1}}, \right\} \tag{3}$$

In order to adapt to the key's size, an array with k rows and $2^{\frac{n}{k}}$columns can be expressed as in Table 1.

Table 1. Shared Key Pool of Ks

a_1^0	a_1^0	…	$a_2^{2^{\frac{n}{k}-1}}$
a_1^0	a_1^0	…	$a_2^{2^{\frac{n}{k}-1}}$
:	:	…	:
a_1^0	a_1^0	…	$a_2^{2^{\frac{n}{k}-1}}$

The private key of each user, Kp, is a set of bit columns consisting of k bits as in Equ.(4).

$$Kp = a_1^{b_1} \mid a_2^{b_2} \mid \cdots \mid a_k^{b_k} \tag{4}$$

Here, b_i corresponds to the value of each ith row of the key pool, as shown in Equ. (5), and it is an important value determining each user's private key. b_i is extracted from the public key of the user certificate.

$$B = b_1 \mid b_2 \mid ..... \mid b_k \tag{5}$$

The length of the user's public key is 512 bits if required secrecy is low. For critical information requiring high secrecy, a key length of 1024 bits is used. In general, if a public key of n bits length is used, the value of each item within the key pool falls in the range of $2^{\frac{n}{k}} (0 \sim 2^{\frac{n}{k}} - 1)$. For example, if the public key's length is 512 bits while the private key's length is 128 bits, 512/128=4; therefore, each item is 24=16 which corresponds to a value range of 0-F in hex. Therefore, the individual rows of the actual private key are determined by the public key's value which is in the range of 0-F. The range of each key's value according to the length of the secret key and the length of the public key is summarized in Table 2.

Table 2. Value of Key Pool

Size of Secret Key Size of Public Key	128	256	512
512	0 ~ F	0 ~ 3	0, 1
1024	0 ~ 255	0 ~ F	0 ~ 3
2048	0 ~ 65535	0 ~ 255	0 ~ F

Table 3. Encrypted Shared-key Pool by Ks

a'^{0}_{1}	a'^{0}_{1}	…	$a'^{2^{\frac{n}{k}-1}}_{2}$
a'^{0}_{1}	a'^{0}_{1}	…	$a'^{2^{\frac{n}{k}-1}}_{2}$
:	:	…	:
a'^{0}_{1}	a'^{0}_{1}	…	$a'^{2^{\frac{n}{k}-1}}_{2}$

Therefore, since each user's private key is determined by each user's unique public key, the key value selected from each row by the public key guarantees that each user is assigned a different key. When the shared key pool is generated, a bitwise XOR is executed for encryption of the secrete key (Ks) on $a_i^0, a_i^1, ..., a_i^{2^{\frac{n}{k}-1}}$ ($2^{\frac{n}{k}}$ bits) with each bit of Ks_i for each i th row of the shared key pool, as shown in Equ. (6).

$$a'^{0}_{i} = Ks_i \oplus a^{0}_{i}\,,\quad a'^{1}_{i} = Ks_i \oplus a^{1}_{i}\,,\quad \ldots\ldots\ldots..,\quad a'^{F}_{i} = Ks_i \oplus a^{F}_{i} \tag{6}$$

The shared key pool in an encrypted form can be obtained through the calculation of Equation 6 as shown in Table 3.

Table 3. Encrypted Shared-key Pool by Ks

Then, the encrypted shared key pool is forwarded to the user agent through the network. The agent finds the secret key (Ks) using the user's public key (Kp) from the encrypted key pool in order to decrypt the encrypted content. The moving image content file is decrypted using this secret key (Ks) and then displayed to the user.

3.2 License Authentication Protocol

In order to use copyrighted content, user authentication is required as illustrated in Figure 2. User authentication is carried out through member enrollment, and members can log in to download files. However, users who have received moving image data from other users can also redistribute contents, and therefore, anyone can join and login through PKI-based certificates. While a separate login scheme using ID and password is also supported, even in this process, the PKI-based certificate is always verified for login. If authentication is made through ID/password- or certificate-based login processes, moving images can be downloaded.

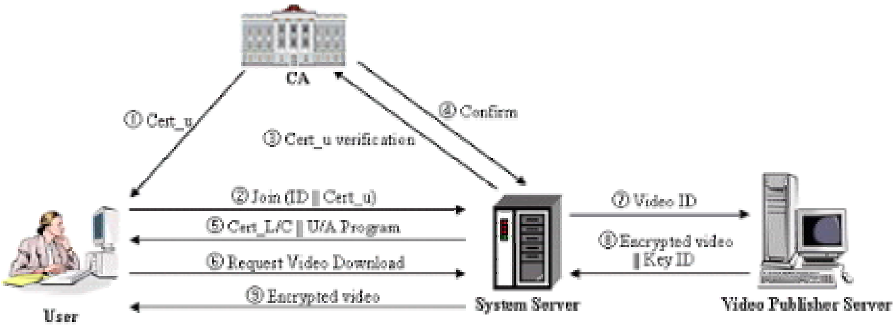


Fig. 2. User Verification Protocol

The user connects to the system server and transmits his certificate Cert_u. The system server verifies the user's certificate Cert_u through the authentication path in the CA server. If the certificate is correct, the user agent program and the server's certificate are transmitted to notify the user that authentication has been completed successfully. However, the moving image data is encrypted, and therefore cannot be executed directly after a download. The agent downloaded during the authentication process must be installed so that the execution of the moving image content can be requested through the agent. When the user executes copyrighted content, the licensing agent verifies the user's license. If a license exists, it is authenticated through the server; if there is no license, a license is issued.

As illustrated in Figure 3, when the user executes an encrypted content, the licensing agent checks whether a license is present. If it is not present, a license is issued according to the license issuing protocol. If a license is present and the client is on-line, authentication for the relevant license is requested to the server. If off-line, the license is authenticated for the client according to the information stored in the user's database, supporting the execution of up to a specific number of content.

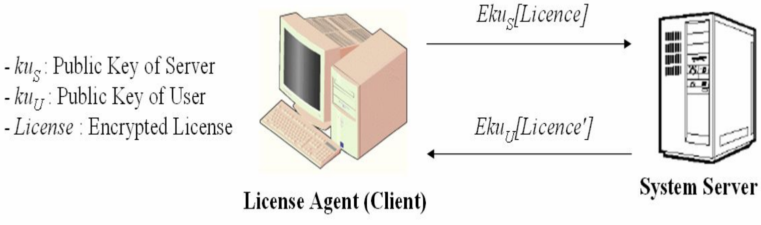

- ku_S : Public Key of Server
- ku_U : Public Key of User
- *License* : Encrypted License
Eku_S[Licence]
Eku_U[Licence']
License Agent (Client)
System Server

Fig. 3. License Authentication Protocol

The system server, when it receives a license authentication request from a licensing agent, compares the relevant license's information stored in the database with the client's license information, and then corrects and authenticates the license information. The license storage database status is shown in Table 4.

Table 4. License Information in Database

Licence ID	user ID	Data ID	Auth.	Auth. Value	Conn. Count	System num.	Private Info.
1	11111	s11111	1	10	2	203.253.21.174 162.192.56.39	12345678
2	22222	a11111	2	04-3-12	1	203.253.27.162	87654321
3	33333	k11111	1	5	1	223.65.198.45	33333333
:	:	:	:	:	:	:	:

The system information of the user's license is verified and added to the server's database. If the license is time-limited up to a certain date, it checks whether the license has expired, and if the license is a count-limited one, the license information is corrected, the corrected license is encrypted using the user's public key, and it is then transmitted. The user agent which has received the corrected license from the server corrects the client side database information based on that license and generates the secret key based on the license's private information and the shared key pool's value. It then decrypts the encrypted content using the secret key for display to the user.

The proposed environment for development is based on an Intel(R) Pentium-IV CPU (2.4GHz), 512MB RAM, and the MS Windows 2000 Server operating system. The programming languages used for implementation were Visual C++ 6.0 and Delphi 7.0. The decryption of the moving image is performed as illustrated in Figure 4, and is done by the agent when the user executes the moving image. When executing the user's content, the user agent decrypts the moving image using a key.

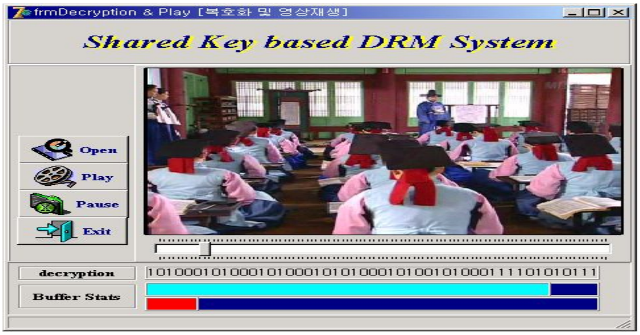


Fig. 4. Decryption Processing for Video Data

4 Performance Evaluation

To evaluate the performance of the system proposed in this thesis, we measured the encryption time of the video itself and the initial play-out delay time according to the decryption time. The conventional method of first decrypting an already encrypted video data file (non-real-time decryption method), and the method proposed in this thesis, that is, playing out while performing decryption in real-time (real-time decryption method), were implemented and their respective execution times were measured with the results shown in Table 5. For accurate time measurement, the video file was divided by minutes. Each delay time is the time required for decrypting the encrypted video data file and play-out. This time is the sum of the video data file's decryption time and the loading time. In general, the video file's loading time differs for each video player; therefore, in this experiment, the loading time was processed together

Table 5. Delay Time Comparison for Execution Time

File Size (MBytes)	Execution Time (Minutes)	Decryption Time (Seconds)	Delay Time of Existing Method (Seconds)	Delay Time of Proposed Method (Seconds)
6.83	1	0.76	2.42	2.42
13.66	2	1.57	5.50	2.42
20.49	3	3.00	9.24	2.42
68.31	10	9.05	25.01	2.42
204.94	30	24.31	49.11	2.42
423.94	60	41.62	82.06	5.50
635.92	90	59.46	104.72	5.50

with the decoding time to calculate the delay time. In the proposed method, video data files are played out concurrently while performing execution by using double buffer scheduling. Therefore, we can see that the play-out delay time significantly decreased compared with the conventional method.

5 Conclusion

In this paper we proposed a new encryption scheme which encrypts the video data's I-frame for encryption of moving image data. The licensing agent performs PKI-based user authentication and encryption/decryption of moving image data using a shared key pool when executing the user's multimedia data. After encrypting a moving image file using a secret key, the user's private information is extracted from operations with the PKI certificate and the shared key pool, and then the secret key is transmitted to the user hidden within the shared key pool so that the user cannot expose the key to the outside by accessing the secret key. If a key is exposed, the path of exposure can be traced. The shared key pool system is a methodology for effectively countering the exposure of the key by a user. If a user executes a moving image file, the licensing user authenticates the license at the system server, calculates the secret key based on the user's personal information and the shared key pool information, and then the moving image file can be decrypted for play-out. Here a double buffer is employed, which enables execution if a part of the file is decrypted, so that decryption can be performed in real-time for the user.

References

1. Joshua, Duhl and Susan, Kevorkian: Understanding DRM system: An IDC White paper. IDC (2001)
2. Jai Sundar, B., Spafford E.: Software Agents for Intrusion Detection. Technical Report, Department of Computer Science, Purdue University (1997)
3. Dubl, J.: Digital Rights Management: A Defination. IDC (2001)
4. Dubl, J., Kevorkian, S.: Understanding DRM system: An IDC White Paper. IDC (2001)
5. Kentaro, Endo: The Building up of National Regional and International Registers for Works and Objects of Related Rights. Proc. of International Conference on WIPO, Seoul, Korea October (2000) 25-27
6. Gupta, V. K.: Technological Measures of Protection. Proc. of International Conference on WIPO, Seoul, Korea October (2000) 28-29
7. Vora, P., Reynolds, D., Dickinson, L., Erickson, J., Banks, D.: Privacy and Digital Rights Managements. A Position Paper for the W3C Workshop on Digital Rights Management, January (2001)
8. Mulligan, D. K. and Burstein, A.: Implementing Copyright Limitations in Rights Expression Languages. In 2002 ACM Workshop on Digital Rights Management, Washington DC, November 18 (2002)
9. Erickson, J. S.: Fair use, DRM, and Trusted Computing. Communications of the ACM, vol. 46, no. 4, April (2003) 34–39

10. 10.Microsoft's Press Releases of the PocketPC 2002 Launch, Available at ww.microsoft.com/presspass/events/pocketpc2002/default.asp, Oct 8 (2001)
11. John, Linn: Trust Models and Management in Public Key Infrastructures. Technical Notes and Reports of RSA Laboratories, November (2000)
12. Russ, Housley and Tim, Polk, Planning for PKI, John Wiley & Sons (2002)

Scheduling Method for a Real Time Data Service in the Wireless ATM Networks

Seung-Hyun Min[1], Kwang-Ho Chun[2], and Myoung-Jun Kim[1]

[1] Dept.of Computer Sicence , Chungbuk National University, Korea
{imturtle,khchun}@iita.re.kr
[2] Dept.of of Electronic & Information Eng. Chonbuk National Univ., Korea

Abstract. This paper proposes an improved Traffic-Controlled Rate Monotonic (TCRM) priority scheduling algorithm as a scheduling method in order to transmit real-time multimedia data in a wireless ATM networks. A real-time multimedia data transmission scheduling policy is applied using a different method to uplink or downlink states according to the wireless communication environment, and guarantees the requirement of QoS for both real-time and non-real-time data. In addition, it deals the issue of fairness to share and distribute insufficient wireless resources. Moreover, an issue of inefficiency for non-real-time data, which is a demerit of a TCRM, can be solved using an arbitrary transmission speed by configuring a Virtual Control (VC) in a Base Station (BS).

1 Introduction

In recent years, the use of mobile devices, such as cellular phones, PCS, and other devices, has rapidly increased, and the need for wireless multimedia services will significantly increase in the very near future according to the increase in various multimedia services. Studies have been conducted to convert the ATM method to an integrated environment of wired and wireless services in order to provide these wireless multimedia data. In addition, studies on a wireless ATM network, which extends and supports various multimedia data services provided by the existing wired ATM networks to wireless sections, have been processed with the ATM Forum as the central figure throughout the world [1].

Although a wireless ATM networks is a worthwhile way to provide user's mobility in addition to the existing advantages of a wired ATM, flexible assignment of bandwidth, and guarantee of the QoS for various services, it has some differences to a wired ATM network in the context of special properties, such as limited bandwidth, high transmission delay rate, bit error rate, and mobility. The cell structure of wireless ATM network considered at the present time has some problems, which include a frequently occurred hand-over, due to a decrease in the size, such as micro/pico cells, and QoS guarantee owing to a sudden change in traffic. Thus, a number of studies have been conducted to apply the ATM to wireless environments [2].

Multimedia services can be represented by certain properties, which consider its specific qualities using specific standards, such as data transmission rates, acceptable

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 811 – 820, 2005.

transmission error rate, and acceptable maximum delay time. Multimedia services may have different requirements according to whether they are provided under a wired network or wireless network, even though they offer the same service. Although the objective of wireless ATM networks is to provide similar services of wired ATM networks, there are some differences between wired networks and wireless networks in their physical properties and inevitable limitations.

This paper proposes a scheduling method, which reduces the loss of non-real-time data, while a limitation of delay time for real-time data is guaranteed, in order to ultimately use the limited bandwidth in wireless ATM networks, and to provide high quality services similar to that of wired ATM networks by classifying the used multimedia data as real-time data, such as voice and video, and non-real-time data, such as text and image.

2 Related Works

Multimedia data can be divided into two categories, such as static data, which requires a strict error control according to the property of the multimedia data, and dynamic data, which requires a real-time data transmission. In the case of the static data, such as text and images, it isn't sensitive to the flow of time, but requires a perfect error control. Conversely, the dynamic data, such as voice and video, isn't sensitive to the strict error control, but requires not only the properties of real-time and continuity, but also the synchronization between data. In order to achieve an effective transmission of real-time multimedia data, it is necessary to guarantee the delay limitation of real-time data, and to minimize the loss of non-real-time multimedia data [5].

A TCRM method [3] applied in wired ATM networks satisfies a transmission speed of multimedia data required by users only using a simple traffic controller and rate-monotonic priority scheduling [4]. This TCRM assigns real-time data to each real-time channel using a traffic controller, and transmits the data by assigning the priority according to the speed of real-time data using the rate monotonic priority scheduler.

In the case of the application of a preempt RMS algorithm to a scheduler, a cell transmission process may be preempted by other cells when cells are transmitted. Thus, a non-preempt RMS algorithm is to be used. The disadvantage of the TCRM policy is that it never transmits non-real-time data when real-time data continuously exists in real-time channels. Due to this disadvantage, a storage space in a buffer is infinitely required to store the non-real-time data. Otherwise, there is a loss of the non-real-time data.

3 Scheduling Policy in Wireless ATM Networks

A TCRM scheduling policy in wireless ATM networks can be applied using the TCRM scheduling used in wired ATM networks by extending it to wireless ATM networks.

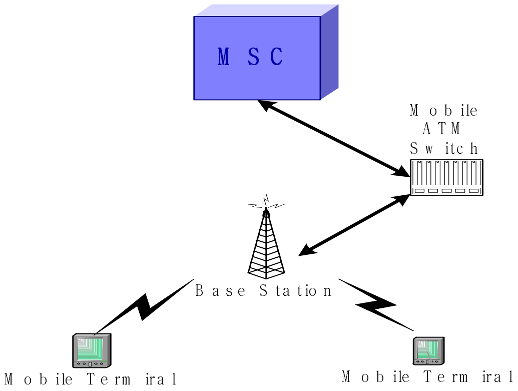


Fig. 1. Communication system in wireless ATM networks

All ATM services, which include all types of data in real-time and non-real-time, should be transmitted under certain network environments, which ensure the QoS for a multimedia environment, where the requirements of bandwidth has dynamically changed. In addition, channels should be fairly and effectively used in the network. Thus, a TCRM scheduling policy should be applied by dividing it into uplink and downlink states according to the wireless ATM network environment. The wireless ATM environment proposed in this paper provides services to several Mobile Terminals (MTs) based on a single Base Station (BS) as illustrated in Fig.2.

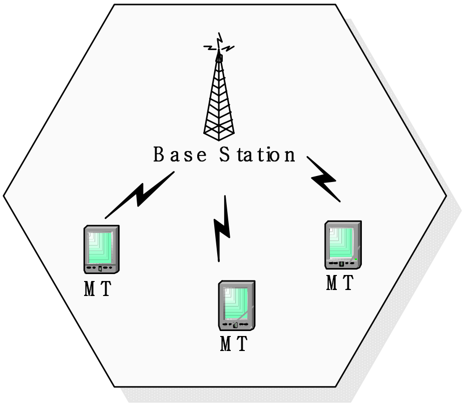


Fig. 2. Scheduling environment in wireless ATM networks

- A channel, which transmits data from each MT to the BS, is called an uplink channel, and a channel, which has a reverse transmission, is called a downlink channel.
- Each MT is able to transmit data to the BS, and to receive data from the BS. However, it is impossible to transmit or receive data between MT.
- An uplink channel is shared by all MTs.

- Because a downlink channel is a type of broadcasting channel monopolized by the BS, following scheduling of the BS can perform the transmission.
- A multiple access method uses a dynamic time division multiple access for both uplink channels and downlink channels. A dynamic TDMA method generally presents a high flexibility in the acceptance of a required bit rate to connect each link by assigning a proper number of time slots according to the condition of the present traffic condition.

A leaky bucket model [8], which is used to configure the input traffic in this paper, is applied, and the leaky bucket model can be expressed as (a_i, p_i). The entire scale of the cell configured in a bucket is presented as a_i, and the network penetration speed is presented as p_i. The scale of the network packet is fixed the same as the cell used in an ATM network. To provide a real-time communication service, an UNI is required for each ATM switch. The UNI requires a buffer space of a_i to protect the loss of cells, and transmits the cell to the entrance of a network with the speed of p_i.

3.1 Method for the Consideration of Non-real-time Traffic Transmission

The disadvantage a TCRM policy has in achieving real-time data transmission is that it never transmits non-real-time data when real-time data continuously exists in real-time channels. Due to this disadvantage, a storage space in a buffer is infinitely required to store the non-real-time data. Otherwise, there is a significant loss of non-real-time data. In order to solve this problem, a VC is to be configured in the BS, which generates a Reservation Buffer (RB), and reduces the loss of non-real-time data using an arbitrary fixed speed of p_k. Real-time data configures a channel according to the speed of the user's requirements, and transmits data with a regulated transmission speed. However, it is difficult to recognize how many cells are stored in a RB, and how long it takes to store it in the RB in non-real-time data, because the data stored in the RB uses variable speeds that differ from the real-time data. Thus, the RB, which stores non-real-time data cells, is organized by a FIFO queue, and the VC configures the minimum threshold (i.e., 30% of the buffer size) and maximum threshold (i.e., full buffer size) and transmits the cell to a scheduler with an arbitrary fixed speed of p_k when the RB approaches the maximum threshold. Equ. (1) presents the required space of a RB for non-real-time data.

$$RB + 1 = U - \frac{p_k}{L} \tag{1}$$

where U is the total number of cells, which are entered in a RB from a number of arbitrary connections, M, at a certain time of t, and L is the size of a single cell. If U is smaller than p_k/L, a RB will require a space to store a single cell. As the configuration of an arbitrary fixed speed of p_k, if the value of p_k increases, the bandwidth of non-real-time data will increase. Thus, the bandwidth of real-time data decreases as much as the increased value, and the link usability of real-time data decreases by p_k.

The loss rate of cells can be reduced by controlling the value of p_k to transmit cells from a RB to a scheduler, or value of the cell storage space of a RB. Then, the problem of inefficiency in non-real-time data, which is a disadvantage of the TCRM, can be solved by this control. Fig. 3 presents the structure of the scheduling policy.

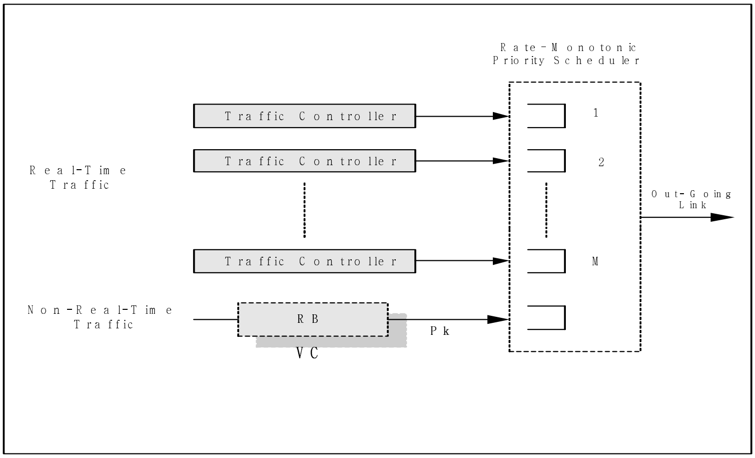


Fig. 3. Structure of the proposed scheduling policy

3.2 Scheduling Method in the Uplink

An uplink is the data transmission from MT to the BS. In the case of the uplink, the BS doesn't recognize whether the real-time data existed in the MT, or not. Thus, a loss of bandwidth is inevitable when data is periodically transmitted. In the case of the uplink, the BS performs a polling in advance for the MT, and transmits data when data existed in the MT by assigning a specific bandwidth. The transmitted data to the BS using this method applies a TCRM policy in an ATM switch. The polling period can be calculated using Equ. (2).

$$\text{Polling period} = M/p_i \tag{2}$$

where M is Packet size of a single packet , p_i is Data transmission speed in the channel i. Polling process is the following thing

1) The BS generates a polling token for a data transmission signal from the MT. The generation rate of the polling token presents the same rate for the packet generation of an actual data transmission signal.
2) For a data transmission signal, the first polling token is generated after p seconds. The time of p can be configured by a certain delay condition.
3) In a connection process, the BS performs several processes as follows.

- A polling token is generated in a VC, which corresponds to the actual data transmission signal of MT.
- The VC generates a polling token, which corresponds to a packet generated in an actual MT, and configures the priority using RMS according to the polling period.
- When the polling order is decided using RMS, the token is entered in a PQ (Polling Queue).
- The VC checks real-time traffic among the polling token in the PQ, where a token is removed in the case of existing real-time traffic, and performs a polling for the real-time traffic.

4) After finishing a polling for actual traffic, a packet is transmitted when the MT has a packet to transmit it to the BS. If there are no packets to transmit, an EOF (End Of File) signal is transmitted.
5) When the BS receives an EOF signal from the MT, the BS generates a polling token again after p seconds.
6) A polling is performed by calculating the usability of a link for non-real-time traffic when the transmission of real-time is finished or a new connection is configured. Then, a packet is transmitted.

3.3 Scheduling Method in the Downlink

A downlink is the data transmission from the BS to MTs. This link is a type of broad casting channel monopolized by the BS, and the transmission in this link can be achieved using a specific scheduling of the BS. Because the BS recognizes that there are some packets to transmit to MTs, a TCRM policy can be used in this case, except for a polling method, which is conducted in advance at an uplink. In order to configure a channel to achieve a real-time transmission, the user's requirement of the cell transmission speed of p_i and acceptance test is required. Then, the channel configuration is performed according to the traffic model. To calculate the arrival time of cells in a traffic controller, a type of logical arrival time is used. The calculation of the logical arrival time can be achieved using the arrival time of the previous cell in the same channel. The logical arrival time of the nth cell existing in the sth channel can be expressed as presented in Equ. (3).

$$X_n = \begin{pmatrix} A_{1,s} & n = 1 \\ \max(X_{n-1,s} + \frac{L}{p_i}, A_{n,s}) & n \geq 2 \end{pmatrix} \tag{3}$$

where X_n is arrival time of the nth cell, $A_{n,s}$ is arrival time of the nth cell in the sth channel, p_i *is* transmission speed of the channel i, and L is size of a single cell. Because the cell arrival time and cell transmission time are equal, the buffer space only requires a buffer space to store a single cell. Proposed TCRM scheduling process is the following thing.

1) The data transmitted from the MT can be divided into real-time traffic and non-real-time traffic using a traffic controller.
2) If the transmission speed of p_i for a real-time traffic packet passes the acceptance test, a real-time channel with the period of L/p_i can be configured at a traffic controller.
3) The non-real-time traffic data recognized by a traffic controller is to be stored in a RB, which is generated in the VC.
4) Non-real-time traffic data transmits an arbitrary transmission speed of p_k to the scheduler at the RB, and generates a virtual task with the period of L/p_i in VC.
5) A high priority is assigned to the real-time traffic data, which has a small period of L/p_i.
6) The real-time traffic data can be scheduled by a rate monotonic scheduler, and transmits the data according to the priority. If a non-real-time traffic task enters a

scheduler when a task can be transmitted in the VC, scheduling it transmits the non-real-time task (The scheduling is achieved using a non-preemptive method.)
7) If there is no real-time traffic data in a real-time channel, all bandwidths will be used as the non-real-time data, in which the non-real-time data can be output as a type of FCFS.

4 Acceptance Control

An acceptance test is applied to configure a new real-time channel in order to transmit multimedia data in wireless ATM networks without any affection to other real time channels. Then, the channel can be assigned if the test is passed. We perform the acceptance test for an uplink and downlink, respectively.

When a number of real-time channels, n, existed in a certain link, the set of the real-time channel can be expressed as {i, i=1,2,...n}. All channels of j have a higher priority than that of the channel i for 1<j<i. The scheduling is only achieved when the finishing time of the transmission of all packets of pj, which minimally have a higher priority than that of pi, is smaller than the transmission time to transmit pi. A single time of polling should be performed to transmit a single packet for an uplink. Thus, an equation for an uplink can be defined by adding the transmission time of C, which is required to transmit a single packet, to the time of C0, as presented in Equ. (4).

$$\sum_{j=1}^{i-1} (C + C_0)\left\lceil \frac{L/p_i}{L/p_i} \right\rceil + 2(C + C_0) \le \frac{L}{p_i} \tag{4}$$

where C is transmission time for a single cell, *L/p*$_i$ is link delay limitation of a cell in the channel *i* .

Equ. (4) has to present a lower value than that of the transmission time of the channel *i* in the transmission time of all packets after polling it to all channels of *j,* which have a high priority, in all of the worst cases.

In the case of the BS, an acceptance test for a downlink can be performed when data is to be transmitted as one-way to MT, or data cannot be transmitted when the polling is achieved, because there are no packets. In this case, the acceptance test can be performed after finishing the acceptance test for the scheduling for an uplink, when the BS reconfigures a channel to transmit data to MT. If a newly connected link of k is a downlink, an equation of the acceptance test for the downlink can be noted as presented in Equ. (5).

$$\sum_{i=1}^{u} (C + C_0)\frac{L}{p_i} + \sum_{i=1}^{d} C\frac{L}{p_i} + C\frac{L}{p_k} \le Nb + \sum_{i=1}^{r} C\frac{L}{p_i} \tag{5}$$

where *Nb* is the entire bandwidth of a network, *u* is connection index to connect an uplink, *d* is connection index to connect a downlink, and *r* is connection index that doesn't transmit packets when a polling is performed. If the amount of bandwidth used in all uplinks and downlinks presents a smaller value than that of the sum of the whole usable amount in a network, in which packets cannot be transmitted when the polling is achieved because there are no packets, the downlink schedule can be satisfied.

$$\sum_{i=1}^{d} C\frac{L}{p_i} + C\frac{L}{p_k} \le Nb \tag{6}$$

In the case of the one-sided downlink from the BS to MT, an equation of the acceptance test can be noted as presented in Equ. (6).

Real-time data guarantees all delay limitations, and transmits non-real-time data using the rest of the usability of bandwidth. The usability is calculated to transmit non-real-time data by calculating the entire usability when a new channel is configured, or removed. Because a RMS scheduling is used to achieve a polling, the calculation focuses on the RMS. The least upper bound to achieve a polling is *ln2*.

$$Nr \le In\ 2 - \sum_{i=1}^{u} (C + C_0)\frac{L}{p_i} + \sum_{i=1}^{d} C\frac{L}{p_i} \tag{7}$$

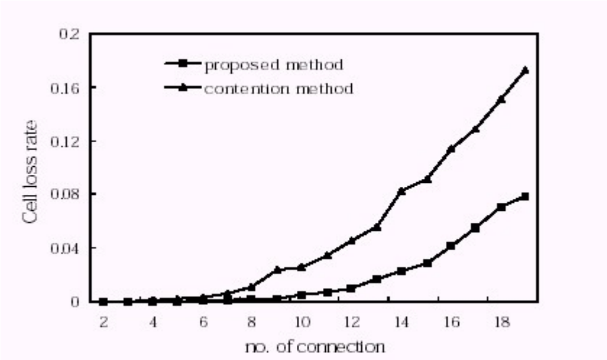


Fig. 4. Cell lose rate

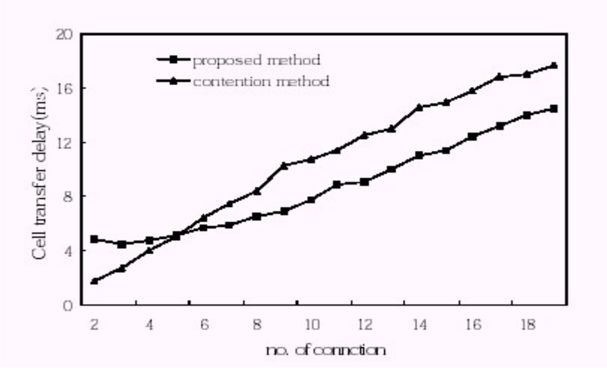


Fig. 5. Cell transfer delay

If the usability of a bandwidth in a network satisfies Equ. (7), non-real-time data can be transmitted while the delay limitation of actual data is satisfied. The arbitrary transmission speed of p_k to transmit this non-real-time data can be obtained using the factor of *Nr.*

For simulation test, packet generation rate of a data transmission signal (cell delay limitation) is 25Mbps, Connected channel is *i, (i=1,2,..20),* packet size of a single packet is *2ms.* The proposed scheme shows cell lose rate and cell transfer delay in Fig 4 and Fig. 5.

5 Conclusions

This paper presents a scheduling algorithm according to the statues of uplink or downlink, which appear as the property of a wireless ATM, by extending a TCRM policy, which is used to transmit real-time multimedia data in the existing wired ATM networks, to wireless ATM networks. In the case of the uplink, it is difficult to recognize whether or not real-time data exists in MT, and then a loss of bandwidth occurs when data is periodically transmitted in this condition. Thus, the TCRM policy was applied after polling. This paper achieved the polling based on a RMS scheduling, and a channel, which is passed through the given acceptance test, guarantees the delay limitation of real-time data to transmit real-time multimedia data. However, it has the disadvantage that the entire bandwidth is to be configured by 69% after considering a worst case, in which the least upper bound of RMS is 69%. In the case of the downlink, data transmission is performed at one-side from the BS to MT, and then a TCRM policy is directly used.

References

1. WATM workinggroup:Baseline Text for Wireless ATM specifications. Montreal, Quebec, Jul.(1997)
2. Mahmoud Naghshineh, Anthony S. Acampora,:QoS Provisioning in Micro-cellular Networks Supporting Multimedia Traffic. IEEE INFORCOM (1995)
3. Kweon, S. K. and Shin, K. G.: Traffic-controlled Rate-monotonic Priority Scheduling of ATM Cells. In Proceeding of the 15th IEEE INFORCOM Mar (1996)
4. Liu, C.L. and Layland, J.W.: Scheduling Algorithms for Multiprogramming in a Hard Real Time Environment. J.ACM 20(1) (1973) 46-61
5. Geert, J. Heijenk, Xinli Hou, and Ignas G.Niemegeers,: Communication Systems Supporting Multimedia Multi-user Applications. IEEE network, January/Febuary (1994) 33-44
6. Qsama Kubbar and Hussein T. Mouftah,: Multiple Access Control Protocols for Wireless ATM : Problems Definition and Design Objectives. IEEE Communications Magazine, vol. 35, Nov. (1997) 93-99
7. Kweon, S. K. and Shin, K. G. : Real-Time Transport of MPEG-Video with a Statistically Guarnateed Loss Ratio in ATM Networks. IEEE trans, parallel and distributed systems, vol 12. no.4 Apr. (2001)

8. Knightly, E., Wrege, D., Liebeherr, J., and Zhang, H. :Fundamental Limits and Tradeoffs of Providing Deterministic Guarantees to VBR Video Traffic. In Proc. of ACM SIGMETRICS (1995) 98-107
9. Turner, J. S.: New Directions in communications (or Which Way to the Information Age?). IEEE communications Magazine, Vol. 25, No. 8, October (1996) 8-15
10. Kandlur, D. D., Shin, K. G., and Ferrari, D.: Real-time Communication in Multi-hop Networks. In Proc. 11-th Int'l conf. Distributed Comput. Systems, May (1991) 300-307

Road Change Detection Algorithms in Remote Sensing Environment

Hong-Gyoo Sohn[1], Gi-Hong Kim[2], and Joon Heo[1]

[1] School of Civil and Environmental Eng., Yonsei Univ., Seoul, Korea
{sohn1, jheo}@yonsei.ac.kr
[2] Department of Civil Eng., Kangnung National Univ., Gangneung, Korea
ghkim@kangnung.ac.kr

Abstract. This paper describes an automatic change detection of roads using aerial photos and digital maps. The task is based on the idea that one can derive information about the changes strictly from its imagery once the geometric relationship among data sets is correctly recovered. The goal of research is achieved by using the Modified Iterated Hough Transform (MIHT) algorithm, the result of which not only solves the orientation parameters of the aerial camera but also filters out blunders from all possible combination of their entities. To examine the effectiveness of the MIHT algorithm, a digital road map and an aerial photo are used to detect changes. Experimental results demonstrate the potential of the MIHT algorithm for detecting changes of the Geospatial Information System (GIS) data.

1 Introduction

Recent urbanization being accelerated at rapid rate has made spatial information of urban features more complicated to interpret and detect its changes. Monitoring of these changes is one of the most argumentative issues nowadays in GIS. To update the GIS data from imagery it is necessary to establish the geometric relationship between them. Once this relationship is correctly recovered, one can then derive information strictly from its imagery. Orientation procedure, such as resection, is the prerequisite for establishing relationship. It is conducted with the given corresponding entities of two data sets [8].

The problem is that the relationship is difficult to be determined when the corresponding entities are not known. Some attempts, so-called matching techniques, have been made to identify and measure the corresponding points in two data sets automatically. In image to image case, there are some matching techniques established quite well such as area-based and feature-based matching. In area-based matching, the similarities of two images, especially gray levels, are derived either through the cross-correlation [6] or least-squares approach [1]. On the other hand, feature-based matching [4] is based on the extraction of image features, for example the shape, sign and strength of edges, through an interest operator including F五rstner operator [3], Moravec operator [9], Canny edge detection operator [2] to select point-like or other well-defined features. Habib and Kelly proposed a new statistical approach named MIHT not only to detect the matching entities but also to simultaneously solve orien-

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 821 – 830, 2005.

tation parameters of imagery using linear features [5]. This is a robust technique to estimate parameters when the corresponding entities are not identified but linear features and mathematical function of two data sets are given.

The main goal of this paper is to automate the task of change detection using aerial photos and digital maps. We investigate the potential of automatic change detection algorithms based on the MIHT and apply this algorithm in domestic GIS data which are 1:5000 aerial photos and 1:1000 digital maps.

2 Modified Iterative Hough Transform

Hough transform is a method used to estimate parameters by way of a voting scheme [7]. The basic principle of this approach is to switch the roles of parameters and spatial variables. This technique is particularly useful for computing description of a feature given local measurement. With some modification, the Hough transform can be used to estimate the parameters of a mathematical model relating entities of two data sets assuming no knowledge of corresponding entities and no established correspondence. This is derived from the idea of the Hough transform that the peak location of an accumulator array, in which solutions are located for all candidates, is the most possible location that includes the correct answer.

The MIHT algorithm assumes that there is no knowledge of corresponding entities and correspondence but the relation of two data sets defined by a mathematical model. This mathematical model yields an observation equation. Because corresponding entities are not known, the evaluation begins with all possible matches of data sets.

The parameters of the mathematical model can be estimated simultaneously or sequentially. If the parameters are simultaneously solved, however, It has to be carefully considered that the number of parameters determines the dimension of the accumulator array of the Hough transform. For example, if there are i entities in data set one and j entities in data set two, solving n parameters simultaneously would generate $\frac{i \cdot j!}{(i \cdot j - n)!n!}$ combinations, leading to combination explosion. In addition, more than two-dimensional accumulator array not only decreases the speed of the computation but also creates the memory problem of the computer.

To overcome this problem, each parameter is solved sequentially in an iterative manner, updating the approximations at each step. Consequently, the accumulator array can be manipulated in one or two dimensional space and then generates only $i \cdot j$ combinations of entities reducing the computational complexity. On top of that, the parameters can be estimated with high accuracy.

The convergence rate towards the correct parameters depends on the independency of the parameters and the non-linearity of the transformation function. The effect of the non-linearity is similar to a least squares adjustment of non-linear model. Highly non-linear model converges more slowly requiring more iteration. On the other hand, the independency of the parameters is more crucial by the reason that parameters are estimated sequentially. Therefore, the order of the parameter estimation has to be determined in a manner to minimize the influence of the parameters adjusted previously to the next parameters to be adjusted.

2.1 Mathematical Model

The first step to detect changes occurred in two data sets is to determine the geometric relationship. Single photo resection (SPR) is a photogrammetric technique commonly used to estimate the Exterior Orientation Parameters (EOPs), which establish both the position (X_0,Y_0,Z_0) and the rotation (ω,φ,κ) of an image with respect to the object space coordinates system, so that the geometric relationship between image and object space data is defined. In SPR, the collinearity model (see Equ. (1)) is used to relate points in the image with corresponding points in the object space, and the relation is expressed as a function of the EOP.

$$\begin{bmatrix} x-x_0 \\ y-y_0 \\ -f \end{bmatrix} = \lambda M(\omega,\varphi,\kappa) \begin{bmatrix} X-X_0 \\ Y-Y_0 \\ Z-Z_0 \end{bmatrix} \tag{1}$$

Traditionally, the parameters are estimated by way of a least squares adjustment involving measured control points in the image. At least three control points are required to estimate the six EOPs. More than three points increase the redundancy and strengthen the solution of the parameters.

2.2 Observation Equations of SPR Using MIHT

To estimate EOPs of SPR the observation equation of the collinearity model has to be composed of each parameter group. Equation (1) can be rewritten as

$$\begin{aligned} x &= x_0 + f\frac{m_{11}(X-X_0)+m_{12}(Y-Y_0)+m_{13}(Z-Z_0)}{m_{31}(X-X_0)+m_{32}(Y-Y_0)+m_{33}(Z-Z_0)} \\ y &= y_0 + f\frac{m_{21}(X-X_0)+m_{22}(Y-Y_0)+m_{23}(Z-Z_0)}{m_{31}(X-X_0)+m_{32}(Y-Y_0)+m_{33}(Z-Z_0)} \end{aligned} \tag{2}$$

For applications with more unknown variables to be solved for from a series of minimally sufficient or redundant observations, this non-linearity of the equation is expanded using Taylor series approximations. Equ. (2) can then be represented as

$$\begin{aligned} F_x &= (x-x_0) - f\frac{U}{W} = 0 \\ F_y &= (y-y_0) - f\frac{V}{W} = 0 \end{aligned} \tag{3}$$

where

$$U = m_{11}(X-X_0)+m_{12}(Y-Y_0)+m_{13}(Z-Z_0) \tag{4}$$

$$V = m_{21}(X-X_0)+m_{22}(Y-Y_0)+m_{23}(Z-Z_0) \tag{5}$$

$$W = m_{31}(X-X_0)+m_{32}(Y-Y_0)+m_{33}(Z-Z_0) \tag{6}$$

In Equ. (3), the photo coordinates x and y are considered as the observations, the elements of interior orientation x_0, y_0, and f are considered to be known value from calibration data, and the variables are considered as unknown parameters. Consequently, the linearized form of Equ. (3) is given by

$$v + B\Delta = F \tag{7}$$

where v is image coordinates residuals $(= [v_x, v_y]^T)$, B is the matrix of partial derivatives of the two functions of Equ. (3) with respect to each of the six exterior orientation elements and the three coordinates of the object point. Δ is the vector of the approximations for each parameter, and F is given by

$$F = \begin{bmatrix} -F_x^0 \\ -F_y^0 \end{bmatrix} = \begin{bmatrix} -(x - x_0) + f \cdot U / W \\ -(y - y_0) + f \cdot V / W \end{bmatrix} \tag{8}$$

The elements of matrices B, Δ, and F are consisted depending on the parameters to be solved as follows:

Observation Equation for (dX_0, dY_0)

$$\begin{bmatrix} \frac{\partial F_x}{\partial X_0} & \frac{\partial F_x}{\partial Y_0} \\ \frac{\partial F_y}{\partial X_0} & \frac{\partial F_y}{\partial Y_0} \end{bmatrix} \begin{bmatrix} dX_0 \\ dY_0 \end{bmatrix} = \begin{bmatrix} -F_x^0 \\ -F_y^0 \end{bmatrix} \tag{9}$$

where, F^0 is a function of $(X_0^0, Y_0^0, Z_0^0, \omega_0^0, \varphi_0^0, \kappa_0^0)$ and the partial derivatives of matrix B are

$$\frac{\partial F_x}{\partial X_0} = \frac{f}{W^2}(m_{11}W - m_{31}U) \tag{10}$$

$$\frac{\partial F_x}{\partial Y_0} = \frac{f}{W^2}(m_{12}W - m_{32}U) \tag{11}$$

$$\frac{\partial F_y}{\partial X_0} = \frac{f}{W^2}(m_{21}W - m_{31}V) \tag{12}$$

$$\frac{\partial F_y}{\partial Y_0} = \frac{f}{W^2}(m_{22}W - m_{32}V) \tag{13}$$

Observation Equation for $(d\kappa)$

$$\begin{bmatrix} \frac{\partial F_x}{\partial \kappa} \\ \frac{\partial F_y}{\partial \kappa} \end{bmatrix} (d\kappa) = \begin{bmatrix} -F_x^0 \\ -F_y^0 \end{bmatrix} \tag{14}$$

where *F⁰* is a function of $(X_0^{new}, Y_0^{new}, Z_0^0, \omega^0, \varphi^0, \kappa^0)$ and the partial derivatives of matrix B are

$$\frac{\partial F_x}{\partial \kappa} = -\frac{f}{W} V \tag{15}$$

$$\frac{\partial F_x}{\partial \kappa} = \frac{f}{W} U \tag{16}$$

Observation Equation for (dZ_0)

$$\begin{bmatrix} \dfrac{\partial F_x}{\partial Z_0} \\ \dfrac{\partial F_y}{\partial Z_0} \end{bmatrix} (dZ_0) = \begin{bmatrix} -F_x^0 \\ -F_y^0 \end{bmatrix} \tag{17}$$

where *F⁰* is a function of $(X_0^{new}, Y_0^{new}, Z_0^0, \omega^0, \varphi^0, \kappa^{new})$ and the partial derivatives of matrix *B* are

$$\frac{\partial F_x}{\partial Z_0} = \frac{f}{W^2}(m_{13}W - m_{33}U) \tag{18}$$

$$\frac{\partial F_y}{\partial Z_0} = \frac{f}{W^2}(m_{23}W - m_{33}V) \tag{19}$$

Observation Equation for $(d\omega, d\phi)$

$$\begin{bmatrix} \dfrac{\partial F_x}{\partial \omega} & \dfrac{\partial F_x}{\partial \phi} \\ \dfrac{\partial F_y}{\partial \omega} & \dfrac{\partial F_y}{\partial \phi} \end{bmatrix} \begin{bmatrix} d\omega \\ d\varphi \end{bmatrix} = \begin{bmatrix} -F_x^0 \\ -F_y^0 \end{bmatrix} \tag{20}$$

where *F⁰* is a function of $(X_0^{new}, Y_0^{new}, Z^{new}, \omega^0, \varphi^0, \kappa^{new})$ and the partial derivatives of matrix *B* are

$$\frac{\partial F_x}{\partial \omega} = \frac{f}{W^2}\left[(m_{13}\Delta Y - m_{12}\Delta Z)W - (m_{33}\Delta Y - m_{32}\Delta Z)U\right] \tag{21}$$

$$\frac{\partial F_y}{\partial \omega} = \frac{f}{W^2}\left[(m_{23}\Delta Y - m_{22}\Delta Z)W - (m_{33}\Delta Y - m_{32}\Delta Z)V\right] \tag{22}$$

$$\begin{aligned} \frac{\partial F_x}{\partial \varphi} = \frac{f}{W^2}[(&\Delta X \sin\varphi \cdot \cos\kappa - \Delta Y \sin\omega \cdot \cos\varphi \cdot \cos\kappa + \Delta Z \cos\omega \cdot \cos\varphi \cdot \cos\kappa)W \\ &+ (\Delta X \cos\varphi + \Delta Y \sin\omega \cdot \sin\varphi - \Delta Z \cos\omega \cdot \sin\varphi)U] \end{aligned} \tag{23}$$

$$\frac{\partial F_y}{\partial \varphi} = -\frac{f}{W^2}[(\Delta X \sin\varphi \cdot \sin\kappa - \Delta Y \sin\omega \cdot \cos\varphi \cdot \sin\kappa + \Delta Z \cos\omega \cdot \cos\varphi \cdot \sin\kappa)W + (\Delta X \cos\varphi + \Delta Y \sin\omega \cdot \sin\varphi - \Delta Z \cos\omega \cdot \sin\varphi)V] \tag{24}$$

where $\Delta X_i = X_i - X_0$, $\Delta Y_i = Y_i - Y_0$, $\Delta Z_i = Z_i - Z_0$.

3 Data Sets

To detect and update the changes of road data, following information is needed: (1) A sequential of 3D points along the ground features, (2) A sequential of 2D points along the image features, (3) The interior orientation parameters of the camera.

For a sequential of 3D points along the ground features, digital map distributed in January 1999 (see Fig. 1), which is expected to be generated from an aerial photo acquired previously, is used as data set to be updated. For a sequential of 2D points along the image features, aerial photo acquired in May 1999 is used as the reference data (see Fig. 2 and Table 1).

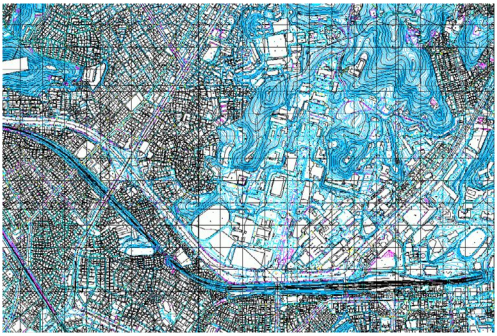


Fig. 1. Digital map of the study area

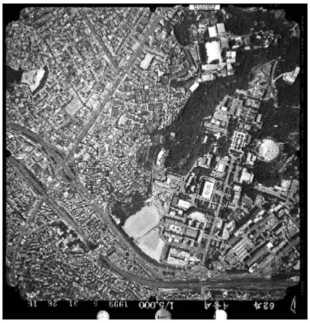


Fig. 2. Aerial photo of the study area

Table 1. Information of aerial photo

Aerial camera	RC-30
Focal length	152.85 mm
Photo scale	1:5,000
Scanning resolution	32×32 μm^2
Image size : width	11732 pixels
Image size : height	11240 pixels

These data were selected because some changes were expected to be occurred by the frequent redrawing of the road middle lines. Scale of data sets is 1:1,000 and 1:5,000 each. Interior orientation parameters of the camera are obtained from manufacturer's calibration certificates.

Fig. 3 shows the center lines of the roads, which are especially classified as a major road longer than 4 *m*, and 309 vertices are extracted. However, the road data of digital maps are usually given as 2D points. In this study, Z values of the road vertices are estimated using Digital Elevation Model (DEM) interpolated by 1 *m* using the kriging method (see Fig. 4).

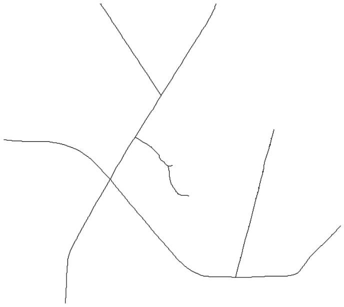


Fig. 3. Major road center lines

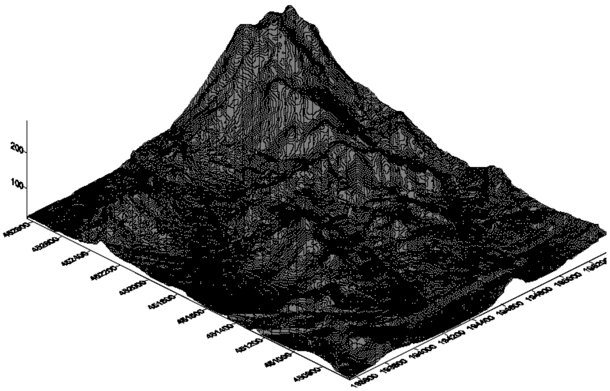


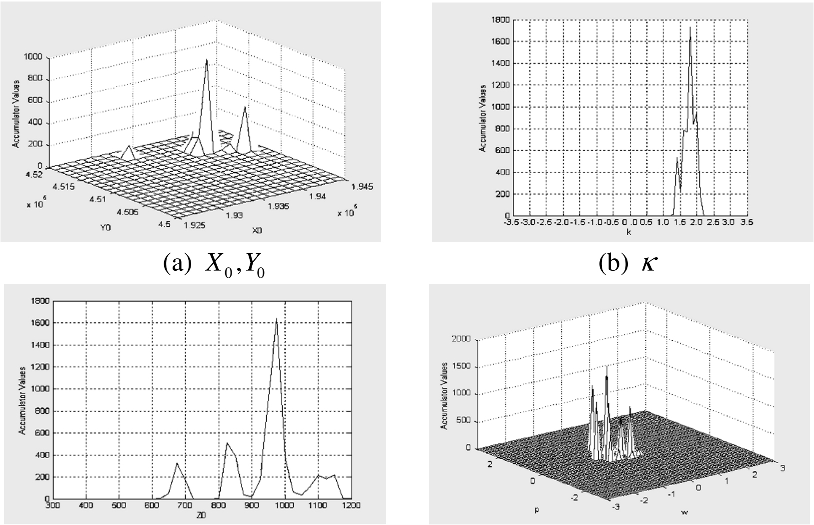
Fig. 4. DEM of the study area

On the other hand, a 2D point sequence along the image road network must be extracted. In a digital environment, the extraction process can be established by applying a dedicated operator (e.g., a Canny or any other operator for road network extraction). However, the performance of those operators has not yet qualified to extract the complete road body and be applied to the MIHT algorithm. In this work, therefore, 2D image features have been digitized manually and 1656 pixels of the digitized road were extracted.

4 Experiments

Before the SPR to solve EOPs in digital environment, points in the image coordinates system were relocated into the photo coordinates system. This relationship is defined by the affine transform using the given calibration data.

The experiment began with the all possible combinations of image and ground space data along corresponding regions. To improve the speed of the operation, the cell of the next iteration was rebuilt to include only the entities existing in the cell of the approximation. Fig. 5 shows accumulator arrays and their peaks.



(a) X_0, Y_0 (b) κ

(c) Z_0 (d) ω, φ

Fig. 5. Accumulator array and their peaks

In order to exclude the blunders that could be generated by errors such as digitizing errors and errors of digital map, more than three consecutive points in image space detected having no corresponding entities are regarded as data to be updated. As a result, 34 pairs of points were detected to be updated (see Fig. 6 and Table 2).

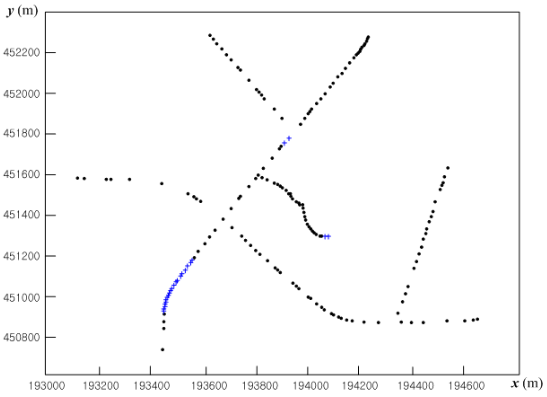


Fig. 6. Results of the road center line changes (blue cross marks: to be updated)

Table 2. Final result of the change detection using the MIHT

Total number of image points	1656
Total number of object points	309
The number of matched points	253
The number of points to be updated	34
The number of blunders	22
Percentage of matched points	81.88 %
Percentage of detected change	11.00 %

After detecting the matching entities, EOPs to determine the geometric relationship between the image and ground were simultaneously estimated. The EOPs are listed in Table 3.

Table 3. Result of the estimated EOPs change detection using the MIHT

X_0 (m)	Y_0 (m)	Z_0 (m)	ω (radian)	φ (radian)	κ (radian)
194098.05	451441.44	958.51	-0.0069	-0.0009	1.5983

5 Conclusions

The results of experiments demonstrate that the road center line change information of the digital map can be successfully detected from the aerial photos. The MIHT algorithm has detected 11% of the whole features as entities to be updated. These changes may be real changes in road network or mis-drawing road center lines in digital maps.

The resolution of the scanned images and the accuracy of the edge detection operator, however, block the exact extraction of linear feature information. If the extraction of linear features will be completely automated in the near future, better result of automatic change detection is expected.

References

1. Ackermanm, F.: Digital Image Correlation, Performance and Potential Application in Photogrammetry. Photogrammetria 11(64) (1984) 429-439
2. Canny, J.: A Computation Approach to Edge Detection. IEEE Transaction on Pattern Analysis and Intelligence 8(6) (1986) 679-697
3. F五rstner, W. and Gulch, E.: A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centers of Circular Features. Proceedings ISPRS Inter-commission Conference on Fast Processing of Photogrammetric Data (1987) 281-305
4. Grimson, W.E.L.: Computational Experiments with a Feature-based Stereo Algorithm. IEEE Transactions on Pattern Recognition and Machine Intelligence 7(1) (1985) 17-43
5. Habib, A. and Kelly, D.: Single Photo Resection using the Modified Hough Transform. Photogrammetric Engineering & Remote Sensing 67(8) (2001) 29-41
6. Hannah, M.J.: Digital Stereo Image Matching Techniques. International Archives of Photogrammetry and Remote Sensing, 27(B/3) (1988) 280-293
7. Hough, P.V.C.: Methods and Means for Recognizing Complex Patterns. U.S.Patent 3,069,654 (1962)
8. Mikhail, E.M. and Bethel, J.S.: Introduction to Modern Photogrammetry. John Wiley&Sons, Inc (2001)
9. Moravec, H.P.: Towards Automatic Visual Obstacle Avoidance. Proceedings 5th International Joint Conference on Artificial Intelligence (1977)

Safe RFID System Modeling Using Shared Key Pool in Ubiquitous Environments

Jinmook Kim and Hwangbin Ryou

Dept. of Computer Science, Kwangwoon Univ., Seoul, Korea
{jmkim, ryou}@netlab.kw.ac.kr

Abstract. In a Ubiquitous environment many individual devices which have the communication ability with the computation ability are connected with one another. For this change we must be preceded study of Radio Frequency Identification (RFID) afterward research of a sensor network. And we must grasp the threat to protect the RFID system. In this paper we propose safe RFID system modeling based on shared key pool in ubiquitous environment. This proposed modeling is secured to RFID system from many security threat and privacy problem, and it has shorter access delay time. We expect this proposed modeling to be used in ubiquitous environment for the RFID system being safe.

1 Introduction

The ubiquitous environment is changing to a form different from existing a computer communication environment. We can operate as the object and all things is Self-Propelled of the communication at a ubiquitous environment. Therefore the device should gather information by itself and analyze it and process it in the ubiquitous environment, while the user should study how to use the device and delivery it to the third party in the existing computing environment.

The ubiquitous computing comes into our lives and makes us live more convenience life with the help of devices. To support more convenience life whenever and wherever, the sensor, processor, communication and interface are necessary. Also, the security technology is the essential for the ubiquitous computing environment.

The fragility of the ubiquitous computing is more serious than the existing computing network, because temporal and dynamic equipments are interoperating on the network. Therefore the security technologies that are used for the current network environment are not applicable for the ubiquitous. In addition, not only fragility of the security, but also privacy will be more serious issue. In the best ubiquitous environment, user information is gathered without user's recognition and treat it as many kinds of type. During this process, the information of user privacy can be shown to the outside [9-11].

To protect the personal information and resolve the privacy issue, the information protection technology and privacy protection technology that are applicable to the ubiquitous environment are urgent. However, the equipments that uses in ubiquitous environment have the special features, such as low power, low price, small size, mobility and inter-combination [12,15,17,18]. So the existing network security services

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 831 – 840, 2005.

that is applied the public key based security algorithm are not applicable. It is essential to know how the ubiquitous-applicable security technology and privacy protection technology are improving and how to apply them in the ubiquitous environment. To apply the ubiquitous computing to the reality, the sensor network based on RFID as user interface technology should be studied.

In this paper we use RFID as the information transmission and object recognition measurement in ubiquitous environment, and look into the information protection problem and privacy issues during the process of using RFID.

2 Related Works

In RFID environment, radio frequency makes the communication between tag and reader. In this environment, two objects are communicating each other with using international standard frequency. So, it is not difficult to wiretap the communication or forge it for the ill-intentioned user. In addition, it could be used for tracing the proper user's location and steal the identity by using the reader. It necessarily needs that we should grasp the threatening factors for the mobile communication of RFID in advance, and prepare the solution for the emerging ubiquitous environment.

RFID system is composed of the tag that putting a information, reader that gather a information or transferring and back-end system. To protect the communication environment in RFID from these threatening factors, we should analyze the requirements for the security, and use the related technologies properly from the preceded researches for the information protection and privacy issue [13,14, 16].

In Kill command approach, it is the technique that is suggested firstly for information protection in RFID. It sends Kill command when the purpose of using the tag is vanished, so tag information becomes unable to send or receive. In Hash-Lock techniques, it makes tag Lock state after it records the hash function value created randomly from the reader. The Lock state means that the tag responses the only reader that sent the exact hash function value, so that reader can handle the ID value of tag. It can protect the forgery of tag information, but cannot protect privacy problem.

In Random Hash-Lock techniques, it is the technique that gives the random attributes to the reader-create hash function value for resolving the privacy threatening at the Hash-Lock technique. It can resolve the part of threatening privacy problem, but it has much overhead to calculate the hash function value, so the service rejection attack could be easier, because the reader becomes busier. In One-time Pad technique based on XOR operations, it is the early proposed cryptography that is used for the communication between the tag and reader. It encrypts the information in the tag using the key value that has One-Time Pad characteristic with XOR operation, so it can protect the stored information in the tags.

In Re-Cryptography techniques, it is the technique that is proposed for the tracing of the large amount currency in Europe. It encrypts the currency's ID information by the public key. In Blocker-Tag techniques, it supports to set up the protection area and ordinary area for rejecting the other reader's access as the user's necessary in the communication between the tag and reader. Upon the request, we can make the tag not access from the other reader, and if user changes it to the ordinary area mode, all the other reader can use the tag. However, it has the disadvantage that user should remember that long and difficult information to set up the tag as another area mode.

For Privacy problems, Group signature techniques proposed the member to sign it as representative of the group and nobody knows who signed it. However, any trouble is made, the group manager can figure out the signed member. Blind signature techniques proposed for the electronic currency as the basic idea of the privacy. It's based on RSA. In this technique, when the user signs the message the user cannot know what the message is. Distributed cryptography techniques proposed threshold cryptography that it divides the information of secret key to all of managers for protecting the information of shared secret key. So, all of manager have to get together for acquiring the information of secret key.

3 Proposed Safe RFID Systems

3.1 System Architectures

We consider how to prevent RFID security problems, so we will propose safe RFID system, and design it. First of all, the proposed safe RFID system shows how to prevent the forgery of tag information. Additionally, we suggest the privacy protection technique that prevents stealing user's identity.

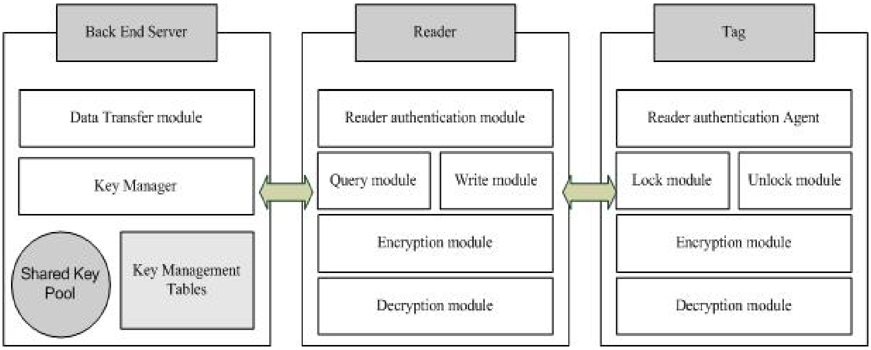


Fig. 1. Safe RFID system architectures

The safe RFID system based on Shared Key Pool has three parts [1-8]. Firstly, the back-end server has the modules that creates keeps and allocates the Shared Key. The second part is the RFID reader that has the module for reading and storing the tag information. It’s the basic function of reader, and the reader in this paper has the encrypting and decrypting modules that work between the reader and tag. We will assume that the last part has the functionality like semi-passive tag for the future environment. The existing passive tag had many restrictions for the low price in the real world. Therefore, we will use the semi-passive tag to overcome it. It will resolve the security problem and privacy with the improved operation ability and low storing power.

3.1.1 Back-End Server

First of all, back-end server contains random number creator, Shared Key Pool, Key Manager, Key Management Table. The random number creator creates many random numbers, and stores at the Shared Key Pool, and transfers created Shared Key to the reader upon reader's request. Key Manager transfers Shared Key to the reader, and records the information of requested reader and allocated key at the Key Management Table.

The function of Shared Key Pool which constitutes Back-end server is generating secret key that is used at first for tag to communicate with readers. A number of secret keys more than number of tags which must be managed by Back-end server are generated periodically and store in pool. And by request readers, shared keys are allocated to readers and tags, Key Manager module performs managing this. In this process, for managing generated key such as allocation, register, modification, and delete Key Management Tables are remained.

3.1.2 Reader

For the Reader, it had the querying module and writing module for information, and it contains authentication module for Shared Key of reader's group and encryption module for the tag information. It encrypts tag ID by using the Shared Key that came from the Back-end server and locks it. During the encryption, It use the symmetric algorithm (i.e., RC5) to encrypt tag ID. It can reduce the load of system and improve the speed of operation by this algorithm. It encrypts the tag ID, and transfers it. To prevent the illegal forgery, it makes the tag lock state. Also it transfers hash value of reader ID by using Shared Key to check the reader, so it assurance the confidentiality.

In safe RFID system proposed in this paper, reader which is core of communication and authentication of reader which participated in communication before performing communication must be preceded. For this, each reader takes reader authentication module for authenticating itself. This module which performs authentication for readers transfers packets that consist of combination with user information and product information about reader to back-end server.

After a server receives authentication request information from readers, the server searches tables, identifies authentication of reader and transfers tags and a secret key for communication to identified readers.

Authenticated reader from the server transfers part of secret key which is allocated by server for communication with tag and communication request information including own authentication information to reader. It performs in Query module.

When tags identify authentication of reader and are ready for communication, in earnest communication, using encryption module and decryption module information which is ready for communication between readers and tags is encrypted and decrypted and then tags perform communication. For this encryption and decryption communication this paper compares to streaming method of RC5 encryption algorithm which concerned with communication environment of tags as fundamental approach and block chain method. This is for providing flexible approach later by providing various encryptions and decryptions as communication protocol between readers and tags.

3.1.3 Tag

We assume that we will use the semi-passive tag that does not have the limitation of the small operation capability and battery in the existing passive tag. The semi-passive tag that will be used in this paper has the 8-bit operation capability and the same operation power with passive tag, but it has more space to store information that the passive tag. It has the extended storing power that can store not only EPC code, product name and basic information, but also the encrypted tag ID. This algorithm is easy to implement the decrypting operation in the tag, and has the same speed for encrypting and decrypting.

Tag consists of EPC code in distinction from actual object, encryption module which implements encryption algorithm for secure communication and decryption module. At this time, transferring securely encrypted information to only authenticated readers is function of encryption module and decryption module. Decryption module decrypts received information from authenticated reader. A secret key which is initial information of performing encryption and decryption and information which is part of EPC code which is system information about tags are saved in FLASH memory. And wiretapping communication between readers and tags can be occurred so even though communication between authenticated readers and tags is performed, transferring information transfers encrypted data. If somebody wired the communication, it would be difficult to figure the contents of communication out and when decrypted, it is rubbish information.

Secret key is necessary information for authentication and initial connection. Tag ID is corresponding to EPC code. So using secret key and tag ID, user individual and position information don’t get it.

If unauthenticated reader requests communication, it must confirm whether it is authenticated or not, if not, in the state of locking tags, replying temporal information corresponding to EPC code which is generated by using secret key, objects can distinguished but it had making lock and unlock state module for not taking information such as position and user individual information.

3.2 Procedure of Safe RFID System

Operation of safe RFID system proposed is shown. There are two parts of operation. First part is storing tag ID corresponding to EPC code and secret key which is used encryption and decryption between reader and tag. It is called pre-distribution. Back-end server transfers secret key and tag ID that is corresponded EPC code by pre-distribution protocol. Tags receive tag ID and secret key, store and then are locking state. Second part is that readers request tags to communication and then tag response information of serve to reader. Readers request authentication of it. The back-end server checks authentication of reader. If readers are authenticated, tags transfer tag ID, if not, it ignores.

In system proposed, in case many readers request tags it works properly on the other hand when many tags respond to one reader, it will occur collision problem. In this paper, collision problem is not solved. It is not mentioned in this paper.

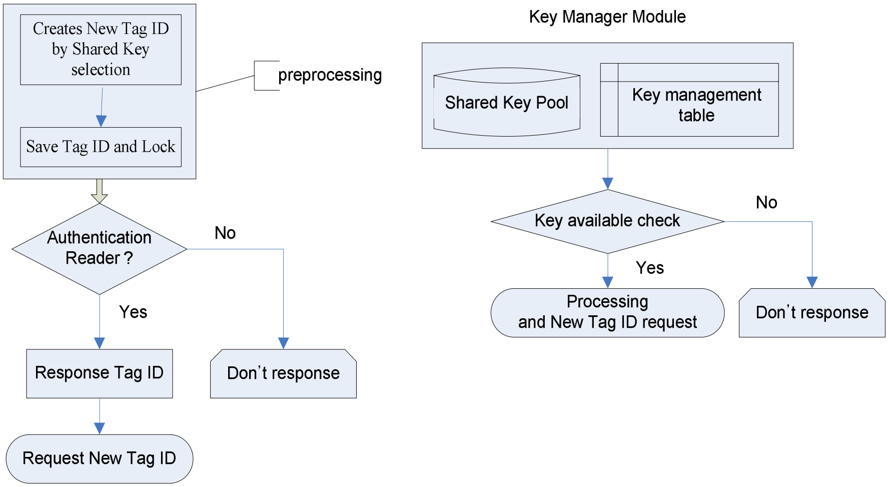


Fig. 2. Procedure of safe RFID system

The figure 2 shows that the procedure of the proposed safe RFID system based on Shared Key Pool. Allocation of secret key and tag ID between tag and reader will consider. The procedures are followed:

- The back-end server choose the secret key in the Shared Key Pool
- The back-end server transfer tag ID that is encrypted by selected Shared Key.
- The Tag store the encrypted tag ID and lock it to prevent the forgery from the other readers.

It is pre-processing about allocating of secret key and tag ID on back-end server by pre-distribu tion protocol. When readers request for communicating to tag, tags respond back-end server information. Back-end server performs authentication of reader. If readers are authenticated, tags transfer tag ID and tag information that encrypted by secret key. Otherwise tags are performed following procedures:

- Store the encrypted tag ID that is come from the right reader, and query the rightness of reader upon the reader's reading request on lock state.
- Check the Shared Key that is sent by reader, and send the encrypted tag ID. If Shared Key was not right, do not send the tag ID.
- Transfer stored tag ID to the reader, and release the lock state. And request the reader send the newly-encrypt tag ID by the new Shared Key.
- Receive the newly-create encrypted tag information, and store it, and becomes lock state again.

Finally, the reader requests back-end server send the encrypted tag ID by using new Shared Key upon the tag's request, and the procedure is finished when the reader receives it.

4 Performance Analysis

In experiment environment, it is difficult to prove the safety of this system. Therefore, we will prove that there is no lowering of performance when we apply our algorithm to RFID system, so that our system is applicable for RFID system.

We assume that the back-end server and reader communicate on wired environment, because our purpose of this experiment is providing the safe RFID environment. The wireless part is just the connection between reader and tag, so we will test only the delay time at the tag. In the realistic environment, the communication time between the tag and the reader is existed, but we will not consider this time, because it's not a major consideration. We will test gcc 2.7.2.1 compiler on ARM7T that has 8bit processor and 512Mbit. It is similar with the safe RFID system. Also we will omit some of parameters which we cannot measure, because the maximum timer time that ARM7T supports is near 130 seconds. We think the experiment result has the small gap with the real RFID system, but we can predict the safety of future RFID system by watching the result under the similar environment.

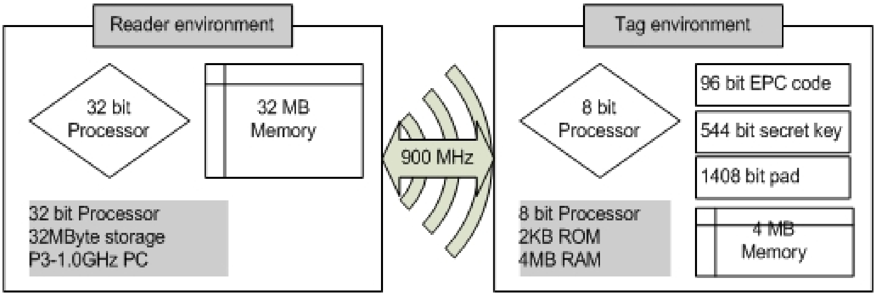


Fig. 3. Experiment environment

Experiment environment of safe RFID system proposed in this paper consists of two parts as shown in figure 3. This system operates on wireless environment.

First part, that is tags, is distinguished with general tags that are called passive tags. This system will use semi-passive tag which is proposed. It is difficult to set specification of semi-passive tag. However, passive tag cannot provide security service so semi-passive tag is proposed.

Semi-passive tag proposed is analogous to EPC Code version 1.2 which is standard of RFID. Processor in semi-passive tag will be processing power of node in sensor network. Therefore processors are expected to have 8-bits processing power. Tag and reader perform encryption and description. It needs data saving space whose size is considered in two ways.

First, we consider ROM (FLASH) size because ROM store secret key and EPC Code. Second, we consider RAM size because RAM store encryption algorithm (RC5) and temporal data.

ROM stores EPC code which is 96 bits, secret key which is 8 of 56 bits (544 bits) and encryption algorithm which is 1KB. Therefore size of ROM is bigger than 2KB so that proposed safe RFID system suggests that ROM size is 2KB. This size of ROM needs low-cost.

RAM has same size of general passive tag. Therefore RAM size suggests 4MB which is the same as general node size of sensor network. We will be modeling about readers which encrypt, decrypt and transfer information at PC so that modeled readers are similar to ARM7T or Pentium 3 – 1.0GHz. In this paper we will consider Pentium 3 - 1.0GHz.

We proposed experiment environment. In this environment confidentiality service can be provided as well-known by using RC5 security algorithm. Therefore we will observe our delay of execution time. In result, we can figure out capability of safe RFID system.

We compared execution time of safe RFID system and access time of RC5 in PC. This experiment calculates average value about 10 times of encryption and decryption. The domination is micro sec.

Table 1. Execution time of processing (micro-sec)

Environment	Encryption(RC5)	Decryption(RC5)
PC Only	482	482
Safe RFID system	780	850

Table 1 shows PC Only execution time and execution time of safe RFID system. On safe RFID system proposed, there is delay time of communication between tag and reader so that above result is available. Table 1 shows simple data encryption and decryption execution time about 64 bit data block so that we consider execution time in more complex system. And we consider communication with many reader and tag. In this case, many problem appear and complexity of tag and reader.

First line of table 1 indicates encrypt and decrypt time using RC5 security algorithm and second line indicates encrypt and decrypt time when RC5 security algorithm is performed in proposed system. As table 1 indicates the result, it is not quite difference between performing RC5 security algorithm in PC and performing RC5 security one in proposed system. This proves capability of proposed system.

However the experiment environment which we proposed is simple. In other words we just compare the execution time of performing RC5 algorithm. But the environment which RFID system will be operated in is very complex and has a number of considerable problems. The real RFID system environment wants more tags and readers that will communicate. Each readers and tags also will attempt to communicate. They have low rate of communication. As you know, general RFID system sometimes has less than 10% of success rate.

We overlook lack of connection between tag and reader and collision problem. But these problems do not mentioned in this paper.

5 Conclusions

Ubiquitous computing has been developing emerging. Like this security issue concerned with ubiquitous computing is more risky than traditional computing environment. But there are not solutions for security problems.

This paper suggests for safe RFID system which provides confidentiality services between readers and tags and authentication service and solves user privacy problem and user position privacy.

It represents modeling of safe RFID system and is simulating about secure system. For this purpose, modeling shows confidentiality and authentication. Confidentiality service ensures that secure communication between tags and readers. Authentication is ensured by secret key, tag ID and agent module. Especially for confidentiality service RC5 security algorithm is used. RC5 security algorithm needs less system resources.

At last, we evaluate performance of proposed safe RFID system. The result of evaluation presents that proposed system has availability and future works are needed.

References

1. Eschenauer, L. and Gligor, V.: A Key-management Scheme for Distributed Sensor Networks. ACM CCS'02, Nov. (2002) 41-47
2. Chan, H., Perrig, A. and Song, D.: Random Key Pre-distribution Schemes For Sensor Network. IEEE Symposium on Security and Privcy (2003)
3. Du, W., Deng, Jing, Yunghsiang, S. Han, Pramod, K. Varshney,: A Pairwise Key Predistribution Scheme for Wireless Sensor Networks. Proceeding of the 10th ACM conference of Computer and Communication Security, October 27-30, Washington D.C., USA (2003)
4. Zhu, S., Xu, S., Setia, S. and Jajodia, S.: Establishing Pair-Wise Keys for Secure Communicatin in Ad Hoc Networks: A Probabilistic Approach. 11th IEEE International Conference on Network Protocols(ICNP'03), Atlanta, Georgia, November 4-7 (2003)
5. Kalidindi, R., Parachuri, V., Kannan, R., Durresi, A. and Iyengar, S.: Sub-Quorum Based Key Vector Assignment : A Key Pre-Distribution Scheme For Wireless Sensor Networs. Intnl. Conf. on Wireless Networking(ICWN'04), Las Vagas, July (2004)
6. Liu, D. and Ning, P.: Establishing Pairwise Keys in Distributed Sensor Networks. 10th ACM Conference on Computer and Communications Security(CCS'03), Washington D.C., October (2003)
7. Dirk Balfanz, Drew Dean, Matt Franklin, Sara Miner, and Jessica Staddon: Self-healing Key Distribution with Revocation. Proceeding of the IEEE Symposium on Research in Security and Privacy, May (2002) 241-257
8. Duncan, S. Wong and Agnes, H. Chan,: Efficient and Mutually Authenticated Key Exchange for Low Power Computing Devices. In Advances in Cryptology – ASIACRYPT (2001)
9. Weis, S. et al.: Security and Privacy Aspects of Low-cost Radio Frequency Identification Systems. Security and Pervasive Computing , LNCS 2802 (2003) 201-212
10. Rivest, R. L.: Approaches to RFID Privacy. RSA Japen Conference (2003)
11. Sarma, S., Weis, S., and Engls, D.: RFID Systems, Security & Privacy Implications. AutoID Center. white Paper (2002)

12. Stephen, A. Weis: Security and Privacy in Radio-Frequency Identification Devices. MIT, May (2003)
13. Ari Juels and Ronald Rivest L. and Michael Szydlo,: The Blocker Tag: Selective Blocking of RFID Tags for consumer Privacy. RSA Laboratory, MIT
14. Ari Juels, John Brainard,: Soft Blocking: Flexible Blocker Tags on the Cheap. RSA Laboratory
15. Ari Juels,: Minimalist Cryptography for Low-Cost RFID Tags. RSA Laboratory
16. Philippe Golle, Markus Jakobsson, Ari Juels, Paul Syverson, : Universal Re-encryption for Mixnets. RSA Laboratory (2004)
17. Sakata, S.,: Security Technology for Mobile and Ubiquitous Communication. IEICE Magazine, Vol.87, no.5, May (2004)
18. Otsuka, T. and Onozawa, A.,: User Privacy in Ubiquitous Network: Anonymous Communication Technique for Ad-hoc Network. Technical Report of IEICE ISEC (2003)

Robust 3D Arm Tracking from Monocular Videos

Feng Guo[1] and Gang Qian[1,2]

[1] Department of Electrical Engineering
[2] Arts, Media and Engineering Program, Arizona State University,
Tempe, AZ, 85287, USA
{feng.guo, gang.qian}@asu.edu

Abstract. In this paper, we present a robust method to tackle the ambiguities in 3D arm tracking, especially those introduced by depth change (distance of the arm from the camera), and arm rotation about humerus (upper arm bone). In a particle filter framework, the arm joint angle configurations are monitored and the occurrences of the ambiguous arm movements are detected. Inverse kinematics is applied to transfer invalid joint angle configurations from unconstrained movement space into constrained space. Experimental results have demonstrated the efficacy of the proposed approach.

1 Introduction

3D arm tracking from monocular videos is one of the most active and challenging research areas in human motion analysis. The existence of near unobservable and unobservable movements, such as depth change and stretched arm rotation about the humerus, makes the problem even harder to solve. Although Particle filter has been used in arm tracking extensively to capture the multimodality of the posterior distribution of the arm movement parameters, there is no guarantee that the particle filter is able to capture the ambiguities and a large number of particles will be expected to fully represent the solution space. Most important, there is no systematic way to detect the occurrence of ambiguous movements. Recently, some efforts have been made to improve the sampling efficiency of particle filters. For example, learned dynamical model was introduced to present movement constraints[2]. In [3] a continuation principle based on annealing was used to introduce the influence of narrow peaks in the fitness function. In [8], a scaled covariance sampling approach was introduced to search the cost surface along the most uncertain direction and generate samples efficiently.

However, to effectively handle ambiguities in 3D arm tracking from monocular video, explicit considerations have to be taken. In this paper, we present an efficient and robust method, which explores the specific joint angle configurations for ambiguity and singularity. The samples are deployed adaptively so that the local minima will be covered more effectively during sampling. In addition, the joint angle constraints are considered and inverse kinematics is used to transfer invalid samples in the unconstrained arm joint angle space into valid constrained space, which will further improve the sampling efficiency.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 841–850, 2005.

Relation with Existing Work Kinematic jumping processes were used in [7] to deal with depth ambiguity tracking. Kinematic reasoning was used to enumerate the tree of possible depth change. The proposed approach is similar to this in spirit. However, more specific structure of the arm movement configurations are used in the proposed approach. Hence, the proposed approach can handle ambiguities both resulting from depth change and from arm rotation . In [6], the Condensation algorithm was used to successfully track an arm through kinematic singularity. However, in our experiments, we have found out that for Condensation algorithm to successfully track arm through such a singular movement, specific models of dynamic noises for each angle are needed. However, such movement-specified models are not suitable for general movements, since it will increase local uncertainty and will cause more ambiguity. Instead, our approach considering specific mode and inverse kinematics can use more general model.

2 3D Arm Tracking Using A Particle Filter

In our tracking approach, the upper arm and forearm are modeled as truncated cones which are connected by the elbow joint, as shown in Figure 1 (a). The state vector given by $\mathbf{X}_t = [\varphi_x, \varphi_y, \varphi_z, \varphi_e, T_x, T_y, T_z]^T$. It contains global configuration of the arm $[\varphi_x, \varphi_y, \varphi_z]$ and $[T_x, T_y, T_z]$ which respectively represent the rotation angles and translation of the upper arm coordinate system with the camera coordinate system. φ_e is the relative rotation angle of forearm with upper arm.

The 3D arm model can be projected on image plane to generate predicted edges using joint angle samples. Here the method discussed in [9]is explored to obtain four straight lines as the projection edges.

A second order auto regressive process is used to model the dynamics. The dynamic equation is:

$$\begin{bmatrix} \mathbf{X}_t \\ \dot{\mathbf{X}}_t \end{bmatrix} = F \begin{bmatrix} \mathbf{X}_{t-1} \\ \dot{\mathbf{X}}_{t-1} \end{bmatrix} + V_t \tag{1}$$

where $\mathbf{X}_t$ is the state vector and $\dot{\mathbf{X}}_t$ is the velocity of the state vector. F is dynamic matrix, process noise matrix is $V_t = [\mathbf{0}, v_i \mathbf{1}]^T$, where v_i is angle velocity $\dot{x}_i$, following a Gaussian random variable with distribution $N(0, \sigma_i^2)$.

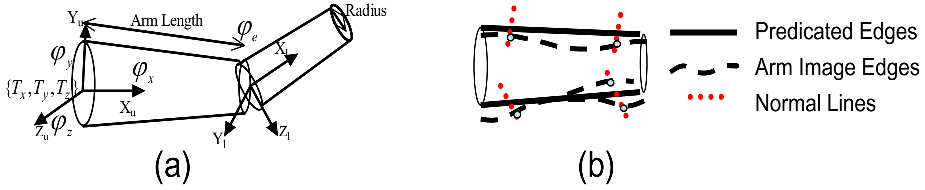


Fig. 1. (a)The 3D arm model (b)Edge matching process

The likelihood based on arm configuration is $p_{image}(z_t|x_t^{(i)})$. Both the edge orientation and intensity of the detected edges are used to compute image likelihood, as shown in Figure 1(b). For one projection line, a set of independent normal lines are generated to measure the likelihood of detected edge points. Along each normal line, the detected edge points are located and the corresponding orientations are calculated. If the difference between edge orientation of the point and the orientation of projected contour is less than a preselected threshold θ, the point is set as edge candidates. Here $\theta = 25$ degrees. This will reduce the clutter noise from the image background. The resulting likelihood function is multi-modal. We combine the distance measure with edge intensity measure. Let K be the number of peaks. For each peak, the distance similarity measure is same with [1] and is given by:

$$p(z_k|c) = e^{-\frac{f^2(d_k;\mu)}{2\sigma^2}} \qquad k = 1 \cdots K \tag{2}$$

where $f(d_k;\mu) = min(d(z_k,c),\mu)$, $d(z_k,c)$ is the distance of point k to the projected contour c, μ controls the clutter-resistance of the tracker, σ^2 is the variance of model and input edge disparity.In these K candidates, the relative weight of each candidate point can be obtained as $\pi_k = \frac{I_k}{N_m}, k = 1, \cdots, K$, where I_k is the edge intensity value and N_m is the normalization factor. Given the clutter probability, for each normal line l, the combined likelihood is obtained by

$$p^l(z_t|x_t^i) = \sum_{k=1}^{K} \pi_k p(z_k|c) + U(1 - \sum_{k=1}^{K} \pi_k) \tag{3}$$

Different from [1], U is uniform outlier distribution, not the Poisson process. $1 - \sum_{k=1}^{K} \pi_k$ is background clutter probability,here we choose 0.05. The overall likelihood is

$$p_{image}(z_t|x_t^i) = \prod_{l=1}^{L} p^l(z_t|x_t^i) \tag{4}$$

where L is the number of normal lines.

3 Ambiguous Arm Movements and Ambiguity Adaptive Sampling

configurations have almost the same observations. Using the previously presented 3D arm model in Figure 1(a), there are at least the following scenarios when ambiguous movement can occur:

- The rotation angle along the upper arm bone φ_x is difficult to be observed in any camera view when the elbow angle φ_e is close to be straight. This is one of main singular movements.
- The depth change of the arm movement is controlled by the angle φ_y. It makes the arm move forward and backward when facing the camera. It is near unobservable with only edge information. The ambiguities exist when φ_y changes small and forward/backward flips to $-\varphi_y$.

- The angle φ_z controls the height of the arm projection edge. It is easy to track these kinds of movements because the observations of the edges are obvious.

If there is only one mode for sample distribution, most of the samples will be trapped in the local minima valley of the matching cost. When the arm leaves ambiguous or singular configuration, it is largely possible that the likelihood peak will be the tail of the samples distribution and further the tracking will lose. If the samples can be generated in multiple modes in ambiguous and singular conditions, the correct posterior distribution will be generated once the movement leaves these conditions. This will make the tracking consistent.

To perform consistent 3D arm tracking in the presence of ambiguous or singular movements, a two-step strategy is used: first, the conditions of ambiguity are checked; if an ambiguous movement is detected, an adaptive sampling scheme is invoked.

Ambiguity Detection. It is easy to see that the ambiguous and singular conditions are mainly based on the values of arm bone rotation angle φ_x, elbow angle φ_e and depth rotation angle φ_y. For example, when $\varphi_x = 90°$ and φ_e is small, the arm will be almost straight. The combination of any value of φ_x and small value of φ_e will be ambiguous. Therefore, by looking the values of these three angles, ambiguous movements can be detected. From tracking results of the previous three frames $[t-3:t-1]$, $\hat{\varphi_x}$, $\hat{\varphi_y}$ and $\hat{\varphi_e}$ are computed as the average of φ_x, φ_y and φ_e in these three frames. Define $R_{t-1} = [\hat{\varphi_x},\ \hat{\varphi_y},\ \hat{\varphi_e}]$. If the angles in R_{t-1} fall into the aforementioned scenarios of ambiguity and singularity, it means that the current movement is ambiguous. Because of the continuous characteristic of the human movement, the current movement is also in the same ambiguous scenario. Hence, the new samples for current time instant will be generated adaptively using R_{t-1}.

Adaptive Sampling. Because the joint angles are almost independent, we draw samples for each angle separately.

Sampling method for φ_y

For the depth control angle $\hat{\varphi_y}$, if it is not near zero, the image projection will produce two possible 3D configurations . So new φ_y is sampled as follows:

```
– If $\hat{\varphi_y} \approx 0$,
    • Draw sample $\varphi_y \sim N(\hat{\varphi_y}, \sigma_1^2)$. Empirically, $\sigma_1 = 15°$.
– else
    • Draw sample $\varphi_y$ from two Gaussian distribution $N_1(\hat{\varphi_y}, \sigma_2^2)$ and $N_2(-\hat{\varphi_y}, \sigma_2^2)$. Empirically, $\sigma_2 = 10°$.
```

Sampling method for φ_x and φ_e

The humerus rotation angle $\hat{\varphi_x}$ is the main factor in singularity. When it is far away from 90° and the elbow angle $\hat{\varphi_e}$ is far away 0°, no adaptive steps will be implemented. Otherwise, φ_x and φ_e will be sampled from the possible kinematic positions as followed:

$\theta_1 \in [0, 1]$ is threshold to control sample which is from whole range of φ_x. θ_e is threshold near zero to cover elbow angle. μ_1, μ_2 and μ_3 are mean values with value 30°, 90° and 150°. $\sigma = 20°$ is sample variance from empirical data.

```
– Draw uniform random number $n_1$ and $n_2$
– If $n_1 < \theta_1$
    • Draw sample $\varphi_e \sim U(-\theta_e, \theta_e)$
    • If $n_2 < 1/3$
        ∗ Draw sample $\varphi_x \sim N(\mu_1, \sigma^2)$
    • Else if $n_2 < 2/3$
        ∗ Draw sample $\varphi_x \sim N(\mu_2, \sigma^2)$
    • Else
        ∗ Draw sample $\varphi_x \sim N(\mu_3, \sigma^2)$
– Else
    • Draw sample $\varphi_x \sim N(\hat{\varphi_x}, \sigma^2)$
    • Draw sample $\varphi_e \sim N(\hat{\varphi_e}, \sigma^2)$
```

With these steps, the multiple modes with approximately-correct image projections will be obtained as prior to generate posterior distribution. In [5], a small number of samples leap in state space as candidates for a sudden shape change in each iteration step. While our approach deploys the samples at all possible hypotheses, not changing from one mode to another one.

4 Invalid Sample Transfer Using Inverse Kinematics

Anatomical kinematic constraints limit the joint movements. In traditional particle filter-based articulate limb tracking algorithms, these constraints are enforced to generate physically valid samples. For example, in [6], at regions close to the endstop (angle limits) in the valid state space, the state velocity was reversed proportionally to a reversal coefficient drawn from a uniform distribution. Thus all the samples are physically valid, i.e. within the possible state space defined by the physical anatomical limits. However, at the same time, as we will explain below, it also prohibits tracking recovery from singular movements.

Consider the scenario mentioned above, where φ_x slowly changed with arm outstretched. Assume that φ_x changed from one position to the others, namely from $\varphi_{x(a)}$ to $\varphi_{x(b)}$. Let the gap be $\Delta\varphi_x = \varphi_{x(b)} - \varphi_{x(a)}$ during the tracking. Assume that φ_x is relatively large, say 60 degree. Since the change of φ_x is unobservable in the image, nearly all entries in current samples corresponding to φ_x will be far away from $\varphi_{x(b)}$, the true rotation angle when small dynamic noise is applied to φ_x. With the above anatomical constraints enforced in sample generation, the corresponding φ_x in samples are way off from the ground truth. Consequently, tracking will fail right after the arm moves out of such a singularity. One way to solve this problem is to allow the existence of *physically-invalid samples* which can correctly track the arm profiles on the image plane and then transfer these invalid samples back to valid sample space later. Hence, tracking can be successfully recovered after singular movements. We apply this "unconstrained+transfer" strategy to our framework.

In this section, we present one approach which transfer invalid sample more efficiently. In our approach, we use the joint angle limits in a way similar to the one presented in [2], except that there is no hard lower bound applied in sample generation for the elbow angle φ_e. The physical range of φ_e is $[\varphi_{e,min}, \pi]$, where $\varphi_{e,min}$ here is chosen as 15 degrees. When there is a predicted sample, with φ_e out of this physical range, it will not be corrected immediately. Instead, these invalid samples will be utilized to keep track of the arm profiles on the image plane and the forearm distal point (i.e. the wrist) in the 3D space.

Once a physically-invalid sample has been used to track the arm, it then can be mapped or transferred back to the valid joint angle space.This step of sample transfer is done through inverse kinematics. First, forward kinematics is used to obtain the 3D position of shoulder, elbow and wrist from kinematic chain using the joint angle configuration in the invalid sample, since the 3D shape of the arm is known and modeled using connected truncated cones. Suppose the related joint angles are $[\varphi_x, \varphi_y, \varphi_z, \varphi_e]$ and translation $[T_x, T_y, T_z]$, the joint position in camera coordinate system can be obtained.

With these joints positions and rotation of initial position with respect to the camera coordinate system, similar to [7], we can infer the actual joint angles from the inverse kinematics, as shown in Figure 2.

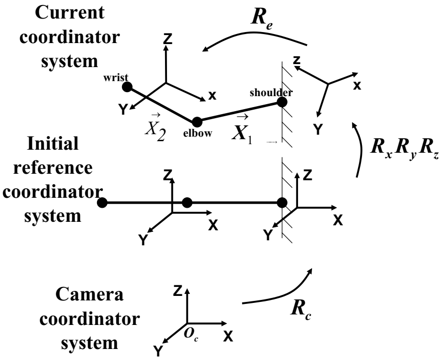

Current coordinator system
R_e
Z
x
wrist
Y
$\vec{X}_2$
elbow
$\vec{X}_1$
shoulder
z
x
Y
$R_x R_y R_z$
Initial reference coordinator system
Z
X
Y
Z
X
Y
Camera coordinator system
Z
O_c
X
Y
R_c

Fig. 2. Mapping using inverse kinematics

The elbow angle is the dot product of vector of $\overrightarrow{x}_1$ and $\overrightarrow{x}_2$, which are normalized vector along local x axes direction. It should satisfy the inequality

$\varphi_e > \Psi_{min}$. The rotation angle φ_x of upper arm can be obtained using inverse kinematics. To calculate the shoulder rotation angles, we use a simple Euclidean coordinate transformation. Given previous coordinate rotation with initial reference system, we know vector $\overrightarrow{\mathbf{x}}$ in camera coordinate system as $\overrightarrow{\mathbf{x}}_1$ and current coordinate system as $[1 \quad 0 \quad 0]^T$. We can calculate the rotation for this vector from these two systems to previous coordinate system. And we can obtain the rotation angles φ_y and φ_z.

For example, let the previous system be the initial reference coordinate system. Let $R_xR_yR_z$ be the rotation matrix from the initial system to the current system, and R_c be the rotation matrix from the camera system to the initial reference coordinate system. We have

$$(R_xR_yR_z)^{-1}\begin{bmatrix}1\\0\\0\end{bmatrix}=\begin{bmatrix}cos(y)cos(z)\\cos(y)sin(z)\\-sin(y)\end{bmatrix}=R_c\overrightarrow{\mathbf{x}}_1 \tag{5}$$

Regarding the forearm, using kinematic chain we have

$$R_x^{-1}R_e^{-1}\begin{bmatrix}1\\0\\0\end{bmatrix}=\begin{bmatrix}cos(e)\\cos(x)sin(e)\\sin(x)sin(e)\end{bmatrix}=R_yR_zR_c\overrightarrow{\mathbf{x}}_2 \tag{6}$$

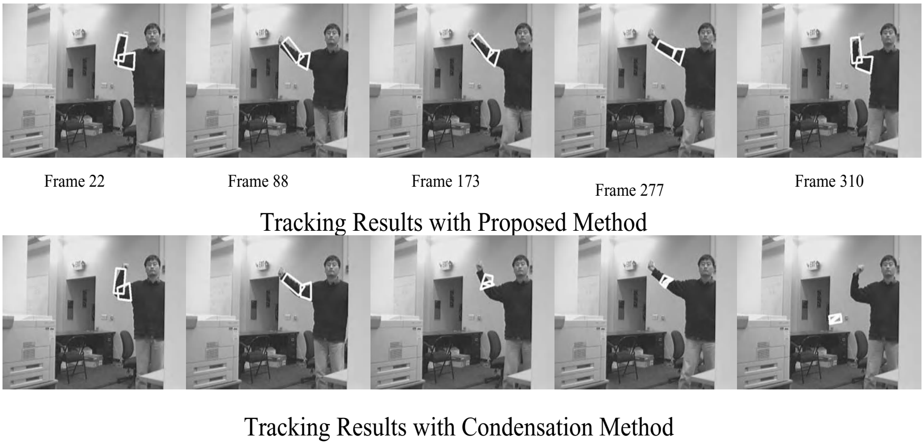
Based on the above two equations, the invalid samples can be easily transferred to the corresponding values in the valid joint angle space.

Different from [7] is that we only get one angle value for elbow angle and upper arm bone rotation angle. This is because we have joint constraints to set the angle in specific range. Also only samples with strong weight will be transferred using this method. For those samples without strong weight, it means the 3D position is not consistent with image projection and transferring is useless. Here we choose weights with values greater than 0.2 maximum weight value.

5 Experimental Results

The current implementation is developed using C++ on a *Pentium IV* 3.0GHz PC running Windows XP. The tracking system runs at 25 frames/sec with 640 × 480 color images for 32 normal lines searching along whole arm edge and 800 particles. Three examples are included here. Results from original Condensation algorithm and the proposed approaches were obtained and compared.

Example 1: Large Shoulder Movement. In the first example, a sequence with the ambiguous movements and singularity discussed previously was used. This sequence has 470 frames (see the attached *video1*). In this video, the shoulder position was quickly changing. Both approaches used 1500 motion samples and the same levels of dynamic noises. As shown in the upper row of Figure 3, the proposed approach tracked arm movements consistently and it can also recover from large tracking errors. In contrast, the Condensation algorithm eventually lost the tracking.

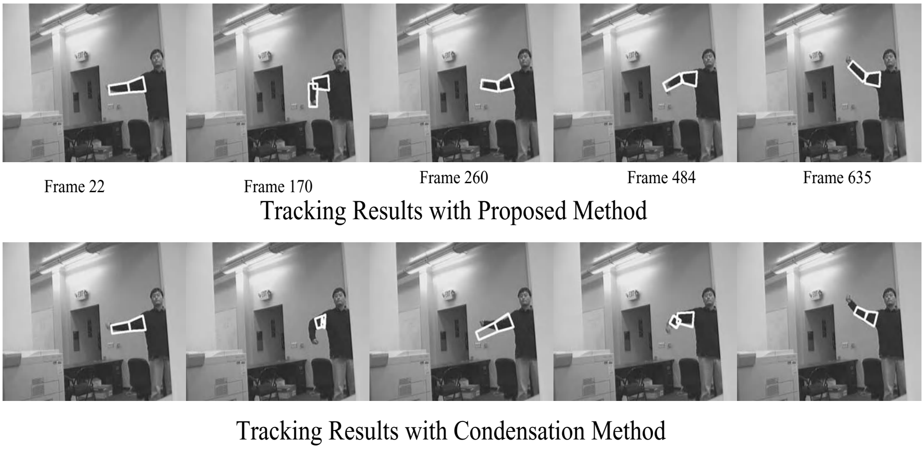


Tracking Results with Proposed Method

Tracking Results with Condensation Method

Fig. 3. Results for *Example 1*

Example 2: Small Shoulder Movement. The sequence used in the second example has 820 frames (see the attached *video2*). This sequence contains more ambiguous movements. But the shoulder position changes very small. Both approaches used 300 motion samples and same levels of dynamic noise. The upper row of Figure 4 shows the results obtained using the proposed approach tracked arm movements, which is more accurate than that obtained using the Condensation method.



Tracking Results with Proposed Method

Tracking Results with Condensation Method

Fig. 4. Results for *Example 2*

Example 3: Effect of Different Sample Numbers. A sequence with 985 frames was used in the third example (see the attached *video3*). The proposed approach used

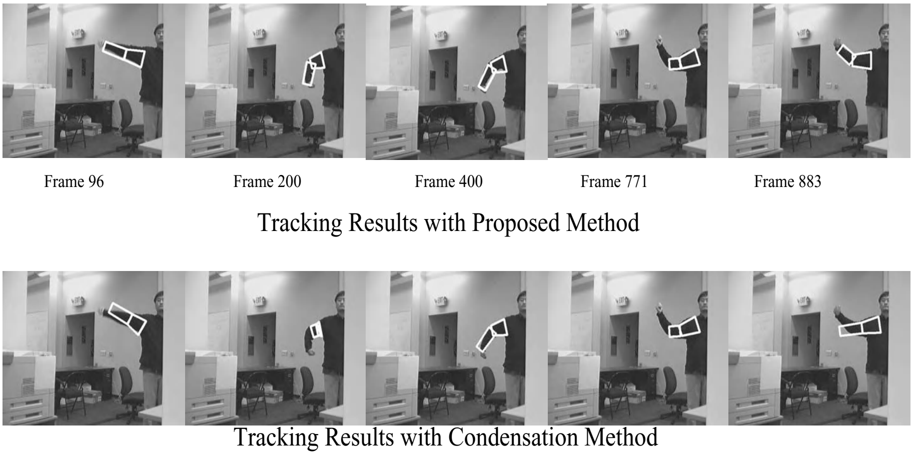

Frame 96
Frame 200
Frame 400
Frame 771
Frame 883
Tracking Results with Proposed Method
Tracking Results with Condensation Method

Fig. 5. Results for *Example 3*

300 motion samples during the tracking, while 500 samples were used for the the Condensation method. Both had the same levels of dynamic noises. As shown in 5, the proposed approach still outperformed the Condensation algorithm, even a smaller number of samples were used.

6 Conclusions

We present an efficient and robust method to tackle depth ambiguity and singularity in arm tracking. In our approach using a particle filter for 3D arm tracking, joint angles conditions for the occurrence of depth ambiguity and singularity are checked. Different sampling methods are used to generate samples efficiently in ambiguity and singularity scenes. Also we explore the joint angle constraints and use inverse kinematics method to transfer invalid samples from unconstrained movement parameter space into valid constrained space. Experimental results have demonstrated the efficiency of our approach in terms of explicit consistent tracking for ambiguous and singular scenes and small number of particles. This approach can be extended to the whole body tracking. For example, the specific angle configurations for the arms and legs can be explored in the whole body framework. Because each limb is sampled independently, the adaptive sampling for each limb will be used separately.

References

1. Isard M.and Blake A.: Condensation-Conditional Density Propagation for Visual Tracking. *IJCV*,Vol.29, No.1,(1998)5-28
2. Sidenbladh H., Black M. and Fleet D.: Stochastic Tracking of 3D Human Figures Using 2D Image Motion. *In ECCV*, (2000)

3. Deutscher J., Blake A. and Reid I.: Articulated Body Motion Capture by Annealed Particle Filtering. *CVPR* Vol.2, (2000) 126-133
4. MacCormick J., Isard M.: Partitioned Sampling,Articulated Objectes and Interface-quality Hand Tracking. *ECCV*,Vol.2, (2000) 3-19
5. Heap T., Hogg D.:Wormholes in Shape Space: Tracking Through Discontinuous Changes in Shape. *ICCV*,(1998) 344-349
6. Deutscher J., North B., Bascle B. and Blake A.: Tracking through Singularities and Discontinuities by Random Sampling. *ICCV*,(1999) 1144-1149
7. Sminchisescu C. and Triggs B.: Kinematic Jump Processes for Monocular 3D Human Tracking. *CVPR*, Vol.1, (2003) 69-77
8. Sminchisescu C. and Triggs B.: Covariance Scaled Sampling for Monocular 3D Body Tracking. *CVPR*, Vol.1,(2001)447-454
9. Stenger B.: *Model-Based Hand Tracking Using A Hierarchical Bayesian Filter.* Ph.D. Thesis, University of Cambridge, St. John¡s College, (2004)

Segmentation and Tracking of Neural Stem Cell

Chunming Tang[1] and Ewert Bengtsson[2]

[1] Info. & Commun. Eng. Coll., Harbin Engineering University, Postfach 15 00 01,
Harbin, China
april1971@vip.sina.com.cn
[2] Centre for Image Analysis, Uppsala University, Sweden, Postfach s-752 37
Uppsala, Sweden
ewert@cb.uu.se

Abstract. In order to understand the development of stem cells into specialized mature cells it is necessary to study the growth of cells in culture. For this purpose it is very useful to have an efficient computerized cell tracking system. In order to get reliable tracking results it is important to have good and robust segmentation of the cells. To achieve this we have implemented three levels of segmentation: based on fuzzy threshold and watershed segmentation of a fuzzy gray weighted distance transformed image; based on a fast geometric active contour model by the level set algorithm and interactively inspected and corrected on the crucial first frame. For the tracking all cells are classified into inactive, active, dividing and clustered cells. A special backtracking step is used to automatically correct for some common errors that appear in the initial forward tracking process.

1 Introduction

Neurogenesis has been seen in adult brains from both animals and humans [1]. To date, little is known about the basic regulatory mechanisms of neurogenesis. In order to understand this regeneration of brain cells, cultured cells are studied. In this way, some properties of neuronal stem cells as they develop over time can be discovered. For this purpose efficient methods for tracking cells in cultures are needed. However, to get good results several special adaptations and heuristics are necessary. Estimating the motion of objects from a sequence of images consists of two major steps: (i) segmentation and (ii) tracking, which typically comprises the position and velocity [2] of the object. This paper will focus on these two steps. The segmentation is quite difficult for this application since the cells are unstained, as the stain would be harmful to the living cells. The contrast is thus quite low. The cell images are also acquired with auto-focus, which sometimes yields a poorly focused image. Therefore, the segmentation of the images sequence will be implemented in three levels, which is presented in section 3. The characteristic of tracking unstained stem cells is that the irregularly and widely varying shapes of the cells, both within an image and across a series of images, makes it impossible to use masking or direct matching techniques to distinguish the cells. In our tracking system the position of the centroids of each region in subsequent frames are considered in an initial automatic

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 851 – 859, 2005.

processing step and regions that can be matched are linked as be presented in section 4. Subsequent analysis steps try to link remaining unmatched regions using special heuristics. Finally the automatic tracking results can be visually checked and corrected if necessary.

2 Image Acquisition

Images were captured from a computer-controlled microscope attached to a cell culture system with carefully controlled environment for the cells. The time interval between images was about 15 minutes, yielding total sequences of on the average 70 frames, each illustrating the behavior of the cells under influence of a catalyzing chemical substance. The cells were unstained, and the image acquisition was completely automatic with auto-focus applied for each image frame.

3 Segmentation

3.1 Initial Automatic Segmentation of the Whole Sequence

To segment the images in the whole sequence into individual cells we first smooth the image with a small (3x3) Gaussian filter to reduce noise. We then perform a fuzzy threshold as follows: All pixels with intensity below a lower threshold t_l are set to 0 and all pixels above a higher threshold t_k are set to 1. Between t_l and t_k image intensities are linearly rescaled to the range [0,1]. The thresholds we have chosen are for $t_l = \mu + 0.3\sigma$ i.e just above the background level. μ is the mean value and σ is the standard deviation of the background intensity. Similarly we have chosen $t_k = \mu + 4\sigma$. That is high enough to guarantee that pixels brighter than that are really well inside the cells. Through the use of a fuzzy approach, the method becomes less sensitive to the exact values of these threshold levels, than what would have been the case if a standard crisp threshold had been used. On the fuzzy threshold image, we apply a fuzzy gray weighted distance transform [3] to incorporate both the shape (roundness) and the intensity of the cells. This gives us a good "landscape" to segment using the watershed algorithm [4]. We use the extended *h*-maxima transform [5] to find suitable seed points for the watershed algorithm, where *h* is fixed. We require the seeds to have intensity above a threshold *h* in the fuzzy distance transformed image to use them for the watershed transforms. In that way small and faint objects are automatically removed. After segmentation of each of the frames in the sequence of stem cell images, a series of labeled images can be obtained through application of the standard connected component-labeling algorithm.

3.2 Second Stage Segmentation on Poorly Focused Images

Since the focusing during the image acquisition is fully automatic, some images will have very poor focus, as shown in figure1b. The automatic segmentation algorithm described in the previous section sometimes fails in such images, giving results such as in figure2b. This is detected through a simple test. If the number of detected cells in the current image is less than 0.15 times the number of total cells in the whole sequence which means the number of segmented cells no more than 3, we have a segmentation failure. We then try another more computationally expensive segmentation algorithm to try to recover the lost cells. The algorithm we use is a fast geometric active contour model based on the level set algorithm. Geometric snakes are based on the theory of curve evolution in time according to intrinsic geometric measures of the image and implemented via level set algorithms [10]. This implementation approach helps to automatically handle changes in topology and hence, unknown numbers of multiple objects can be detected simultaneously. There has been a number of works based on the geometric snake and level set framework, e.g. [7]-[9], [12]. In our implementation of the process we have omitted the curvature term in order to avoid that regions are merged between the adjacent cells and also to speed up the convergence. At any time, the moving front is simply the level set $\{\Phi(x,t)=0\}$. The evolution of Φ is described by the partial differential equation (PDE):

$$\Phi_t = F\,|\nabla\Phi| + \nabla g(|\nabla I|) \bullet \nabla\Phi \tag{1}$$

The speed term *F* consists of two components: $F = F_a + F_i$. The term F_a is the advection term, which defines a uniform speed of the front in the normal direction, corresponding to the inflation force in classical snake models [12]. The $F_a(I_\sigma)$ standing for F_a in images is defined as follow:

$$F_a(I_\sigma) = \begin{cases} +1, & if \quad I_\sigma \geq T \\ -1, & otherwise \end{cases} \tag{2}$$

With: $$I_\sigma = G_\sigma * I \tag{3}$$

which denotes the gray level image convolved with a Gaussian smoothing filter whose characteristic width is σ. In this application, σ is standard deviation (in pixels) which is set to 1 and 2. T in (2) is defined as follow:

$$T = \frac{t}{length(h(:))} + \delta \tag{4}$$

Where, t is a threshold, derived from the method of Minimum Error Threshold [11]. h is a histogram which shows the distribution of data values. The range of δ is [0.00, 0.05]. The second component F_i of the speed term *F* is defined as:

$$F_i = -|\nabla I_\sigma| \bullet F_a \tag{5}$$

With: $|\nabla I_\sigma|$ normalized to [0, 1]. The second term of (1) depends on the gradient of the factor g (.), which denotes the projection of the attractive force vector on the surface normal. The stopping function *g* should tend to zero when reaching edges [12]. Thus *g* is defined as:

$$g(x, y) = \frac{1}{1 + |\nabla(I_\sigma)|} \tag{6}$$

Since our goal is to try to capture all cells in the image the initial contour is defined as a rectangle of the same size as the image, shown as the first subplot in figure 3. Figure 3 shows the contours detected in the image of figure 1 after each time step of evolution. The parameters σ in (3) and δ in (4) are adjusted as the following two loops through the contour evolution:

```
 For sigma=1:2
   For delta=0.00:0.01:0.05
              ...
   End
End
```

In order to shorten the running time and make the segmentation as robust and automatic as possible, two additional stopping criteria have been added based on a priori knowledge of the application. In these two loops, if one of the following situations happen, the evolution on the parameter pair is stopped and the next pair will be processed:

1. No implicitly defined surface exists after the first time step
2. If it is found that the initial boundary rectangle has not evolved after the first time step.

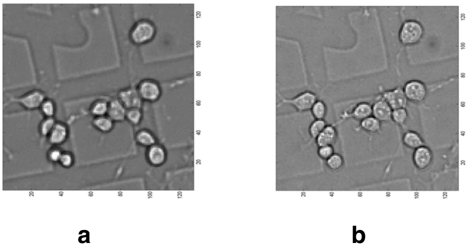


Fig. 1. Different focus in two contiguous frames

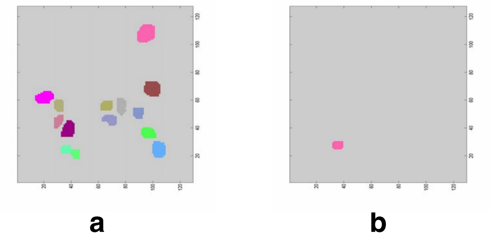


Fig. 2. The corresponding automatically segmented images using algorithm of section 3.1

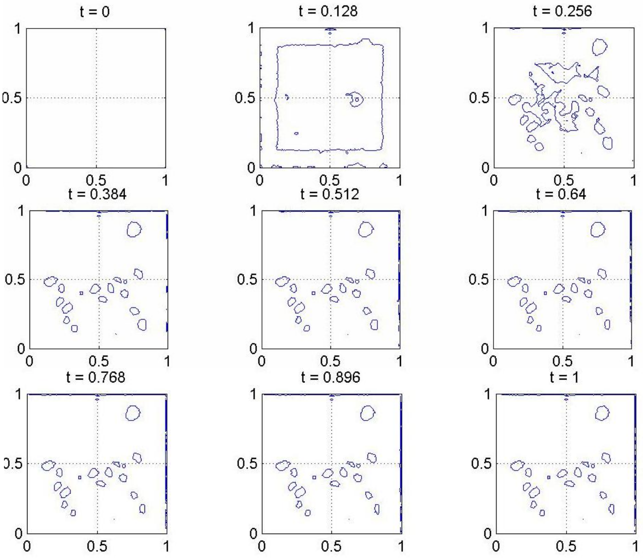


Fig. 3. Evolving contours for segmenting figure 1(b) via equation 1

3.3 Interactive Segmentation on the Crucial First Frame

Since the cell tracking is based on propagation of cell identities from the first frame throughout the entire sequence it is important to have a correct starting frame [13]. The first image frame is shown in original gray level and segmented versions as illustrated in figure 4a and 4b. If under or over segmentation has happened it is usually obvious to human vision. The user is therefore given an opportunity of interactively correcting the segmentation as illustrated in figure 4c. The whole set of cells presented in the sample will thus be tracked. At this time the operator also tells the system which cells is part of clusters.

4 Tracking

4.1 The Tracking Algorithm for Inactive Cells

Most cells will only move short distances between two contiguous images. Those cells are called inactive cells. For those cells overlap tracking algorithm leads to good results [13]. When all the overlapping regions have been matched the lists of cell ID numbers of these two successive tracking images are compared. There may then be some regions in the previous frame that seems to have disappeared and some newly appearing regions that need to be handled. New appearing and disappearing cells ID are saved in two 2-D matrixes, called *newappear* and *disappear* respectively. Biologically cells do not disappear from one frame to another except when moving through the image border. Disappearing cells are thus likely to be caused by under-segmentation. In order to avoid that some cells are lost in the tracking, all disappearing cells' regions are copied from the previous frame into the tracking image except for the positions at the border of the image. When later on it is found that part of a disappearing cell's region coincides with another cell in a new tracking image, the region ID of the former is replaced by that of the latter.

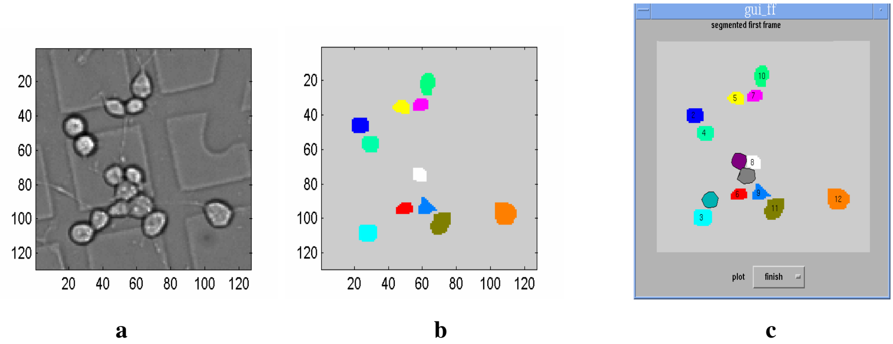


Fig. 4. Interactive initialization of the tracking on the first frame. a) 1st frame, b) Segmented image of 1st frame, c) Interactively corrected segmentation of 1st frame.

4.2 How to Detect and Handle New Cells

When moving through the sequence new cells may appear. There are three distinct reasons for this that has to be recognized and dealt with in different ways:

1 There has been a cell division, creating two cells from a parent cell
2 A cell has moved into the scene through one of the image borders
3 A cell has been split through over-segmentation

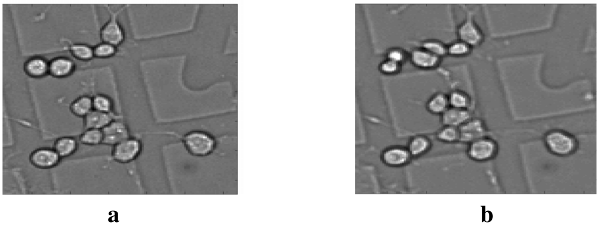


Fig. 5. A scene of two contiguous gray level images with a cell division. (a) Before division, (b) After division.

It is a very important aspect of stem cell tracking to detect cell division and to save the information about the division. For that purpose it is useful to note some biological facts that influence the cell appearance in the images: before a cell division, the cell keeps almost stationary for several successive frames. Then it becomes very round just before the frame where the cell begins to divide because the surface tension of the cell membrane reaches maximum, as shown in Figure 5(a). There are two kinds of division. One is called symmetric division. The other is called asymmetric division. The former division is that the two daughter cells very symmetrical around the perpendicular bisector of the line linking the two centroids, as shown in Figure 5(b). The area of the parent cell is almost equal to the sum of the two daughters' areas. One daughter cell keeps its parent cell's original position while the other one is pushed away some distance. We use these features in an analysis of each new cell region that appear during the tracking to see if it can be a cell division. Details of this algorithm can be found in [13].

4.3 The Tracking Algorithm for Active Cells

If the new appearing j_{th} cell neither is coming from the border nor is meeting the conditions of a division cell, then we check whether it is an *active cell* which could not be detected by overlap. Active cells are defined as cells that move rapidly, more than D_{slow} pixels per frame. Biologically they are of special interest. All the disappearing cells in the current frame registered in the 2-D *disappear* matrix $C_d(i)$ become candidates in order to match the active cell. To reach the best matching, features are computed for each candidate. The following formula is used to find the best candidate [6]:

$$C_{\min} = \underset{Cd(i)}{\operatorname{argmin}} f(C_{d(i)}) = \underset{Cd(i)}{\operatorname{argmin}} \{\alpha D_{j,Cd(i)} + \beta A_{j,Cd(i)} + \gamma P_{j,Cd(i)}\} \quad (7)$$

In formula (7), $D_{j,Cd(i)}$ is the distance between the i_{th} candidate and the new appearing j_{th} cell. $A_{j,Cd(i)}$ is the difference in area between the i_{th} candidate and the new appearing j_{th}. $P_{j,Cd(i)}$ is the difference in perimeter between the i_{th} candidate and the new appearing j_{th}. The, β and γ are weights. Finding the minimum $f(C_{d(i)})$ means finding the cell in the *disappear* matrix that best matches this new appearing j_{th} cell. Among the three factors, the distance is the most important. The other two weights depend on the segmentation algorithm. We have used α=0.6, β=0.2, γ=0.2. The region label for the j_{th} new cell ID is changed to the best matching candidate's ID. There are two important parameters in the tracking algorithm relating to cell speed: D_{fast} and D_{slow}. These parameters are based on a model of the cells behavior in terms of Brownian motion. By fitting such a model to the data we have found the parameters to be: D_{slow}, the distance moved by the majority cells is no more than 10 pixels; D_{fast} the distance moved for active cells is no more than 18 pixels [13].

4.4 Backtracking

After performing the cell tracking in a forward matter as described so far, some errors may have appeared that can be detected and corrected through some backtracking through the sequence. When some cells disappear due to under-segmentation and some move a lot because they are being active, the previous algorithms may not work properly. The most apparent mistake is that two cells exchange their ID in two successive tracking frames. A special processing step will detect and correct this error: Build a subset of all the cells in the k_{th} +1, tracking image that are more than D_{slow} pixels apart by scanning all cells between two contiguous tracking images. Study the orientation of the distance vector between cell pairs in this subset to see if any two has similar magnitude and opposite directions. Those are likely to have been exchanged by mistake and this can be corrected by swapping the ID numbers. Some cells will not appear to move at all between several successive frames in the sequence. Such cells are called static cells. The reason may be either biological or a processing artifact. There are some cells that are static because of the replacement, presented in the end of section 4. The reasons are over-segmentation or some cells moving in from the border then out again after several frames. We set up a 2-D matrix called *static_v*, which is formed, by frame number and *cells-ID*. It records the ID numbers of all cells that keep static positions between two successive frames. For each frame after the third one in the sequence we compute for each cell C_{st} in *static_v*, the number called *t_static* of successive frames that the cell stays static.

If 3<= *t_static*<=4, we find the first frame where this cell became static. Then we check whether there is some new cells C_{new} appearing in this frame (including the cells moving in from the border but not including the cells from division). If the distance $|C_{new} - C_{st}| < D_{fast}$, the region of the static cell is updated to zero while the region of the new cell is updated to the ID of this static one in all frames where this static cell exists. After this, the corresponding values in *static_v* are set to zero. If no new cell appeared in the first frame, the static cell is kept for another few frames. If *t_static* >=5, and there are some static cells that were not marked as belonging to clusters in the initial processing, these static cells are deleted for all these frames. After that, the corresponding values in *static_v* are set zero. These static cells are regarded as coming from over-segmentation.

5 Results and Conclusion

Figure 6 and 7 illustrate two successful applications of the described procedure: creating a complete trace of a set of neural stem cells over 71 frames. In Figure 7, the traces marked with a triangle are daughter cells after division. These traces do not include cells that remain clustered. The plot shows some errors due to under-segmentation causing unrealistically large distances between points, such as, 1 to 2 and 2 to 3. One of the merits of the system is that it tracks all cells present in the cell

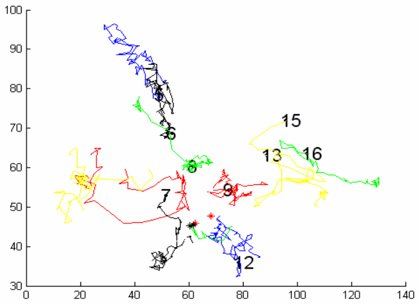


Fig. 6. Trace of cells in a sequence of 70 frames

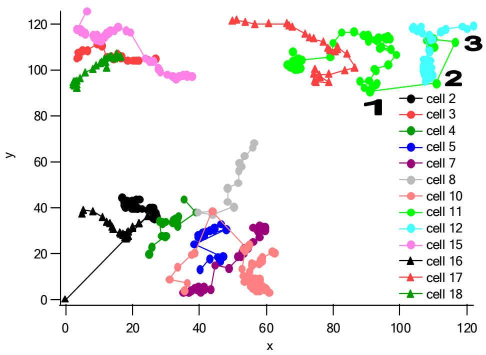


Fig. 7. Trace of cells in another sequence of 71 frames

culture, not just a few selected ones. The system still has some limitations. The judgment about when to split cells during the segmentation stage could be made more accurate through the use of more advanced shape analysis. There is also a possibility of using feed-back from the tracking stage to the segmentation stage, i.e. to look more carefully for cells in regions where the tracking algorithm suggests there should be a cell but the initial segmentation did not find one.

Acknowledgment

The authors gratefully acknowledge Joakim Lindblad and Carolina Wählby of the Center for Image Analysis of Uppsala University for developing the initial automatic segmentation step and to Karin Althoff and Johan Degerman of Chalmers University for recording the stem cell image sequences.

References

1. Eriksson, P., et al.: Neurogenesis in the Adult Human Hippocampus. Nat. Med., 4 (1997) 1313-1317
2. Kirubarajan, T., Bar-Shalom, Y.: Combined Segmentation and Tracking of Overlapping Objects with Feedback. Multi-Object tracking, 2001 IEEE Workshop (2001) 77-84
3. Saha, P. K., Wehrli, F., Gomberg, B. R.: Fuzzy Distance Transform: Theory, Algorithms, and Applications. Computer Vision and Image understanding, Vol. 86, No. 3 (2002) 171-190
4. Vincent, L., Soille. P.: An efficient Algorithm Based on Immersion Simulations. IEEE Trans. on Pattern Anal. and Machine Intelligence, Vol. 13, No. 6 (1991) 583-597
5. Soille, P.: Morphological Image Analysis: Principles and Applications. First edn. Springer- Verlag, Berlin Heidelberg New York (1999) 170-171
6. Xia, L., Mingming, H., David, W. P. Benoit, M. D.: Automatic Tracking of Proteins in Sequences of Fluorescence Images. Proceedings of SPIE---Vol. 5370, Medical Imaging 2004: Image Processing (2004) 1364-1371
7. Xu, C., Yezzi, J., Prince, J.: On the Relationship between Parametric and Geometric Active Contours. In: Proc. 34th Asilomar Conf. on Signal Systems, and Computers (2000) 483-489
8. Paragios, N., Deriche, R.: Coupled Geodesic Active Regions for Image Segmentation: a Level Set Approach. In: Proc.6th Eur. Conf. on Computer Vision (2000) 224-240
9. Yezzi, A., Tsai, A., Willsky, A.: A Fully Global Approach to Image Segmentation via Coupled Curve Evolution Equations. J. Vis. Commun. Image Represent, Vol. 13, No. 2 (2002) 195-216
10. Sethian, J.: Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics. Computer Vision, and Materials Science: CUP (1996)
11. Glasbey, C.: An Analysis of Histogram-Based Thresholding Algorithm. Graphical Models and Image Processing, Vol. 55, No.6 (1993) 532-537
12. Xianghua, X., Majid, M.: RAGS: Region –Aided Geometric Snake. IEEE Transactions on Image Processing, Vol. 13, No.5 (2004) 640-652
13. Chunming, T., Ewert, B.: Automatic Tracking on Neural Stem Cells. Proceedings of WDIC 2005 Brisbane, Australia (2005) 61-66

License Plate Tracking from Monocular Camera View by Condensation Algorithm

İlhan Kubilay Yalçın[1] and Muhittin Gökmen[2]

[1] TUBITAK MRC
Information Technologies Institute Kocaeli, Turkey
ilhan.yalcin@bte.mam.gov.tr
[2] ITU Computer Engineering Department
Istanbul Technical University Istanbul, Turkey
gokmen@cs.itu.edu.tr

Abstract. In this paper, we present a novel approach for pose estimation and tracking of license plates from monocular camera view. Given an initial estimate, we try to track the location, motion vector and pose of the object in 3D in the successive video frames. We utilize Condensation algorithm for estimating the state of the object and filtering the measurements, according to the extracted image features. We utilize directional gradients as the image features. Each sample of the Condensation algorithm is projected to the image plane by perspective camera model. The overlapping of the image gradients and the sample boundaries, gives a likelihood for each sample of the Condensation algorithm. Our contribution is utilizing condensation algorithm for rigid object tracking, where the object is tracked in 3D. We demonstrate the performance of the approach by tracking license plates in outdoor environment with different motion trajectories.

1 Introduction

Probabilistic visual tracking is getting popular in the computer vision community. Probabilistic object tracking involves the construction of the probability distribution of the current state of an evolving system, given the previous observation history. For linear Gaussian models, where the probability distribution can be summarized by means and covariances, the calculation is carried out in terms of the familiar updating equations of the Kalman filter [1].

Generally, for nonlinear, non-Gaussian models, there is no simple solution. Several approximation methods have been used in literature. The known method are, the Extended Kalman filter [2], the Gaussian sum filter [3], approximating the first two moments of the probability distribution [4, 5] and numerical integration over a grid of points in the state space [6]. But none of these methods can be applied directly. Generally, these methods need to be tuned for each specific problem.

Recently probabilistic object tracking algorithms especially particle filter is attracting considerable attention. Particle filter can deal with general non-linear

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 860–869, 2005.

and non-Gaussian problems [7]. Condensation algorithm is a variant of particle filter and can deal with non-linear, non-Gaussian tracking problems [8]. Beside its advantages, Condensation algorithm has drawbacks. Sample impoverishment can appear when peaked likelihood or the new measurements exist in the tail of the prior [10]. There have been many attempts to overcome those drawbacks [9]. Much effort is also devoted to reducing the number of samples for representing the prior and posterior distributions [11].

Rigid object tracking has also been researched for a long time. One of the important works is the RAPID object tracker by Haris et. al. [12]. The pose of the object is estimated, while the object boundaries are tracked. The RAPID tracker represents a 3D object as a set of control points, namely high contrast edges. The control points are projected onto the image using the predicted pose of the object obtained from the Kalman filter. The actual pose correction is calculated by minimizing the distance from the control point to the actual image edge found. The pose of the object is tracked over time using a Kalman filter. A constant velocity model is assumed for the object.

Kollnig and Nagel [13] match a synthetic gradient image directly to the gray level gradient image from the video frames. The difference between the synthetic gradient image and gray value gradient of current frame is used to update the 3D pose of the model using a MAP estimator. A Kalman filter stabilizes the tracking. The edges are models with 2D Gaussians and MAP estimator utilizes Gauss-Newton method modified with Levenberg-Marquart iteration.

N. Giordana et. al. [14] introduce a 2D model based tracking. The control points on the contour are determined from the 3D model. These points form a polyhedral shape, which is assumed to correctly model the object appearance in the image. The estimation is done by the minimization of a Bayesian criterion, which is composed of two terms, one for the minimization of the differences from the model gradients and the frame gradients, one for the deformations from the model shape. For gradient matching a gradient optimization algorithm is used. For overall minimization simulated annealing is utilized.

Eric Marchand et. al. extend the work in [14], with additional paradigms [15]. The future point matching is accomplished by ME method which calculates the gradients in the direction of a line between two future points of the 3D shape model. The minimization step consists of two steps for affine transformation and the 3D model gradient match. For optimization an explicit discrete search algorithm is applied.

Moon et al. in [16] apply Condensation algorithm for 3D face and eye gaze tracking. They consider the problem as nonlinear state estimation and apply a special version of Kalman filter with branching particle propagation. They consider the approach as successful but do not give a measure of computational complexity. They employ about 200 particles.

The remainder of the paper is organized as follows. The Section 2 explains the nature of the problem generally. In Section 3, the model of the system is introduced, and the state to be estimated is detailed. Section 4 briefly gives the principals of perspective camera model. In Section 5, Condensation algorithm is explained and its implementation is detailed. Section 6 compares the optimization-based approaches and statistical solution. In Section 7, experiments are given and the results are shown. Section 8 contains the conclusions and the future work to be done.

2 Problem Statement

Tracking is an application area of estimation theory. The problem is estimating the state of a system from a set of observations. Given a state transition function and a measurement function, the estimators try to determine the state of the system with minimum error.

Let us assume that the process is governed by the following non-linear difference equations,

$$x_k = f(x_{k-1}, u_{k-1}, w_{k-1}) \tag{1}$$

$$y_k = h(x_k, v_k) \tag{2}$$

where, w_k and v_k random processes are stochastically independent. The aim of the estimation is recursively determine the $p(x_k|y_{1:k})$ probability distribution, where the measurements $y_{1:k} = \{y_1, y_2, ..., y_k\}$ are known. A recursive estimator will solve the problem in two stages: prediction and update. In prediction step, $p(x_{k-1}|y_{1:k-1})$ is propagated to future via equation (1). The update stage will utilize Bayes rule for minimizing the prediction error considering new observation such as

$$p(x_k|y_{1:k}) = \frac{p(y_k|x_k).p(x_k|y_{1:k-1})}{p(y_{k|}|y_{1:k-1})}. \tag{3}$$

The expected value of the state can be calculated considering $p(x_k|y_{1:k})$by

$$E[x_k|y_{1:k}] = \int x_k.p(x_k|y_{1:k})dx_k \tag{4}$$

A recursive solution with prediction and update stages will solve the problem iteratively. In the non-linear and non-Gaussian system conditions, probabilistic simulation methods are more convenient for the solution.

3 Model

In order to estimate the system behavior, we need to have a prior knowledge about the inner mechanism of the system. This means, we need to model the behavior of the system approximately. Considering the motion of a vehicle license plate in real world, it is trivial that the motion is in three-dimensional space with many alignment possibilities. Defining a 3D location in Cartesian coordinates, we need three dimensions and for determining the alignment of a 3D object in space, we will need three angles such as

$$[x\ y\ z\ \alpha\ \beta\ \gamma]^T \tag{5}$$

where, α is the rotation around x axis, β is the rotation around y axis and finally γ is the rotation around z axis.

Beside the location and alignment of a 3D object, we need to consider its motion in 3D space, for which we will need to define a magnitude and two angles in spherical coordinates such as

$$[r\ \phi\ \theta]^T \tag{6}$$

where, r is the velocity, θ is the azimuthal angle in xy-plane from x-axis, ϕ is the polar angle from z-axis. So the total state vector of our system will have nine variables. On the other hand, for a planar object, which can move in the direction of its normal only as shown in Fig 1, the state vector can be represented with only seven variables because two variables for pose and motion combine. The state vector can be shown as.

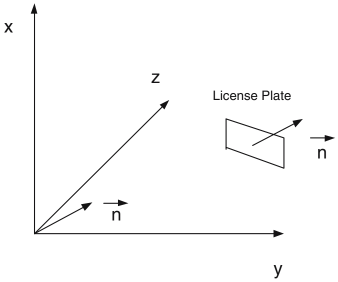


Fig. 1. License plate motion model

$$[x\ y\ z\ r\ \alpha\ \beta\ \gamma]^T \tag{7}$$

The system dynamic equation is non-linear such as

$$x_{k+1} = f(x_k) + w_k \tag{8}$$

which defines the motion in 3D. The measurement equation is also nonlinear such as,

$$y_k = h(x_k) + v_k \tag{9}$$

where w_k and v_k are Gaussian process and measurement noise. The function $f(x_k)$ contains trigonometric functions, while the function $h(x_k)$ is defined by the camera perspective transformation. The system is modeled with first order motion equation, which assumes constant velocity, spherical motion with constant $[\alpha\ \beta\ \gamma]^T$angles. In contrast with the work in [12,13,14,15], we utilize Condensation algorithm for estimating the state of the system. This work is a novel usage of Condensation algorithm for such a tracking problem.

4 Perspective Camera Model

The camera model defines the $h(x_k)$ function. The model contains two types of parameters. The extrinsic parameters are the parameters that define the location and orientation of the camera reference frame with respect to a known world reference frame. The intrinsic parameters are the parameters necessary to link the pixel coordinates of an image point with the corresponding coordinates in the camera reference frame. When both intrinsic and extrinsic camera parameters are known, the full camera projection matrix M is determined. The projection is than just a matrix multiplication defined as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = M. \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} . \tag{10}$$

By this projection the real world coordinates $X\ Y\ Z$ are transformed into image coordinates $x\ y\ z$ and the pixel coordinates on the image p_x, p_y can be calculated by $p_x = x/z$ and $p_y = y/z$.

We utilize full perspective model, where the camera and space coordinates overlap and the 3D origin is the origin of the image plane. The intrinsic parameters are determined by the camera vendor. Depending on the utilized camera lenses, a distortion model can be necessary. Especially radial distortion has significant effects if wide angle lenses are used. We have not utilized any distortion model in our experiments.

5 Condensation Algorithm

The standard Condensation algorithm can be summarized as follows:

1. Initialization: t = 0
   For n = 1...N generate samples from the prior in order to obtain $\{s_0^n, \pi_0^n\}$ where s' are the samples and $\pi's$ are the weights assigned to each sample.
2. Iterate for t=0,1,2,...
   At time step t+1 , construct the n^{th} of N samples as follows:
   (a) Propagate samples using state transition equation to obtain $p(x_{t+1}|y_{1:t})$. From the sample set at time t , where the samples correspond to the location, pose and motion of the license plate:

$$\{s_t^n, \pi_t^n\}, n = 1, 2, ..., N$$

   The new sample set $\{s_{t+1}^n, \pi_{t+1}^n\}, n = 1, 2, ..., N$ is composed according to the equations:

$$s_t^n = \left\{ \begin{matrix} x_t \\ y_t \\ z_t \\ r_t \\ \alpha_t \\ \beta_t \\ \gamma_t \end{matrix} \right\} \qquad s_{t+1}^n = s_t^n + \left(\begin{matrix} x_t + r_t.sin\beta_t \\ y_t - r_t.sin\alpha_t.cos\beta_t \\ z_t + r_t.cos\alpha_t.cos\gamma_t \\ r_t \\ \alpha_t \\ \beta_t \\ \gamma_t \end{matrix} \right) + N(0,\sigma) \quad (11)$$

(b) Calculate the new weights by:

$$\pi_{t+1}^n = \pi_t^n.p(y_{t+1}^n | s_{t+1}^n)$$

(c) Store samples $\{s_{t+1}^n, \pi_{t+1}^n, c_{t+1}^n\}, n = 1, 2, ..., N$ where c_{t+1}^n are the cumulative probabilities given by:

$$c_{t+1}^0 = 0$$

$$c_{t+1}^n = c_{t+1}^{n-1} + \pi_{t+1}^n$$

(d) Normalize by dividing all cumulative probabilities $c_{t+1}^n = c_{t+1}^n / c_{t+1}^N$, i.e. so that $c_{t+1}^N = 1$ and weights $\pi_{t+1}^n = \pi_{t+1}^n / c_{t+1}^N$, so that $\sum_n \pi_{t+1}^n = 1$.

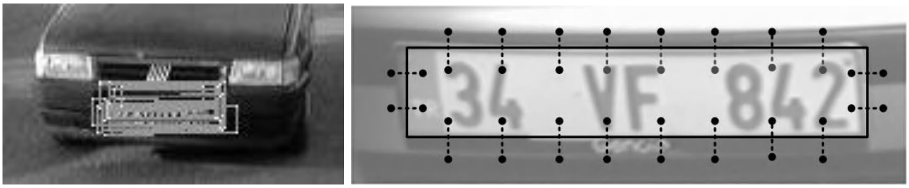
(e) Resample the samples s_{t+1}^n with probability π_{t+1}^n to obtain N samples. For this purpose, generate a random number $r \in [0, 1]$, uniformly distributed. Find the smallest n for which $c_{t+1}^n \geq r$. Add this sample to the new set $\{s_{t+1}^m, \pi_{t+1}^m, c_{t+1}^m\}, m = 1, 2, ..., N$

Initialization is performed when the tracker is started to provide an initial estimate of the position of the target object.

6 Framework

Our approach is a novel method for rigid object tracking. We utilize Condensation algorithm for estimating the location of the rigid object in 3D. This simulation based approach is an alternative to optimization based 3D tracking methods [12,13,14,15]. In optimization based methods the tracking process is accomplished by the minimization of an energy function.

On the contrary, our approach has a probabilistic framework, where the prior and posterior probability distributions are represented by samples. Each sample is a vector, composed of seven variables for the location, alignment and the velocity of the object in 3D as explained in Section 3. Each sample can be considered as a guess for the license plate location, alignment and motion. The samples are projected on to the video frame in Fig 2a. The samples and the associated probabilities approximate the prior and the posterior probability distributions. The probability, associated to each sample is calculated by the likelihood function. We use the directed image gradients for determining the probabilities of each sample $\{\pi_t^n, s_t^n\}$ in Condensation algorithm. The directed gradients are calculated by taking the difference vertical to the proposed license plate borders as



(a) (b)

Fig. 2. a) Samples b) Directed gradients along the object boundaries

shown in Fig 2b. We do not compute the gradients in whole image, thus the computational complexity is reduced significantly. The feature extraction process is reduced to simple subtractions of pixel values.

The likelihood of a sample is calculated as follows,

$$\pi_{likelihood} = number\ of\ gradients\ ,where\ \|\overrightarrow{g_n}\| > C_{threshold} \tag{12}$$

The most significant drawback of the algorithm is sample impoverishment. Several techniques have been proposed for this aim [11]. We follow a similar method suggested by Li et al [10]. The velocity of the license plate r_t is estimated by a Kalman filter individually. For each sample, a separate Kalman filter is utilized. The state vector for Kalman filter is composed of velocity and acceleration magnitudes of the license plate. The transition equation is as flows,

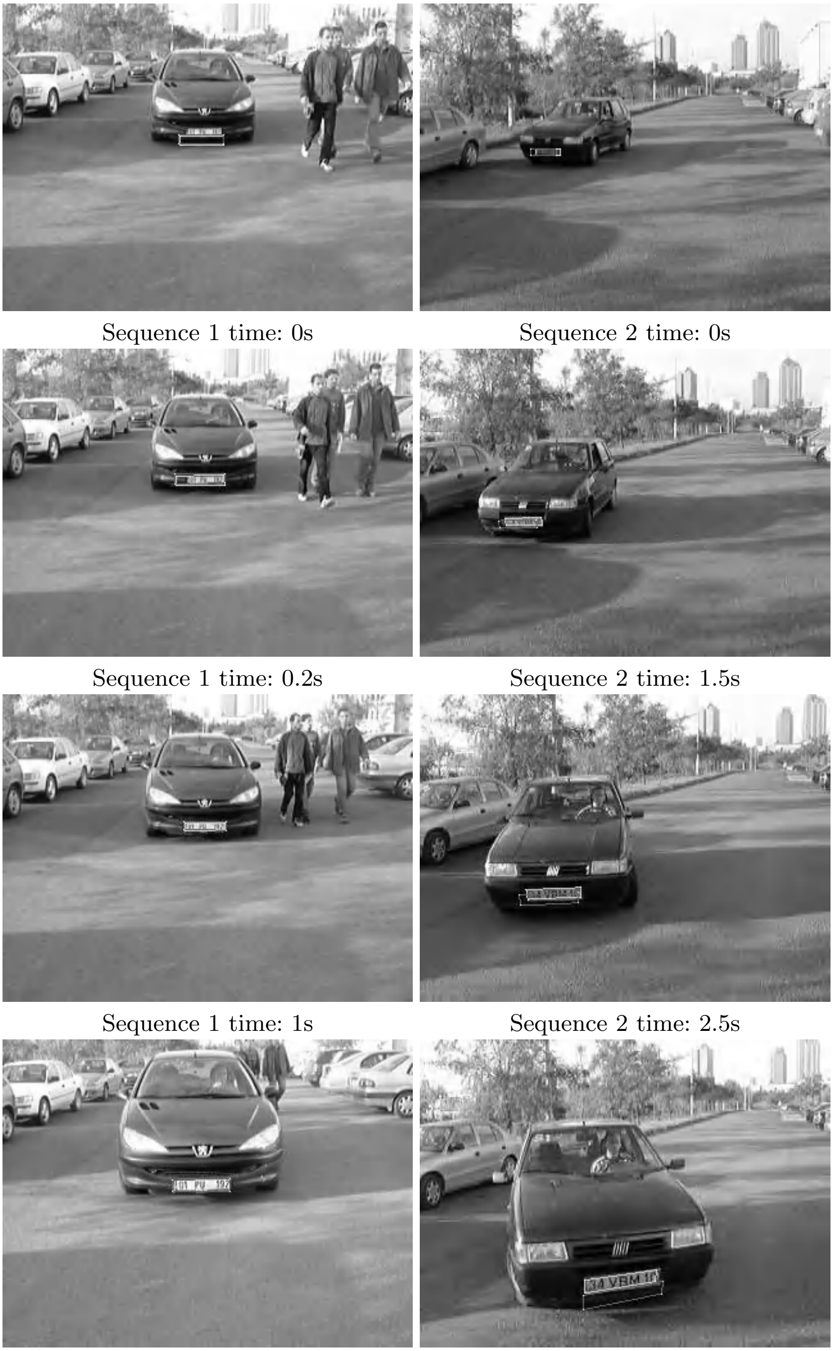
$$\begin{pmatrix} v_{k+1} \\ a_{k+1} \end{pmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} v_k \\ a_k \end{pmatrix} + N(0, \sigma) \tag{13}$$

where T is the sampling period, v_k is the velocity and a_k is the acceleration at time step k.

1. Before the resampling step of Condensation algorithm (after normalization in 2.d Section 5), expected value of the license plate location is stored in an array $\{z_0, z_1, ..., z_n\}$, where z_n is the location of the license plate in 3D.
2. The measurement value y_k^* for each Kalman filter, is obtained from the distance between the location vectors $y_k^* = \|z_k - z_{k-1}\|$. The measurement update formula of Kalman filter is given as,

$$x_{k|k} = x_{k|k-1} + K_k(y_k^* - H.x_{k|k-1}). \tag{14}$$

3. In resampling step, for each sample, the velocity values are updated by drawing a sample from a Gaussian defined by their Kalman filter mean and covariance.
4. The samples with greater likelihood values are selected several times for the next iteration as a result of Factored sampling. By this technique, same location and orientation values are combined with different velocity values. It is practical to set the Kalman filter covariance of the cloned samples to initial values.



Sequence 1 time: 0s

Sequence 2 time: 0s

Sequence 1 time: 0.2s

Sequence 2 time: 1.5s

Sequence 1 time: 1s

Sequence 2 time: 2.5s

Sequence 1 time: 4s

Sequence 2 time: 3.1s

Fig. 3. Results

7 Experiments

In order to evaluate the performance of the proposed approach, we obtained videos from outdoor environment with different motion trajectories. The simplest scenario is shown as sequence 1 in Fig 4. In sequence 1, a car is approaching to the camera without any maneuver. In this sequence the initial state of the license plate is wrongly initialized. The algorithm is robust to wrong initialization; according to the initialization covariance of the samples, the algorithm can tolerate wrong initialization. There is a trade off in defining the initial covariance. If it is defined too large, the algorithm can converge to a wrong location, which is not the license plate. In sequence 1, the utilized number of samples is 128.

The algorithm can also track severe maneuvers. The sequence 2 in Fig 3 shows an example of such a situation. For this example, the number of particles is needed to be increased up to 512.

The algorithm is implemented with Visual C++ and runs on a Pentium 4 2GHz PC. There are two rectangles drawn on the frames in Fig 4; the white rectangle is the filtered position after re-sampling step while the gray one is the estimation for the next frame after propagation of the samples.

8 Conclusions

The Condensation algorithm is applied to rigid object tracking in this paper. Given an initial estimate, the algorithm tries to track the location, motion vector and pose of the object in 3D in the successive video frames. This approach can track a vehicle license plate successfully even with strong maneuver situations. The Condensation algorithm represents the posterior and prior distributions for the state of the license plate.

We utilize a Kalman filter for estimating the velocity of the license plate individually. Utilizing Kalman Filter avoids the sample impoverishment, where the velocity values can be scattered with high variance. Instead of re-sampling the velocities in Condensation algorithm, we draw the velocity values from a Gaussian, whose mean and variance is taken from the Kalman filter. Future research will be focused on reducing the sample numbers and applying the approach to more complex articulated objects.

References

1. Kalman, R. E. : A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME Journal of Basic Engineering Series D 82 (1960) 35-45
2. Moon, J. R. , Stevens, C. F. : An Approximate Linearisation Approach to Bearings-only Tracking. IEEE Target Tracking and Data Fusion Colloquium Digest 96/253 (1996) 8-18
3. Alspach, D. L. , Sorenson, H. W. : Non-linear Bayesian Estimation using Gaussian Sum Approximation. IEEE Trans. Auto. Control (1972) 439-447

4. Masreliez, C. J. : Approximate non-Gaussian Filtering with Linear State and Observation Relations. IEEE Trans. Auto. Control (1975) 107-110
5. West, M. , Harrison, P. J. , Migon, H. S. : Dynamic Generalized Linear Models and Bayesian Forecasting (with Discussion). J. Am. Statist. Assoc. (1985) 73-97
6. Bucy, R. S. : Bayes Theorem and Digital Realiasation for Nonlinear Filters. Journal of Austronautical Science (1969) 80-94
7. Fearnhead, P. : Sequential Monte Carlo Methods in Filter Theory. Merton College University of Oxford PhD Thesis (1998)
8. Isard, M. , Blake, A. : Condensation-Conditional Density Propagation for Visual Tracking. Int. J. Computer Vision vol. 29 no. 1 (1998) 5-28
9. Isard, M. , Blake, A. : ICondensation: Unifying Low-level and Highlevel Tracking in a Stochastic Framework. Proc. 5th European Conf. Computer Vision (1998) 893-908
10. Li, P. , Zhang, T. : Visual Contour Tracking based on Sequential Importance Sampling/Resampling Algorithm. Pattern Recognition (2002) Proceedings 16th International Conference on Volume: 2 11-15 Aug. (2002) 564-568
11. Li, P. , Zhang, T. , Pece, A. : Visual Contour Tracking based on Particle Filters. Image and Vision Computing vol. 21 no. 1 (2003) 111-123
12. Harris, C. J. : Tracking with rigid models. In: Blake, A. , Yuille, A. (eds.) : Active Vision. MIT Press Cambridge MA (1992)
13. Kollnig, H. , Nagel, H.H. : 3D Pose Estimation by Fitting Image Gradients Directly to Polyhedral Models. Intelligent Vehicles Symposium Proceedings (1995)
14. Giordana, N. , Bouthemy, P. , Chaumette, F. , Spindler, F. , Bordas, J. C. , Just, V. : 2d Model-based Tracking of Complex Shapes for Visual Servoing Tasks. In: Hager, G. , Vincze, M. (eds.) : IEEE Workshop on Robust Vision for Vision-Based control of Motion Leuven Belgium (1998)
15. Marchand, E. , Bouthemy, P. , Chaumette, F. , Moreau, V. : IEEE Robust Realtime Visual Tracking us-ing a 2D-3D Model-based Approach. International Conference on Computer Vision ICCV99 Volume 1 Kerkira Greece (1999) 262-268
16. Moon H. , Chellappa, R. , Rosenfeld, A. : 3D Tracking Using Shape-Encoded Particle Propagation. ICCV 2001 proceedings of 8th IEEE Int. Conference Vol 2 (2001)

A Multi-view Approach to Object Tracking in a Cluttered Scene Using Memory

Hang-Bong Kang and Sang-Hyun Cho

Dept. of Computer Engineering, Catholic University of Korea,
#43-1 Yokkok 2-dong Wonmi-Gu, Puchon City Kyonggi-Do, Korea
hbkang@catholic.ac.kr

Abstract. In this paper, we propose a new multi-view approach to object tracking method that adapts itself to suddenly changing appearance. The proposed method is based on color-based particle filtering. A short-term memory and a global appearance memory are introduced to handle sudden appearance changes and occlusions of the object of interest in multi-camera environments. A new target model update method is implemented for multiple camera views. Our method is robust and versatile for a modest computational cost. Desirable tracking results are obtained.

1 Introduction

Object tracking is of interest for a variety of applications such as video surveillance and monitoring systems. In particular, in order to monitor a site effectively, it is desirable to use multiple limited field of view cameras.

Various object tracking algorithms in multi-camera environments have been developed. Comaniciu et al.[1] proposed a flexible multi-camera system for real-time tracking. Kahn et al.[2] suggested a system for people tracking in multiple uncalibrated cameras. They used spatial relationships between camera fields of view to correspond different views of the same person. Triveldi et al.[3] developed 3D tracking system which operated with multiple cameras. Nummiaro et al.[4] proposed object tracking system in multi-camera environments. In their approach, best view was automatically selected for a virtual classroom application.

To deal with sudden changes in 3D pose and occlusions, we propose a new object tracking method in multi-camera environments using memory concepts. If new appearance of the specific object of interest is observed in any camera, we store it into local short-term memory. We also copy the various appearances in each short-term memory into the global appearance memory. The appearance models in the short-term memory are referred to whenever an estimated object state in current frame needs to be determined whether it is new appearance or not. When the estimated state in any camera is new, it is also compared to other appearance models in the global appearance memory in order to determine whether occlusion has occurred or not. The severely occluded object state is not stored in the short-term memory. The novelty of the proposed approach mainly lies in its adaptation to sudden appearance changes of the object of interest in multi-camera environments.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 870 – 879, 2005.

The paper is organized as follows. Section 2 discusses memory-based object tracking method. Section 3 presents our proposed tracking method in multi-camera environments. Section 4 shows experimental results of our proposed method in object tracking.

2 Memory-Based Object Tracking

In this section, we will present our memory-based tracking method to handle sudden appearance changes or occlusions.

2.1 Short-Term Memory Model

To handle sudden appearance changes of the object of interest, we use a short-term memory-based method in updating the target model during tracking. Figure 1 shows a causal FIFO short-term memory based model. The model consists of a memory buffer and an attention span. The memory buffer keeps the old reference (or target) models and the attention span maintains a number of recent reference models [5].

At the initialization step of short-term memory, a copy of the target model in the tracking process enters the memory as a reference model R_i and is placed on the attention span. The estimated target object is compared to the reference models in the attention span. When the similarity value between the estimated target object and one of the reference models in the attention span is lower than the threshold value, the estimated target object enters the short-term memory as a new reference model. Old reference models shift into the next position in the memory and the oldest one is removed from the memory.

Otherwise, the estimated target object is updated using the current reference model in the attention span. By using the short-term memory, the target model is updated appropriately in the case of sudden appearance changes.

2.2 Memory Model in Multi-camera Environments

We extend this memory concept into multi-camera environments to robustly track the object of interest regardless of appearance changes and severe occlusions. In multi-camera environments, it is possible to track the object of interest well even though the objects are severely occluded or lost due to handoff in some camera views. Figure 2 shows the multi-camera environments. Local short-term memory is used for appearance changes in every camera and global appearance memory is used to handle occlusions.

Global appearance memory (GAM) keeps copies of the reference models in the attention span of each local short-term memory. Similar to the local short-term memory, it has FIFO architecture and the oldest one is removed from the GAM. At the initialization step, initial reference models from the local short term memory of all the cameras enter GAM. When the estimated target object is introduced and evaluated as a new appearance model in the local short-term memory, having low similarity value with local reference models, it is then compared to the reference models in GAM by computing the similarity value. If the similarity value here is lower than the threshold value, the estimated target object does not enter the local short term memory because it may have been generated from severe occlusion in that camera view. The reason to compare the new appearance model with reference models in GAM is that various appearances for the object of interest are kept in GAM so that we can prove the face that the appearance model may have been generated from occlusion when the appearance model's similarity value is very low. So, the fact that the appearance model's similarity value is very low means that the appearance model may be generated from occlusion. In this circumstance, the virtual target state is estimated at the intersection of epipolar lines from other camera views. The handoff situation is detected from Field of View (FOV) of every camera. When the handoff has occurred temporary in one camera, the target model update in that camera is suspended.

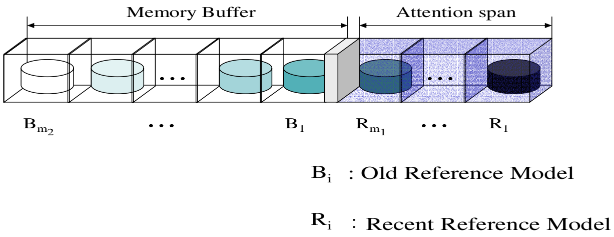


Fig. 1. Short-term Memory Model

3 Face Tracking in Multi-camera Environments

In this section, we will present how to implement our proposed face tracking system in multi-camera environments. Each camera has limited FOVs which are overlapping.

3.1 Initialization

For face tracking, we detect a face from each camera input using Viola and John's method [6]. If a face is detected in one camera, we try to find corresponding object in other camera views using epipolar geometry. Epipolar lines between camera views are calculated using corresponding points and then the corresponding object's position is estimated along the epipolar lines. If more than one face is detected in more than one camera, we determine whether these faces are the same face or not by intersecting epipolar lines. In other words, for each face of interest in one camera view, we detect the corresponding face in other camera by searching candidate along the epipolar line.

3.2 Color-Based Particle Filter

Our tracking method is based on color-based particle filter [7]. We define sample state vector **s** as

$$s = \{x, y, E_x, E_{y,} k\} \tag{1}$$

where *x, y* designate the location of the ellipse, E_x ,E_y the length of the half axes and *k* the corresponding scale change. The dynamic model can be represented as

$$s_t = As_{t-1} + r_{t-1} \tag{2}$$

where ***A*** defines the deterministic component of the model and $\boldsymbol{r}_{t-1}$ is a multivariate Gaussian random variables.

As target models, we use color distributions because they achieve robustness against non-rigidity, rotation and partial occlusion. The color distribution $Py = \{p_y^{(u)}\}_{t=1,..,m}$ at location y is calculated as

$$p_y^{(u)} = f \sum_{i=1}^{I} w\left(\frac{\|\mathbf{y} - \mathbf{x}_\mathbf{i}\|}{k}\right)\delta\left[h(\mathbf{x}_\mathbf{i}) - u\right] \tag{3}$$

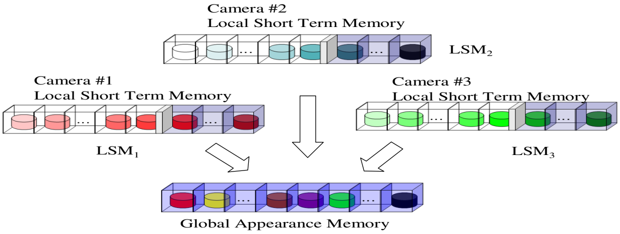


Fig. 2. Global Appearance Memory

where I is the number of pixels in the region, δ is the Kronecker delta function, $k = \sqrt{E_x^2 + E_y^2}$ is used to adapt the size of ellipse, f is normalization factor and w is the weighting function such that smaller weights are assigned to the pixels that are further away from the region center.

As in color-based particle filter [7], the tracker selects the samples from the sample distribution of the previous frame, and predicts new sample positions in the current frame. After that, it measures the observation weights of the predicted samples. The weights are computed using a Gaussian with variance σ:

$$\lambda^{(n)} = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{d^2}{2\sigma^2}} \tag{4}$$

where d is the Bhattacharya distance. Bhattacharyya distance is computed from Bhattacharya coefficient which is a popular measure between two distributions p(u) and q(u). Considering discrete densities such as our color histogram $p=\{p^{(u)}\}_{u=1,..,m}$ and $q=\{q^{(u)}\}_{u=1,..,m}$ the coefficient is defined as

$$\rho[p,q] = \sum_{u=1}^{m} \sqrt{p^{(u)}q^{(u)}} \tag{5}$$

and the Bhattacharyya distance is

$$d = \sqrt{1 - \rho[p,q]} \tag{6}$$

The estimated state at each time step is computed by

$$E(S) = \sum_{n=1}^{N} \lambda^{(n)} s^{(n)} \tag{7}$$

3.3 Target Model Updating in Multi-camera Environments

To update target model in multi-camera environments, we compute similarity value using Bhattacharyya coefficient. The Bhattacharyya coefficient in kth camera is represented as

$$\rho[\mathrm{p}^{\mathrm{k}}_{\mathrm{E[S_t]}}, \mathrm{q}^{\mathrm{k}}_{\mathrm{t-1}}] \tag{8}$$

```
GAM : $m_1$-size Global Appearance Memory
LSM$_k$ : $m_2$-size Local Short-term Memory of k$^{th}$ Camera
    For mean state color distribution $p^k_{E[S_t]}$, target model $q^k_{t-1}$.
  (a)calculate $\rho[p^k_{E[S_t]}, q^k_{t-1}]$.
  (b) detect an occlusion or a sudden appearance change
  case 1: Occlusion
  if $\rho[p^k_{E[S_t]}, q^k_{t-1}] < T_1$ and $\rho[p^{k'}_{E[S_t]}, GM[i]] < T_1$ ($T_1$ : Threshold,
          $k' \neq k,\ i = 1,..,m_1$)
      Calculate the intersections of epipolar lines
      virtual target location = intersection point

  case 2: sudden appearance change
  if $T_1 < \rho[p^k_{E[S_t]}, q^k_{t-1}] < T_2$ and $T_1 < \rho[p^{k'}_{E[S_t]}, GM[i]] < T_2$
          ($T_1, T_2$ : Threshold, $k' \neq k,\ i = 1,..,m_1$)
      $q^k_t = \arg\max_{B_k[i]} \rho[p^k_{E[S_t]}, LSM_k[i]]$
      repeat  $LSM_k[n+1] \leftarrow LSM_k[n]$    $n = 1,....,m_2 - 1$
      $LSM_k[0] \leftarrow p^k_{E[S_t]}$
      repeat  $GAM[n+1] \leftarrow GAM[n]$    $n = 1,....,m_1 - 1$
      $GAM[0] \leftarrow p^k_{E[S_t]}$

  case 3: reference target model update
      $q^k_t = (1-\alpha) q^k_{t-1} + \alpha P^k_{E[S_t]}$
        for each bin u where $\alpha$ weights the contribution
        of the mean state histogram $p^k_{E[S_t]}$
```

Fig. 3. Target model update algorithm in multi-camera environments

We use this coefficient to classify three conditions such as occlusion, sudden appearance change, and normal update cases. When the Bhattacharyya coefficient is very low, the target estimate is compared to reference models in GAM. If the computed Bhattacharyya coefficient is lower than the threshold value, we decide the appearance model is generated from occlusion. In the case of occlusion as shown in Figure 3, we calculate the intersection point of epipolar lines to locate virtual target object. In this situation, no appearance model enters the local short-term memory. In the case of sudden appearance changes like Figure 3, new appearance model enteres the local short-term memory and the copy is also maintained in the GAM. Otherwise, the target model is updated like

$$q_t^k = (1-\alpha)q_{t-1}^k + \alpha p_{E(s)}^k \tag{9}$$

where α weights the contribution of the estimate state histogram $P^k_{E(s)}$.

4 Experimental Results

Our proposed face tracking method in multiple camera environments is implemented on a P4-2.4Ghz system with 320*240 image size. We made several experiments in a variety of environments to show the robustness of our proposed method. Figure 4 shows our camera set up situation. Three cameras (left, center, and right) are used and camera FOVs are overlapping. Figure 4(b) shows the center camera view with other camera FOVs.

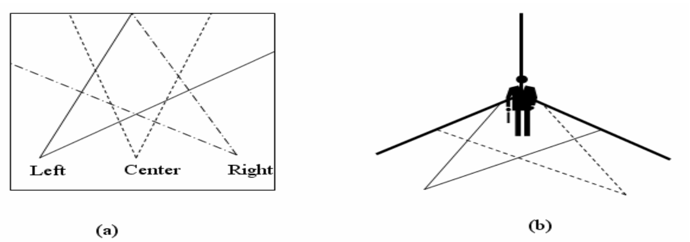


Fig. 4. Camera Setup

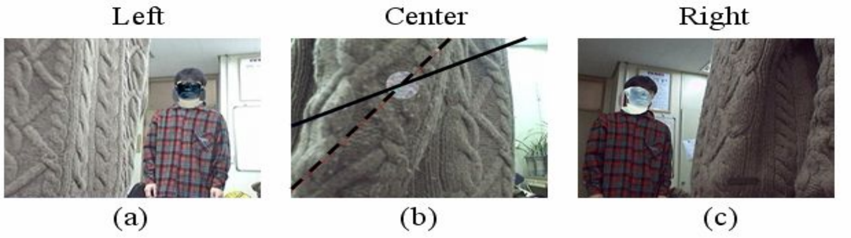


Fig. 5. Epipolar line Intersection

From each camera, faces are detected using Viola and Jones' method [6]. At the initialization step, we determine correspondences between detected faces in each camera using epipolar geometry. Each camera then determines its own target models. Figure 5 shows epipolar lines from other camera views. In Figure 5(b), the target object is occluded by another person. So, the virtual target object is estimated from the intersection of epipolar lines. The solid epipolar line is computed from center view and left view cameras. The dotted epipolar line is computed between center view and right view cameras. As shown in Figure 5(b), the virtual target object position is estimated from the intersection point of epipolar lines. Based on the experiments, the threshold values T_1 and T_2 shown in Figure 3 were set to 0.8 and 0.6, respectively. The size of total memory and the attention span of the local short-term memory were 7 and 3, respectively.

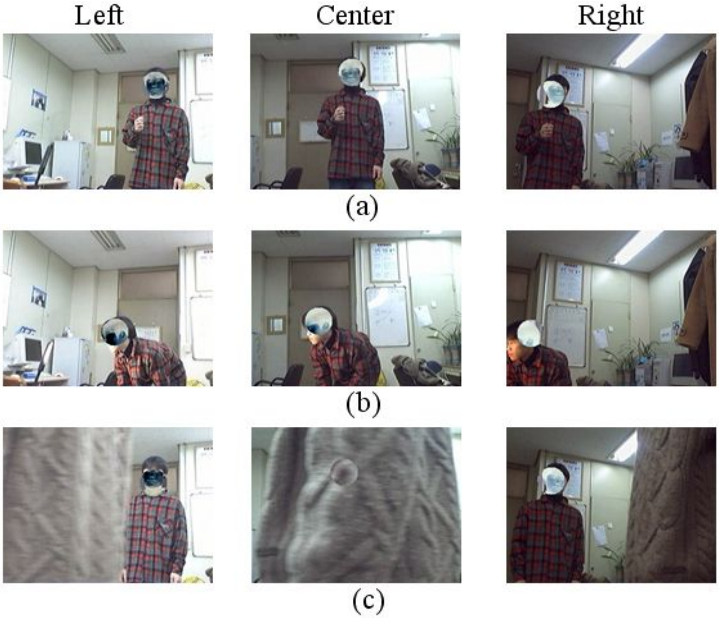


Fig. 6. Experimental Result

We experiment using three video sequences. Figure 6(a) shows the face object in each camera. The object is detected and determined as the same object by computing epipolar lines. In Figure 6 (b), lighting condition has changed because of the stand lamp. The right view is not good because of lightness changes, but the other two views are good enough. So, we can track the object of interest well in spite of illumination changes. Figure 6(c) shows that the object is occluded by other person in the center view. This is determined as the occlusion case based on our algorithm shown in Figure 3. Virtual object position is estimated from the intersection of epipolar lines. Figure 7 shows the object tracking results using particle filter and our proposed method. In the particle filter method like Figure 7(a), the abrupt appearance

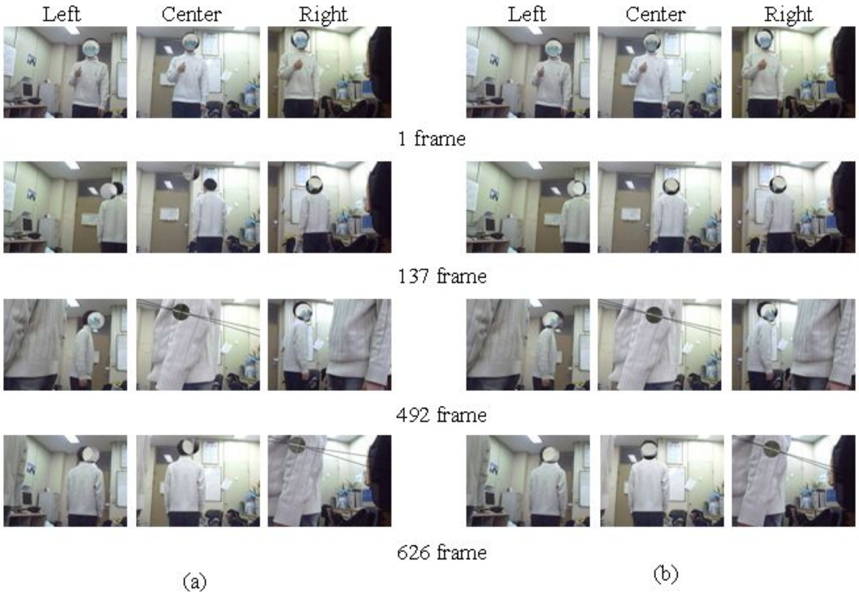
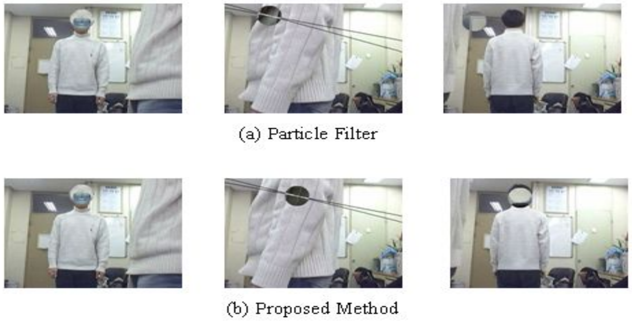


Fig. 7. Experimental Result. (a) particle filter, (b) proposed method


(a) Particle Filter

(b) Proposed Method

Fig. 8. Experimental Result.(a) particle filter, (b)proposed method

changes at 137th frame cause problems in tracking object of interest. However, as shown in Figure 7(b), our proposed method shows accurate tracking result regardless of sudden appearance changes because various target models are maintained in GAM. In Figure 8(a), the virtual object in the center view is not correctly estimated because of the wrong position of the target in the right camera view. However, our method like Figure 8(b) shows accurate tracking result. Figure 9 shows multiple object tracking results. Even though multiple objects of interest are occluded or have sudden appearance changes, tracking is executed accurately.

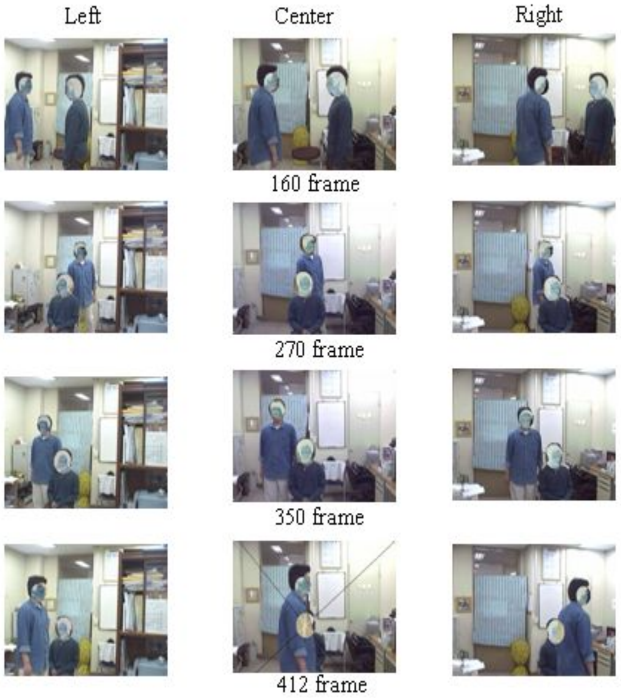


Fig. 9. Multiple Object Tracking Result

5 Conclusions

In this paper, we proposed a novel tracking method to handle sudden appearance changes and occlusions in multi-camera environments. A short-term memory and global appearance memory model are proposed to keep different appearance models in the tracking process. New target update methods in multi-camera environments are also designed. We performed face tracking experiments in various situations and the proposed tracking method showed good results even when sudden appearance changes occurred. Compared with other algorithms, our proposed system shows a better and more robust tracking performance.

Acknowledgement

This work was supported by the Catholic University of Korea Research Fund 2004.

References

1. Comaniciu, D., Berton, F. and Ramesh, V.: Adaptive Resolution System for Distributed Surveillance. Real-Time Imaging, Vol. 8 (2002) 427-437
2. Kahn, S., Javed, O. and Shah, M.: Tracking in Uncalibrated Cameras with Overlapping Field of View. PETS (2001)

3. Trivedi, M., Mikic, I., and Bhonsle, S.: Active Camera Networks and Semantic Event Databased for Intelligent Environments. Proc. IEEE Workshop on Human Modelling, Analysis and Synthesis (2000)
4. Nummiaro, K., Koller-Meier, E., Svoboda, T.: Color-Based Object Tracking in Multi-Camera Environments, LNCS 2781 (2003) 591-599
5. Kang, H. and Cho, S.: Short-term Memory-based Object Tracking, LNCS 3212 (2004) 597-605
6. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. Proc. CVPR '01 (2001)
7. Nummiaro, K., Koller-Meier, E. and Van Gool, L.: A Color-Based Particle Filter. First International Workshop on Generative-Model-Based Vision, in Conjunction with ECCV'02 (2002) 53-60

A Robust 3D Feature-Based People Stereo Tracking Algorithm*

Guang Tian[1], Feihu Qi[1], Yong Fang[1], Masatoshi Kimachi[2], Yue Wu[2],
Takashi Iketani[2], Xin Mao[1], and Panjun Chen[1]

[1] Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, China
Tianguang@cs.sjtu.edu.cn
[2] OMRON Corporation, Japan

Abstract. This paper presents a 3D feature-based people tracking algorithm which combines an interacting multiple model (IMM) algorithm with a cascade multiple feature data association algorithm. The IMM algorithm in this paper only uses an adaptive Kalman Filter and two dynamic models consisting of a constant velocity model (CV) and a current statistics model (CS) to predict the 3D location of people maneuvering and update the prediction with corresponding measurement. The cascade multiple feature data association algorithm in this paper utilizes three hypotheses, including the nearest distance hypothesis, the velocity consistency hypothesis, and the intensity consistency hypothesis, in turn to determine which trajectory a measurement should be assigned to. Experimental results demonstrate the robustness and efficiency of the proposed framework. It is real-time and not sensitive to the variable frame to frame interval time. It also can deal with the occlusion of people and do well in those cases that people rotate and wriggle.

1 Introduction

With recent advances of computer technology the visual tracking has become a popular area for research and development because it has important application in real time world, including visual surveillance, multimedia, computer games, film, and video indexing. The visual tracking is usually used to detect and track people and monitor their activities in a given environment. The goal of our project is to track people indoors, such as subway station, shopping center and so on. The task of person tracking is complex in those sites because some times it is very crowded and people are interacting with each other there. It is all known that the trajectory of people is nonlinear [1]. Meanwhile, real time and accuracy are required in our application. The interacting phenomenon of people is more difficult for single view method to deal with in people tracking. In order to solve those problems, some researchers use multi-view approach in their applications

* Our work is supported by SVS 04 of OMRON Corporation, JAPAN and the National Natural Science Founds of China(N.o. 60271033).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 880–889, 2005.

[2]. They generally use the density stereo match to obtain the 3D model of person and have developed many tracking algorithms based on those models. The tracking results by those methods are very good, but it is very difficult to calibrate multiple camera, and too time consuming to be adopted in the real time application. The literatures about 3D feature-based people tracking are few. We only find a paper about a 3D feature-based tracker for multiple objects tracking developed by TANG et al [3]. However, This tracker only can track the rigid object not the people. Moreover, they use the dynamic feature point set of an object to represent the object, and match the dynamic set of an object in the 3D and 2D space in successive frames respectively. It is time consuming too.

We develop a 3D feature-based people tracking system to track people. Our application consists of feature point detection, stereo matching, the feature point of people grouping, and people tracking. In our work, a Susan feature point detection algorithm [4] and robust point stereo matching algorithm are adopted to get 3D feature points of the surveillance region. We classify those 3D feature points into the points of human body and the points belonging to background. Then those 3D feature points in human body are clustered for each person. We choose the centroid of each cluster as the representation of each person. At last we use a 3D feature-based people tracking algorithm that combines the IMM algorithm with the cascade multiple feature data association algorithm to track people. In this paper we only introduce the 3D feature point based people tracking algorithm.

The rest of this paper is organized as follows: in Section 2 we present a 3D feature-based people tracking algorithm which combines the IMM algorithm with the cascade multiple feature data association algorithm. The experimental results are shown in Section 3. In Section 4 we draw a conclusion about our work.

2 Robust 3D Feature-Based Tracking Algorithm

This algorithm is recursive and modular. In each cycle it consists of predicting the next position of each trajectory, matching each measurement with each existing trajectory, and updating the current state of each existing trajectory with its measurement. In our algorithm, we take the centroid of the 3D feature point set belonging to a person as the representation of the person. We divide a 3D surveillance area into two parts in terms of the x and z of camera coordinates (x, y, z), the detection area and the tracking area. Meanwhile, our algorithm tracks the centroid of each person on X-Z plane, i.e. ground plane, because the person can be separated well in this plane. The block diagram of our algorithm is shown in Fig. 1

2.1 Design of IMM Estimator

In our prediction algorithm, IMM is used to predict the state of each person maneuvering and update the state of each person with his corresponding measurement in next frame. We use two dynamic models including the CS model [5]

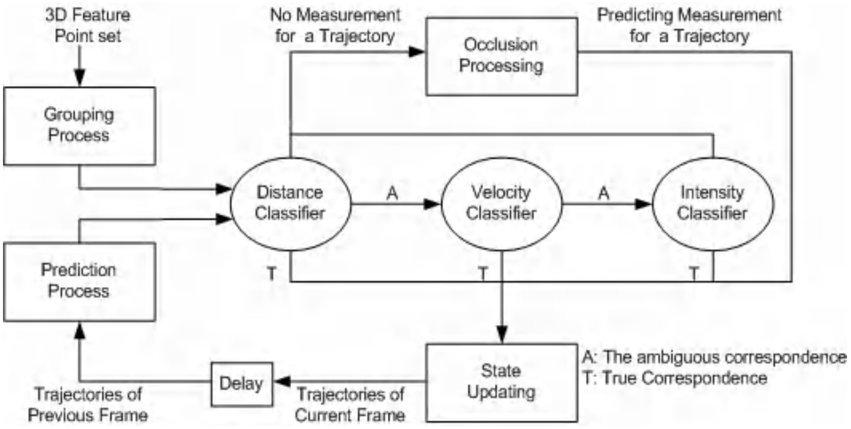


Fig. 1. A 3D feature-based people tracking algorithm

and the CV model, to approximate the people maneuvering. The CS model can track the person who rotates accurately, but is poor in tracking the person who moves in constant velocity. On the contrary, the CV model does well in tracking the person who moves in constant velocity, but poor in tracking the person who rotates. So we combine those two models to obtain an accurate predicting state of each person maneuvering.

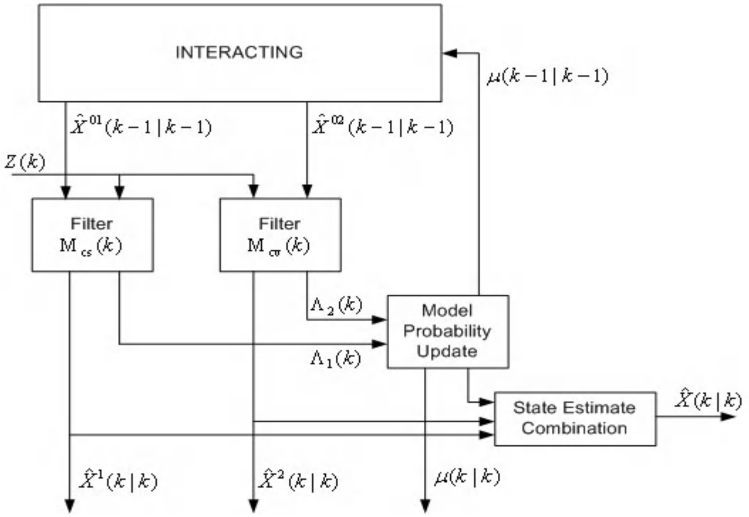


Fig. 2. IMM algorithm

The IMM consists of running a filter for each model, a model probability evaluator, and an estimate combiner at the output of the filters. Each filter uses a mixed estimate at the beginning of each cycle, as illustrated in Fig. 2. An underlying Markov chain is assumed to govern the switching of the models. The details about the IMM and the notations in Fig. 2 can be found in [6] and [7]. Here we only introduce the design of IMM estimator used in our applications. The system dynamic model:

$$X(k) = F[M(k)]X(k-1) + B[M(k)]U(k-1) + V[k-1, M(k)] \ . \tag{1}$$

The measurement model:

$$Z(k) = H[M(k)]X(k) + W[k, M(k)] \ . \tag{2}$$

Where$X(k) = [x, \dot{x}, z, \dot{z}, a_t, a_n]^T$, $Z(k) = [x, z]^T$, $U(t) = [0, 0, 0, 0, c_t\bar{a}_t, c_n\bar{a}_n]^T$, c_t is the time constant of tangential acceleration, c_n is the time constant of normal acceleration, $\bar{a}_t$ is the mean tangential acceleration, $\bar{a}_n$ is the mean normal acceleration. $F[M(k)]$ is the state transition matrix that is the formula of dynamic model. $H[M(k)]$ is the measurement matrix that is the formula of the dynamic model. The model noise vector $V[k-1, M(k)]$ and the measurement noise vector $W[k, M(k)]$ are both zero mean white noise and independent each other.

$$M_{cs} = \begin{pmatrix} 0\,1\,0\,0 & 0 & 0 \\ 0\,0\,0\,0 & d_{11} & d_{12} \\ 0\,0\,0\,1 & 0 & 0 \\ 0\,0\,0\,0 & d_{21} & d_{22} \\ 0\,0\,0\,0 & -c_t & 0 \\ 0\,0\,0\,0 & 0 & -c_n \end{pmatrix}, \ M_{cv} = \begin{pmatrix} 1 & T & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & T & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

Where $d_{11} = \cos\beta$, $d_{12} = -\sin\beta$, $d_{21} = \sin\beta$, $d_{22} = \cos\beta$, $\beta = \arctan(\dot{z}/\dot{x})$. B_{cs} is a 6×6 matrix, all of its elements are zero except b_{55}, $b_{66} \in B_{cs}$, and both two elements are 1. All elements of B_{cv} are zero. All elements of $H(k)$ are zero except h_{11} and h_{23}, and both of them are 1.

In our IMM, the empirical values of the system dynamic model error covariant matrix Q and the measurement error covariant matrix R are adopted in advance. Those values can't describe all real world, and then result in divergence of the filtering computation. In order to ensure the convergence of our algorithm, an adaptive kalman filter for each dynamic model is used in our IMM approach. The iterative formulae are shown as follows.

$$\hat{X}_k = M\hat{X}_{k-1} + K_k(Z_k - HM\hat{X}_{k-1}) \ . \tag{3}$$

$$P_{k|k-1} = SMP_{k-1}M^T + Q \ . \tag{4}$$

$$K_k = P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1} \ . \tag{5}$$

$$P_k = (I - K_kH)P_{k|k-1} \ . \tag{6}$$

where S is amendatory factor.

2.2 Matching Algorithm

In our matching algorithm, three hypotheses have been adopted according to the physical rules, including a nearest distance hypothesis, a velocity consistency hypothesis, and a intensity consistency hypothesis. With those hypotheses, the unique correspondence between a trajectory and a measurement can be ensured in lots of cases. The matching algorithm consists of two stages. One is preselection stage, and the other is selection stage. In the pre-selection stage, the candidates of each trajectory are chosen in terms of the distance in X-Z plane. If the age of a trajectory is less than 4, the current measurements locating in a gate of the previous centroid of a trajectory are considered as the candidates of this trajectory. The reason is that kalman filter will be steady after four times iteration (refer to Fig. 3 and Fig. 4). Otherwise, the current measurements locating in a gate of the predicting location of a trajectory are considered as the candidates of this trajectory. The distance measure is

$$d_1 \quad = \sqrt{(x_i - x_{k-1})^2 + (z_i - z_{k-1})^2},\ i = 0, 1, ..., s \tag{7}$$

$$or$$

$$d_2 = \sqrt{(x_i - x_{k|k-1})^2 + (z_i - z_{k|k-1})^2},\ i = 0, 1, ..., s\ , \tag{8}$$

where the (x_i, z_i) is the projection of the measurement to X-Z plane. The (x_{k-1}, z_{k-1}) is the projection of the previous centroid of the trajectory to $X - Z$ plane. The $(x_{k|k-1}, z_{k|k-1})$ is the projection of the predicting centroid of the trajectory to $X - Z$ plane. The number s is the number of current measurement. d1 is the distance between the (x_i, z_i) and the (x_{k-1}, z_{k-1}), and d_2 is the distance between the (x_i, z_i) and the $(x_{k|k-1}, z_{k|k-1})$.

In selection stage, measurements are assigned to trajectories. If there is only one candidate in the gate of a trajectory, we consider the candidate as the measurement of the trajectory at the moment. If a trajectory has more than one candidate, we use the velocity consistent constraint. If the velocity formed by the trajectory and a candidate is most similar to the velocity of the trajectory among those candidates, the candidate is the measurement of the trajectory. The velocity similarity measure is

$$V_d = ||V_{k-1} - V_i||,\ i = 0, 1, ..., q\ , \tag{9}$$

where V_d is the module of the difference between two velocity vectors. V_{k-1} is the velocity vector of the trajectory at time $k - 1$. V_i is the velocity vector from the previous centroid of the person to the ith candidate. q is the number of the candidates belonging to the person. If a measurement has more than one corresponding trajectory, we think a conflict is presented. We will use the local area correlation coefficient (LACC) in a neighborhood centered at the measurement to decide which trajectory the measurement will be assigned to [8]. The LACC is

$$c_{ij} = \sum_{k=-1}^{n} \sum_{l=-m}^{m} \frac{[I_1(u_i^1 + k, v_i^1 + l) - \bar{I}_1(u_i^1, v_i^1)] \times [I_2(u_j^2 + k, v_j^2 + l) - \bar{I}_2(u_j^2, v_j^2)]}{(2n+1)(2m+1)\sqrt{\sigma_i^2(I_1) \times \sigma_j^2(I_2)}}\ , \tag{10}$$

where I_1 and I_2 are the intensities of the two images. (u, v) is the image coordinates of 3D point. (u_i^1, v_i^1) and (u_j^2, v_j^2) is the ith and jth feature points to be matched, m and n the half-width and half-length of the sliding window,

$$\bar{I}(u,v) = \sum_{i=-n}^{n} \sum_{j=-m}^{m} I(u+i, v+j)/[(2n+1)(2m+1)] \ , \tag{11}$$

the average intensity of the window.

$$\sigma = \sqrt{\frac{\sum_{i=-n}^{n} \sum_{j=-m}^{m} I^2(u+i, v+j)}{(2n+1)(2m+1)} - \bar{I}^2(u,v)} \ , \tag{12}$$

the standard variance of the window. c_{ij} ranges from -1 to 1, indicating the similarity from smallest to largest. If the LACC between a trajectory and the measurement is the maximum among those correspondences between trajectories and the measurement, the measurement is assigned to the trace in current frame.

In detection area, a new trajectory (new person) will be detected if a measurement can't match any other trajectories. One trajectory will disappear if it can't match any measurements in this area for three times continuously.

In tracking area, we think no new trajectory will present and no existed trajectory will disappear. We consider those trajectories as being occluded or stationary if they can't match any other measurements in this area. If an occluded or stationary trajectory can match a measurement in this area, it is considered as activation.

If an occlusion between persons is detected, the algorithm will record the intensity of a neighborhood centered at the trajectory in previous frame. Then the predicted state is taken as the estimating state of the trajectory. If a person is occluded for five times continuously, he is considered as being stationary. If there is a stationary trajectory that satisfies the condition described in the previous paragraph, then we think it acts again.

3 Experimental Results

This algorithm is implemented with C++ and run in a computer that mainly includes P4 2.8G CPU, 512M RAM, and 40G/5400rps Hard Disk. The real scene image sequences are taken as experimental data. The resolution of each image in those image sequences is 512×384. The mean sampling frame rate is 0.125 S. The interval time from frame to frame is variable and relative long in our application. The posture of a person facing surveillance camera is time-varying. The human body is nonrigid. The feature point set of each person in one frame change very violently in next frame. Those result in that the location of the centroid of each person fluctuates dramatically between some successive frames in those image sequences. The value of S in adaptive Kalman filter is much more than 5 at first several steps, and goes down to 1 gradually. The Markov chain governing the transition between the two models in IMM was chosen as $\begin{pmatrix} 0.95 \, 0.05 \\ 0.05 \, 0.95 \end{pmatrix}$.

In first experiment, we use the Cross_A image sequence that contains six persons and 41 frames. The experimental results are shown in Fig. 3 and Fig. 5. Fig. 3 shows the predicting error by our algorithm, and Fig. 5 shows the tracking results. As can be seen from Fig. 3, the predicting error, $||P_k - P_p||$, is very small where P_k is the estimating position of a trajectory and P_p is its predicting position. Through Fig. 5, we can see that no error tracking and missing tracking happen when some of those persons are occluded and are fused together. This shows our algorithm robust to the people interacting. The mean time of each step in this people tracking experiment is not more than 1 millisecond.

In second experiment, we use the Cross_B image sequence that contains three persons and 81 frames. Those persons rotate in circle in turn and the occlusion

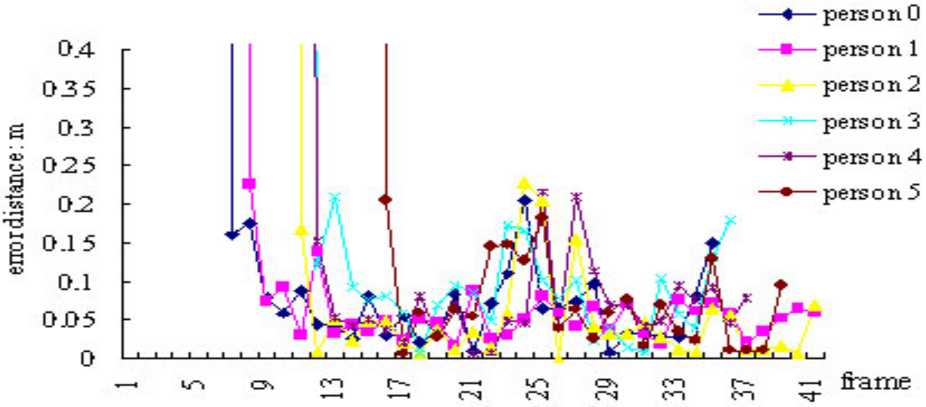


Fig. 3. The predicting error of Cross_A image sequence

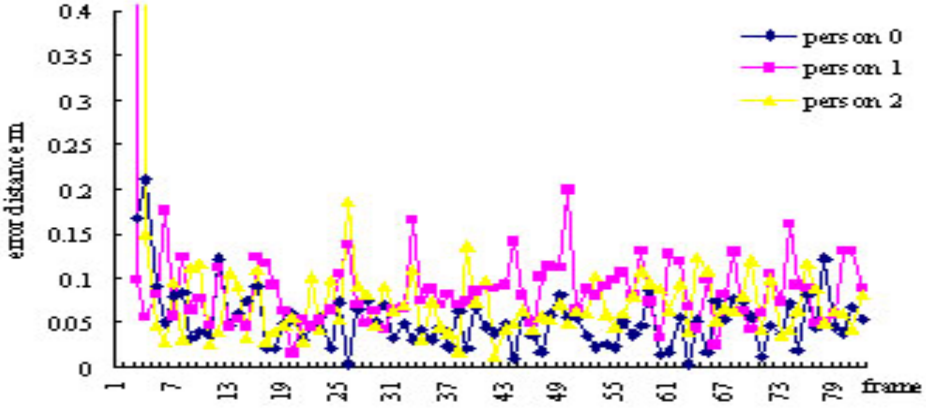
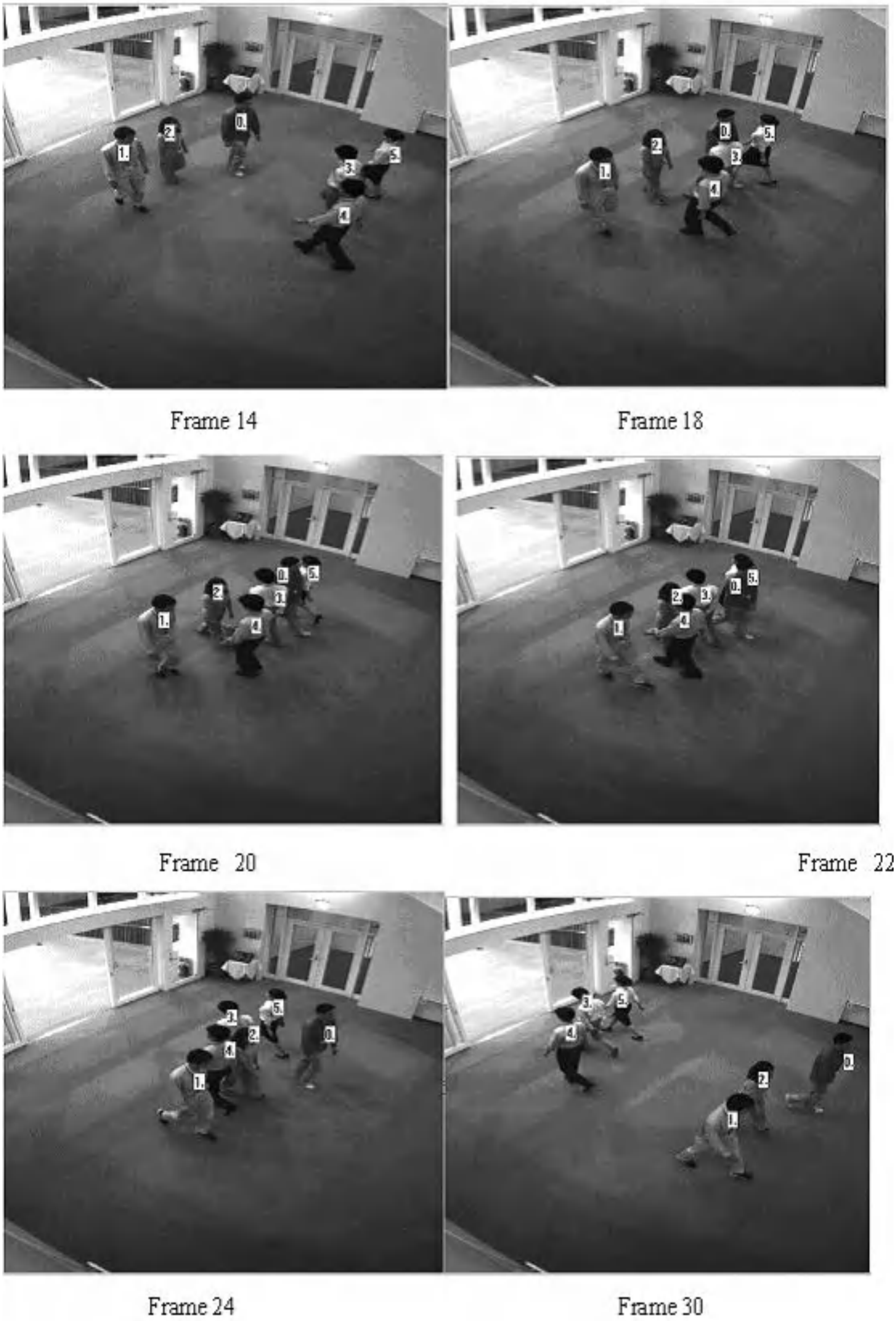


Fig. 4. The predicting error of Cross_B image sequence



Frame 14

Frame 18

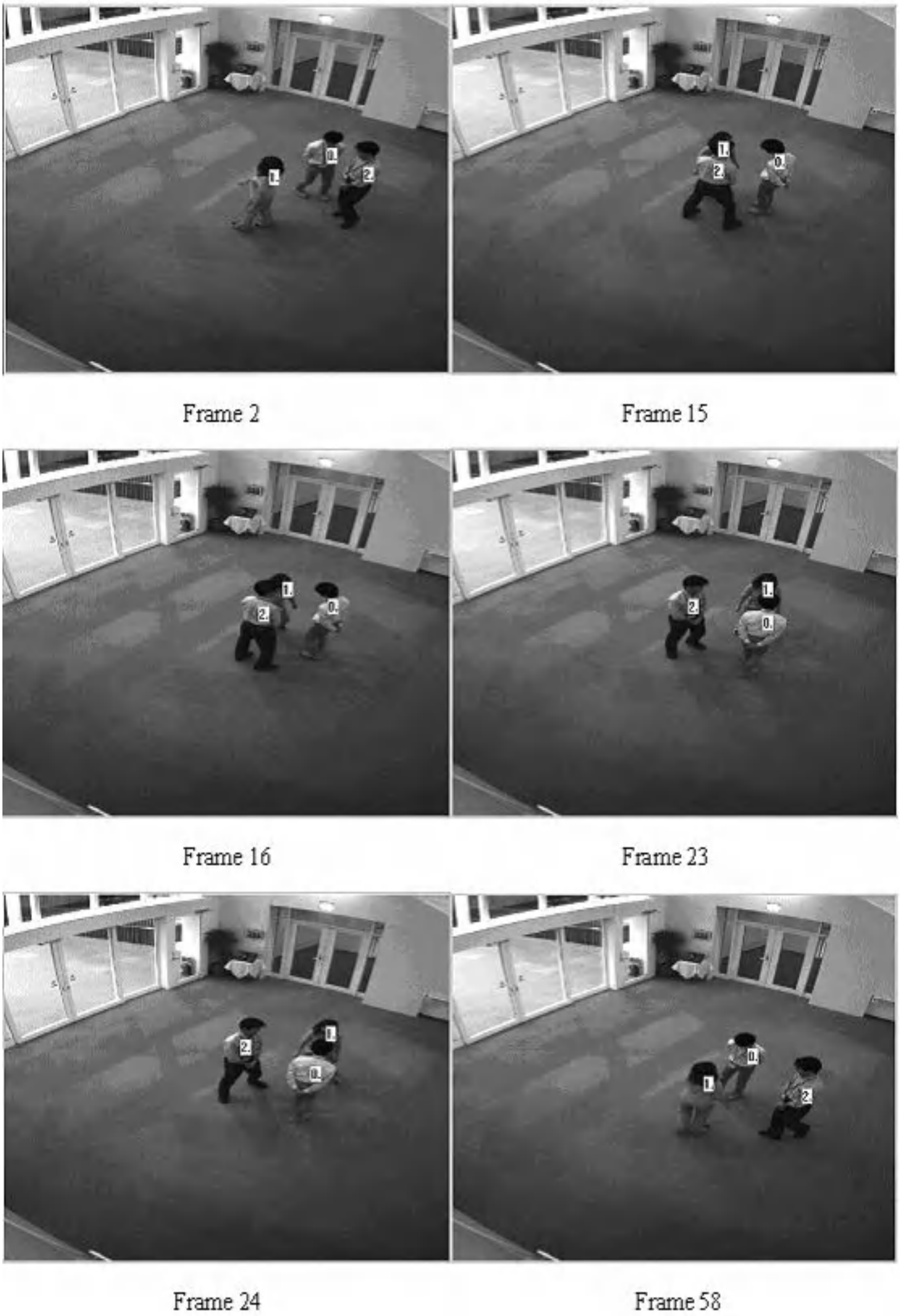
Frame 20

Frame 22

Frame 24

Frame 30

Fig. 5. The experimental result of the Cross_A image sequence by our algorithm



Frame 2

Frame 15

Frame 16

Frame 23

Frame 24

Frame 58

Fig. 6. The experimental result of the Cross_B image sequence by our algorithm

presents some times. The experimental results are shown in Fig. 4 and Fig. 6. Fig. 4 shows the predicting error by our algorithm, and Fig. 6 shows the tracking results. As can be seen from Fig. 4, the predicting error, $||P_k - P_p||$, is very small. Through Fig. 6, we can see that all of those persons are tracked correctly. This shows that our algorithm can deal with not only the occlusion between those persons but also the nonlinear movement of those persons. The mean time of each step in this people tracking experiment is not more than 1 millisecond too.

4 Conclusion

In this paper, we present a 3D feature-based people tracking algorithm which combines the IMM method with the cascade multiple feature data association method. Our algorithm uses the centroid of the 3D feature point set of a person to represent the person. This makes our algorithm real time. The IMM method used in our algorithm can predict the centroid of a person in next time accurately. The cascade multiple feature data association method ensures the unique correspondences between the trajectories and the measurements. The experiments have been done with the real scene image sequences. Although the centroid of each person fluctuates dramatically from frame to frame in those image sequences, the experimental results show that the predicting error resulted by this algorithm is very low even though the person moves nonlinearly. The robustness and efficiency of the proposed framework to the occlusion of people can be seen through those experimental results as well.

References

1. Schulz, D. and Burgard, W. and Fox, D. and Cremers, A.B.: Tracking Multiple Moving Objects with a Mobile Robot using Particle Filters and statistical Data Association. Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (2001)
2. Mittal, A., Davis, L.S.: M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. Int. J. Comput. Vision **51(3)** (2003) 189-203
3. SMITH Tang, C.Y., Hung, Y.P., Shin, Z.: A 3D Feature-based Tracker for Multiple Object Tracking. Proc. Natl. Sci. Counc. ROC(A) **23(1)** (1999) 151-168
4. SMITH, S.M., BRADY J.M.: SUSAN-A New Approach to Low Level Image Processing, Int. J. Comput. Vision **23(1)** (1997) 45-78
5. Zhou, H.R., Kumar, K.S.P.: A Current Statistical Model and Adaptive Algorithm for Estimating Maneuvering Targets. AIAA Journal, Guidance, Control and Dynamics, **7(15)** (1984) 134-148
6. Henk, A.P.B, Yaakov, B.S.: The interacting multiple algorithm for systems with markovian coefficients. IEEE Trans. On Automatic Control, **33(8)** (1988) 780-783
7. Mazor, E., Averbuch, A., Yaakov, B.S., Dayan, J.: Interacting Multiple Model Methods in Target Tracking: A Survey. IEEE Trans. On Aerospace and Electronic systems **34(1)** (1998) 103-123
8. Zhou, J, Shi, J.Y.: A Robust Algorithm for Feature Point Matching. Computers&Graphics. **26** (2002) 429-436

Self-tuning Fuzzy Control for Shunt Active Power Filter

Jian Wu, Dian-guo Xu, and Na He

Harbin institute of technology,
Harbin 150001, Heilongjiang Province, China
wujian@hit.edu.cn

Abstract. The shunt active power filter has been proved to be a useful means to improve power quality. This paper investigates the utilization of a new control strategy for three-phase three-wire voltage source shunt active power filter. The proposed scheme is composed of a self-tuning fuzzy logic controller and a series of generalized integrators. Self-tuning fuzzy controller, which operates through the error between the dc-link voltage and a reference value, regulates the voltage level of the dc-link and its output is regarded as the amplitude of grid reference current. The PI current controllers using stationary-frame generalized integrators are adopted, which can eliminate the steady state errors and compensate current harmonics selectively. In addition, an improved Fourier analytic approach is proposed to obtain the phase of the grid current reference by analysis of the grid voltage. The feasibility of the proposed scheme is validated by experimental results from a prototype.

1 Introduction

In recent years, the proliferation of nonlinear loads results in the deterioration of power quality in power systems [1-3]. Notably, harmonics are becoming a serious problem for both utilities and customers. The application of shunt active power filter (APF) to eliminate harmonic currents and to compensate reactive power for nonlinear loads has attracted much attention.

Since the reference current is a nonsinusoidal signal in APF, a Hysteresis or predictive controller is recognized as a viable solution [4-5]. However, both two methods have their advantages and shortcoming respectively [6], here not in repeat. PI controller, applied extensively in other fields, has been restricted in APF. The reason lies in that the reference signal of PI controller is the direct signal or a single sinusoidal signal, but the reference signal of APF is the sum of several sinusoidal signals. Even though a controller for APF with multiple reference frames can resolve this problem, a mass of coordinate transforms increases the computational effort significantly.

In order to solve above-mentioned problem efficiently, a generalized integrator algorithm for shunt APF is adopted in this paper [7]. Generalized integrator can track the sine signal of specific frequency with zero stable-state error and do not nearly affect the signal of other frequency. Because of this specialty, APF can compensate the harmonic of some certain frequency individually or the whole of harmonics simultaneously.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 890 – 899, 2005.

In APF, the three-phase voltage source type PWM converter using IGBT is often used. Voltage-type active power filter rely on capacity as storage element. Generally speaking, the control structure of APF is composed of detection unit, current control loop, DC voltage regulation and the pulse width modulator (PWM) for the generation of the gate signal of IGBT. For compensating harmonic efficiently, the DC voltage must be maintained at a reference value with small ripple in steady state.

Usually, PI controller is used in DC voltage control because of its simplicity. However, The PI controller requires precise linear mathematical models, which are difficult to obtain and fails to perform satisfactorily under parameter variations, load disturbance, etc. This leads in turn to some disadvantages, such as, great overshoot of DC voltage and acute current concussion when the start of the shunt APF.

In recent years, fuzzy logic controllers (FLCs) have generated a good deal of interest in certain applications. Generally speaking, the advantages or FLCs over conventional controllers are that they can handle non-linearity problem by expert knowledge or the professional experience, and they are more robust than conventional nonlinear controllers.

However, fuzzy controller is not satisfying in some condition because of the subjectivity of fuzzy decision [8]. To overcoming this shortcoming, a parameter a is introduced into fuzzy controller and control rule can be modified online based on a . In this paper, a generalized integrator algorithm based on self-tuning fuzzy controller is implemented to control shunt APF. The DC voltage is regulated to estimate the reference current by a fuzzy controller. The fuzzy controller has a better transient response compared to a conventional PI controller, and the steady state performance of the fuzzy controller is comparable to the PI controller. The PI current controllers using stationary-frame generalized integrators are adopted, which can eliminate the steady state errors and compensate current harmonics selectively. In addition, an improved Fourier analytic approach is proposed to obtain the phase of the grid reference current by analysis of the grid voltage [9]. The validity of the proposed scheme has been verified for the small prototype active power filter system.

2 Generalized Integrator Algorithm Based on Self-tuning Fuzzy Control

The experimental circuit structure of the shunt active power filter, shown in Fig.1, is composed of the grid source, nonlinear load and a voltage source PWM inverter (VIS). The harmonic-producing nonlinear load consists of a typical three–phase diode rectifier with a resistance in series with an inductance. The switching converter of active power filter is connected to the mains by the filter inductance L_f . Meanwhile the Capacitance C_f and the filter inductance L_f compose a low pass filter together in order to eliminate high frequency switching harmonic.

The control strategy of proposed shunt AFP is implemented in the $\alpha\beta$ stationary-frame. A self-tuning fuzzy controller, which operates through the error between the DC-link voltage and a reference voltage value, is used to regulate the voltage level of the DC-link and its output is regarded as the amplitude of grid reference current.

An improved Fourier analytic approach is proposed to obtain the phase of the grid reference current by analysis of the two-phase grid voltage.

Now, the phase and amplitude elements of grid current reference is available, that is, the grid current reference $i_s^*(t)$ has been obtained.

The control of shunt APF is realized by the closed-loop control of grid current, which needn't to detect the harmonics of nonlinear load. The error of grid current reference and the actual grid current are processed in PI current controller using generalized integrators to bring out the output reference voltage of VSI. And then, SVPWM program is implemented to produce the switching signals of IGBT.

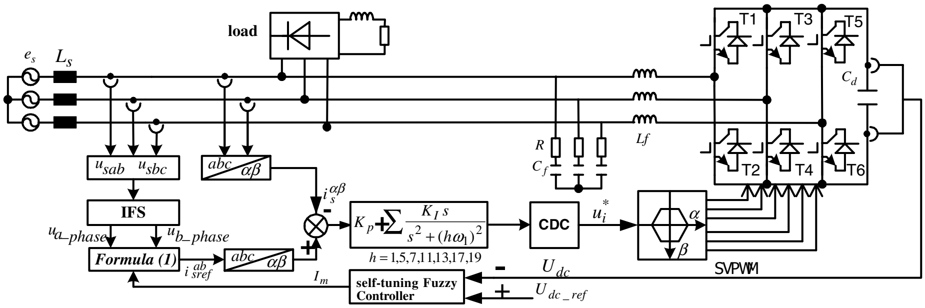


Fig. 1. Control strategy of shunt APF

2.1 Method to Calculate Grid Voltage Phase

Typically, the fundamental component of grid current is given by

$$i_s^*(t) = I_{sm}\cos(\omega_1 t + \varphi_1) \tag{1}$$

Where, $\cos(\omega_1 t + \varphi_1)$ and I_{sm} are the phase and the amplitude of grid fundamental current ($\omega_1 = 2\pi f$, f is the grid frequency).

For the compensation of both harmonics and reactive power simultaneously, both voltage and current should be of similar shape and in phase.

The proposed method employs an improved Fourier analytic approach using a TMS320C32 DSP. Not that the sampling frequency of the APF is fixed at 5kHz, which corresponds to a sampling time of $200\mu s$. The control and sampling cycle is then limited by this value. The time taken to compute the proposed algorithm is $30\mu s$, leaving the remaining $170\mu s$ for the overall APF controller.

Commonly, any non-sinusoidal signal can be expressed as a sum of sinusoidal of various frequencies. Based on this the utility can be expressed as

$$U_s(t) = \sum_{i=1}^{n} U_{nm}\cos(n\omega_1 t + \varphi_n) \tag{2}$$

Where, $U_s(t)$ is the instantaneous grid voltage, U_{nm} is the maximum value of nth-order voltage, φ_n is the phase angle of nth-order voltage, n is the order of harmonics, α_1 is as defined above. The factors for improved Fourier analytic approach are,

$$A_1(t) = \frac{2}{T}\int_{-T/4}^{0} U_s(t+\tau)\sin\omega\tau d\tau \qquad A_2(t) = \frac{2}{T}\int_{-T/4}^{0} U_s(t+\tau)\cos\omega\tau d\tau \tag{3}$$

$$S(t) = A_1(t) - A_2(t - T/4) - A_1(t - T/2) + A_2(t - 3T/4) = U_{1m}\sin(\omega t + \varphi)$$

$$C(t) = A_2(t) + A_1(t - T/4) - A_2(t - T/2) - A_1(t - 3T/4 = U_{1m}\cos(\alpha t + \varphi)) \tag{4}$$

Where, α is the angular frequency of internally generated sine signal by Digital Signal Processor (DSP). Then, the phase and amplitude of grid fundamental voltage can be obtained by (5) and (6)

$$U_{1m} = \sqrt{[S(t)]^2 + [C(t)]^2}\ ,\ \cos(\omega_1 t + \varphi_1) = \frac{C(t)}{U_{1m}} \tag{5}$$

In this paper, three-phase three-wire power system is studied, so the phase of grid voltage can be obtained by detection phase of line-voltage and then moving the phase for 30° prediction.

When the source voltages are imbalanced and/or distorted, the phase of grid currents is determined based on the orthogonal theorem for periodic sinusoidal functions, as shown in (7), in order to maintain the grid current to be balanced and sinusoidal.

$$phase_A(t) = \left\{\frac{1}{2}S_A(t) - \frac{\sqrt{3}}{6}C_A(t) - \frac{\sqrt{3}}{3}C_B(t)\right\}/U_{1m}$$

$$phase_B(t) = \left\{\frac{1}{2}S_B(t) + \frac{\sqrt{3}}{6}C_B(t) + \frac{\sqrt{3}}{3}C_A(t)\right\}/U_{1m} \tag{7}$$

Where, $phase_A(t)$ and $phase_B(t)$ is the instantaneous phase of grid voltage, $S_{A.B}(t)$ and $C_{A.B}(t)$ is the factors of improved Fourier analytic approach correspond to A and B phase.

Comparing this case with the method in [9], the proposed algorithm is superior owing to reducing computation time significantly, and saving the three fourth of operation time nearly. Usually, one or two sampling period prediction is needed in order to compensate computation delay and inaccurateness of detection.

2.2 Current Controller Based on Generalized Integrator

The conventional PI controller cannot track a sinusoidal reference signal without steady-state error. In addition, the grid voltage feed-forward control has to be adopted in order to get a good dynamic response. This subsequently results in the undesired aftereffect of the grid voltage background harmonics presenting in the grid current waveform.

In order to resolve these problems, a PI current controller using generalized integrators in stationary frame is used, depicted in Fig.2. Where, u_i^* is VSI reference voltage, i_i^* is grid reference current, h is the order of harmonics.

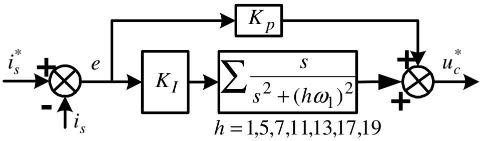


Fig. 2. PI current controller using generalized integrators

By this controller, the 5th, 7th, 11th, 13th, 17th and 19thharmonics could be compensated selectively. Furthermore, the control strategy doesn't need grid voltage feed-forward.

The proposed controller is composed of 14 harmonic controller and the following discrete formula is use to realize the digital generalized integrators.

$$E(z)\frac{K_I(1-z^{-1}\cos\omega T)}{1-2z^{-1}\cos aT+z^{-2}}=U(z) \tag{8}$$

Then, transform to difference equation,

$$U(k)=2U(k-1)\cos aT-U(k-2)+K_I\left[E(k)-E(k-1)\cos aT\right] \tag{9}$$

For reducing the computation time, the generalized integrators are implemented by iterative arithmetic.

Since the output of any generalized integrator controller is sinusoid, it is possible to compensate the control delay by predicting output. An example of control delay compensation (CDC) is depicted in Fig.3. The current output of any harmonic controller is stored in DSP memory and previous output of last harmonics period is sent to present output in advance, n is determined by practice.

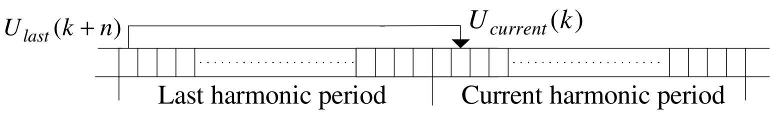


Fig. 3. Control delay compensation

Usually, the grid frequency is allowed to vary in $+0.5Hz \sim -0.5Hz$ range. The integral constant K_I determines the capacity of eliminating steady-state error. For coping with the variations of grid frequency potentially, the integral constant should be oversized properly. As same with the traditional PI controller, the size of proportional gain K_P determines the bandwidth and stability phase margin.

The open-loop transfer function of current loop can be expressed as

$$F(s) = \left(K_p + \sum_{i=1,5,7,11,13,17,19} \frac{K_I s}{s^2 + \omega_i^2} \right) G_d G_f(s) \tag{10}$$

Where G_d is the voltage gain of inverter, G_f is the function of output filter. The magnitude and phase characteristics of the open-loop function and close-loop function are shown in Fig.4 (a)-(b), where the 5th, 7th, 11th, 13th, 17th and 19th harmonics are compensated simultaneously. The PI parameter of current controllers can be determined based on the frequency characteristics of the open loop transfer function.

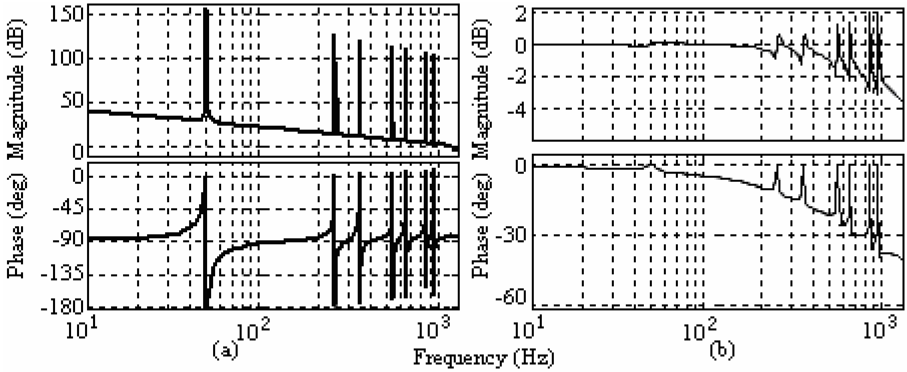


Fig. 4. (a) The magnitude and phase characteristics of the open-loop function. (b) The magnitude and phase characteristics of the close-loop function (K_P =0.1, K_I =4.5, G_d =400, bandwidth is 1200Hz, the phase margin is 98 deg).

2.3 Self-tuning Fuzzy Control

In order to compensate harmonic efficiently, the DC voltage of voltage-type shunt APF must be maintained at a constant level. The great fluctuation of DC voltage will result in excessive compensation or owing compensation. The proper control for inverter may guarantee the DC voltage keeping steady at certain value.

Usually, PI controller is adopted to regulate the DC voltage. By adding an active component to the reference value of grid current, DC voltage could be maintained at the reference value. The shortcoming of this scheme is that it is obligatory to detect fundamental component of load current, which will increase the cost of hardware and software. This paper adopts self-tuning fuzzy control for the DC voltage, getting the amplitude of the grid reference current directly. The system has been simplified hardware structure and software algorithm, the three-phase currents/voltages are detected using only two current/voltage sensors.

In the control system, the DC voltage error ΔU_{dc} and its deviation $\Delta\dot{U}_{dc}$ are selected as the input variable of fuzzy controller, and the amplitude of the grid reference current I_m is regarded as output variable.

In the control system, the DC voltage error ΔU_{dc} and its deviation $\Delta\dot{U}_{dc}$ are selected as the input variable of fuzzy controller, and the amplitude of the grid reference current I_m is regarded as output variable.

Define, $\Delta\dot{U}_{dc}$, I_m and ΔU_{dc} fuzzy sets as

$$\Delta\dot{U}_{dc} = I_m = \{NB, NM, NS, O, PS, PM, PB\} \tag{11}$$

$$\Delta U_{dc} = \{NB, NM, NS, NO, PO, PS, PM, PB\} \tag{12}$$

Their domains respectively is

$$\{-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6\}\ ,\ \{-7,-6,-5,-4,-3,-2,-1,0,1,2,3,4,5,6,7\} \tag{13}$$

Commonly, fuzzy control rule is invariable after designed, which results in that fuzzy controller cannot be suitable to all kinds of case and gets short of versatility. Aimed to overcome this disadvantage, fuzzy controller must have self-tuning function.

According to fuzzy input variable $\Delta\tilde{U}_{dc}$ 和$\Delta\tilde{\dot{U}}_{dc}$,this paper has designed the fuzzy controller system of self-tuning factor, controller principle block diagram as shown in Fig.5. By voltage sensor, System detects the accurate value of the DC voltage U_{dc} . Compared with reference value U_{dc_ref} , the accurate value of error signal can be gotten, and then derivative of error is available. The accurate value of the error and its derivative are transformed into fuzzy variable $\Delta\tilde{U}_{dc}$ and $\Delta\tilde{\dot{U}}_{dc}$, then fuzzy decision is carried out based on self-tuning fuzzy rule. Finally, the output of fuzzy controller I_m is regarded as the amplitude of grid reference current.

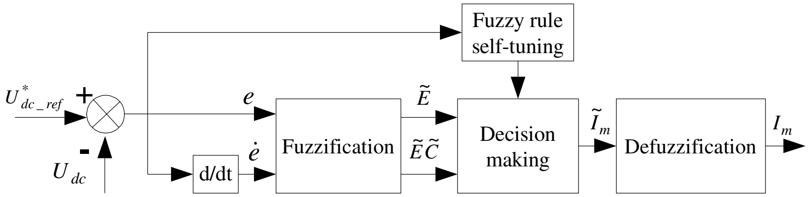


Fig. 5. Self-tuning Fuzzy Logic Controller

The proposed fuzzy control rule cam be expressed as

$$\tilde{I}_m = \alpha\cdot\Delta\tilde{U}_{dc} + (1-\alpha)\cdot\Delta\tilde{\dot{U}}_{dc}\ ,\ \ \alpha = k\cdot\left|\frac{\Delta\tilde{U}_{dc}}{R}\right|^p \tag{14}$$

Where, R , k and p are undetermined parameter. By setting these three parameters reasonably, adjustable factor α can modify fuzzy control rule flexible according to the error of DC voltage.

From the view of application, as the fast change of self-tuning factor, control rule also changes frequently. This often cause control process instability and increasing regulation time. Fig.6 shows the relation between fuzzy input variable $\Delta\tilde{U}_{dc}$ and self-tuning factor. It can be seen that, when error is less, self-tuning factor do not change acutely in greater scope to guarantee control rule unaltered; however, when error is bigger, self-tuning factor change quickly to maintain system stabile.

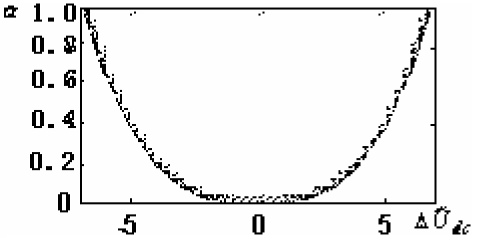


Fig. 6. Self-tuning gene

3 Experimental Results

The proposed control strategy of Fig.1 has been verified in a 10kW three-phase shunt active power filter prototype. The VSI dc-link is composed of a capacitor bank of 6800μF for dc voltage 450V. The Eupec 1200V/200A *IGBTS* are used, driven by the *M57962L* gate drivers. The sampling and switching frequency is set at 5 kHz. The digital control system incorporates a double *DSP* (*TMS320C32* and *TMS320F240*) and *CPLD* circuit. Phase detection and control algorithm are completed by *TMS320C32*, and sampling and SVPWM program are implemented by *TMS320F240.* The parameters of the experiment circuit are shown in Tab.1.

Table 1. The parameters of the experiment circuit

Filter inductance L_f	Grid line voltage	DC-link voltage	Capacity	Nonlinear load
4mH	380V(rms)	700V	6800$^{\mu F}$	L=40mH,R=50(series)

Fig.7 shows the experimental waveform of the PCC voltage, load current, grid current and their spectrums. It can be seen that the 5th, 7th, 11th, 13th, 17thand 19th harmonics have been eliminated effectively. The grid current is measured using HIOKI 3193 power analyzers. As it can be seen in Tab.2, each harmonic content and THD in grid current complies with IEEE-519 standard. Fig.8 shows the experimental waveform of grid current, output current of APF and their spectrums. Fig.9 show the waveforms of DC voltage and grid current when the shunt APF is starting, using self-tuning fuzzy controller. It can be seen that the DC voltage can arrive at its reference

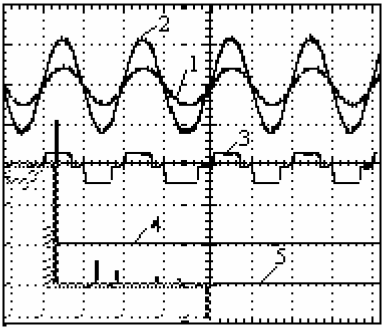


Fig. 7. Experimental results ((1). Grid current 20A/div; (2) the PCC voltage 250V/div; (3). Load current 20A/div; (4). Spectrum of grid current 5A/div; (5). Grid current spectrum, 5A/div 10ms/div)

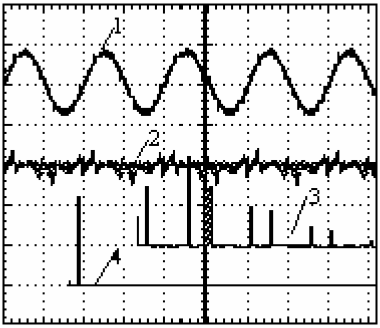


Fig. 8. Experimental results ((1). Grid current 20A/div; (2) Output current of APF 20A/div; (3). Spectrum of output current of APF 5A/div; (4). Grid current spectrum 5A/div. 10ms/div)

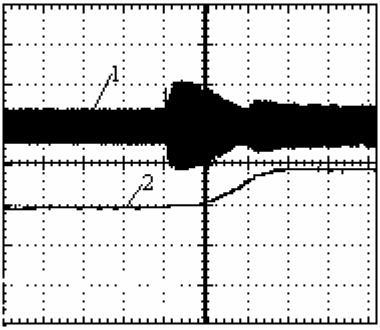


Fig. 9. Experimental results when APF is plunged. ((1). Grid current 20A/div; (2) DC voltage 250V/div. 10ms/div).

value rapidly and there is nearly no overshoot. The ripple of DC voltage is very small. Meanwhile, the grid current concussion becomes weaker.

Table 2. Harmonics content of grid currents

Harmonics order	5th	7th	11th	13th	17th	19th	THD
Not using APF (%)	20.4	12.0	9.2	7.0	5.1	4.1	29.2%
Using APF (%)	2.3	2.3	0.8	0.5	0.4	0.2	4.8%

4 Conclusion

It is shown in this paper that the presented control strategy based on $\alpha\beta$ frame allows the shunt active power filter to compensate load current harmonics successfully and eliminate reactive power disturbance between the grid and load. The proposed control strategy is composed of a self-tuning fuzzy controller and a series of generalized integrators. The inner current loop adopts the generalized integrators and outer voltage loop regulating the DC voltage uses self-tuning fuzzy controller. In addition, an updating Fourier analytic approach is adopted to get the phase of grid reference current. The experimental results have demonstrated the feasibility, effectiveness and advantages of the proposed approach.

References

1. Akagi, H.: New Trends in Active Filters for Power Conditioning. IEEE Trans. Ind Applicat. Nov. /Dec.32 (1996) 1312-1322
2. Bhim Singh, Kamal Al-Haddad, Ambrish Chandra.: A Review of Active Filters for Power Quality Improvement. IEEE Trans on I.E. 46(1999) 960-971
3. Agarwal, P., Chandra, A., Al-Haddad, K., and Srinivasan, K.: Active Power Filter to Compensate only Customer Generated Harmonics Simulation Study. in *Proc. 11th Nat. Power Syst. Conf.*, Bangalore, India (2000)
4. Buso, S., Malesani, L., and Mattavelli, P.: Comparison of Current Control Techniques for Active Power Filters. IEEE Trans. Ind. Electron, 45(1998) 722-729
5. Huang, S. J. and Wu, J. C..: A Control Algorithm for Three-phase Three-wired Active Power Filters under Nonideal Mains Voltages. *IEEE Trans. Power Electron*, July. 14 (1999) 753–760
6. Jeong, S. and Woo, M.: DSP-based Active Power Filter with Predictive Current Control. *IEEE Trans. Ind. Electron.* June. 44 (1997) 329–336
7. Zmood, D. N., Holmes, D. G.: Stationary Frame Current Regulation of PWM Inverters with Zero Steady-State Error. IEEE Trans. On Power Electr, May.18 (2003) 814-822
8. Feng Hsuan-Ming: A Self-tuning Fuzzy Control System Design. IFSA World Congress and 20th NAFIPS International Conference, Canada (2001)
9. El-Habrouk, M.; Darwish, M. K.: Design and Implementation of a Modified Fourier Analysis Harmonic Current Computation Technique for Power Active Filters using DSPs. IEE Proceedings, Electric Power Applications, Jan. 148 (2001) 21 - 28

Optimal Production Policy for a Volume-Flexibility Supply-Chain System

Ding-zhong Feng and Li-bin Zhang

College of Mechatronics Engineering, Zhejiang University of Technology,
Hangzhou, 310032, P.R.China
fdz@zjut.edu.cn

Abstract. A production-delivery system in a supply chain is always expected to reduce its overall production and management cost. In this research, a decision-making model is developed for optimal production rate selection in a single-stage supply chain system with volume flexibility, where raw materials and/or components are procured from suppliers and processed into finished products which are delivered to customers periodically at a fixed quantity with a fixed interval of time. In this model, production rate is perceived as a decision variable and unit production cost becomes a function of production rate. A pragmatic computational approach is presented to solve the proposed model for special unit production cost functions. Finally, a numerical study is conducted to illustrate the optimal solution and computational approach.

1 Introduction

A manufacturing enterprise usually synchronize its production capacity with customers' demand and coordinate the ordering of raw materials with production schedules, so that both raw materials' and finished goods' inventory is reduced. Thus, it usually orders its raw materials and delivers its finished goods in small lots for production system [1].

The customers' demand in the classical economic production lot size (EPL) model is assumed to be continuous and raw materials inventory cost is not taken into account [2]. Some researchers [3-6] developed optimal order and production quantity models for a single-stage production system. Cheng [7] proposed an extension to the EPL model in which demand exceeds supply and the production process is imperfect. Sarker and Parija [8] developed an ordering policy for raw materials to meet the demands of a production facility that supplies a fixed quantity of finished goods to outside buyers at a fixed interval of time. However, their models do not take volume flexibility into account. The production rate of a manufacturing system is assumed to be predetermined and inflexible. In fact, machine production rate can be easily changed [9]. Moreover, unit production cost depends on production rate. In other words, production rate in many cases should be treated as a decision variable. The treatment of production rate as a decision variable is especially appropriate for automation production issues, where the production volume is flexible. Volume flexibility of a manufacturing system is defined as its ability to be operated profitably at different overall output level [10]. Volume flexibility permits a manufacturing system to

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 900 – 909, 2005.

adjust production upwards or downwards within wide limits prior to the start of production of a lot. In a volume-flexible production system, as the production rate is increased, some costs (such as labor and energy costs) are spread over more units while per unit tool/die cost increases. The net result is that unit production cost decreases until an ideal design production rate of the facility is reached. Beyond the design production rate, unit production cost increases [11]. Therefore, it is very interesting to take volume flexibility into account in a production-delivery system. Khouja [12] developed an economic production lot size model under volume flexibility. But his work is limited to extending the basic EPL model. Feng and Yamashiro [13] developed an inventory model for a volume-flexibility production system in a supply chain under lumpy demand.

In this research, a decision-making model for optimal production rate selection in a single-stage supply chain system with volume flexibility is developed to determine a multi-order policy, an optimal batch size, and an optimal production rate. In this model, production rate is a decision variable, and unit production cost becomes a function of production rate.

2 Problem Description and Notations

A volume-flexibility production-delivery system is considered, where a manufacturer purchases raw material from outside suppliers, then processes them, and finally delivers periodically a fixed quantity of finished goods to a buyer at a fixed interval of time. The annual demand of this buyer is known and fixed. The manufacturer's production rate is assumed to be a decision variable, but not small than the demand rate so as to ensure no shortage of products due to insufficient production. Raw materials are nonperishable, and are supplied instantaneously to the manufacturing facility. A decision-making model for optimal production rate selection is to be developed to determine a multi-order policy, an optimal batch size, and a minimized total cost.

Figure 1 shows on-hand inventory of raw materials and finished goods. To model the interactions of finished goods and raw materials, the following notations are introduced and defined:

D_F : Demand for finished goods, units /year.

D_R : Demand for raw materials, units/year.

f : Conversion factor of raw materials into finished goods; $f = D_F / D_R$.

K_0 : Ordering cost of raw materials, \$/order.

K_S : Manufacturing setup cost per batch, \$/batch.

H_F : Holding cost of finished goods, \$/unit-year.

H_0 : Holding cost of raw materials, \$/unit-year.

m : Number of full shipments of finished goods per cycle time.

n : Number of orders for raw materials during the uptime, T_1.

P : Production rate, units/year.

P^0 : Optimal production rate.

P^* : Optimal integerized production rate.

Q_F : Quantity of finished goods manufactured per setup, units/batch.

Q_0 : Quantity of raw materials ordered each time, units/order; $Q_0 = Q_R / n$.

Q_R : Quantity of raw materials required for each batch; $Q_R = Q_F / f = nQ_0$.

L : Time between successive shipments of finished goods.

T : Cycle time; $T = Q_F / D_F = mL$.

T_1 : Manufacturing period (uptime); $T_1 = Q_F / P$.

x : Fixed quantity of finished goods per shipment at a fixed interval of time, units/shipment; $x = Q_F / m = L \cdot D_F$.

$f(P)$: A function that represents unit production cost as a function of production rate.

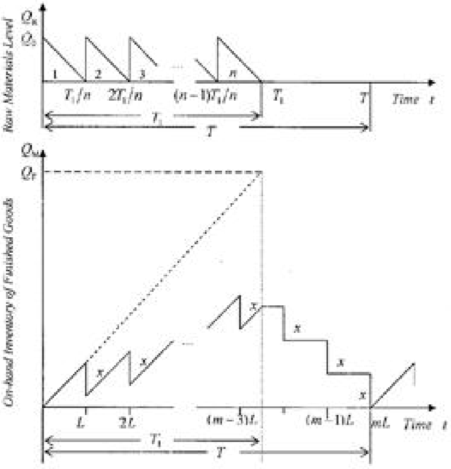


Fig. 1. On-hand inventory of raw materials and finished goods

3 Model Formulation and Optimal Solution

In this research, two types of inventory holding costs are considered: raw materials holding cost (TC_R^H) and finished goods holding cost (TC_F^H). The other associated costs that may incur to the producer (facility) are the ordering cost (TC_R^0) of the raw materials and the manufacturing setup cost (TC_F^S) for each batch. Besides, the incurred production cost is also integrated into the model.

The associated cost with raw materials (TC_R) is composed of TC_R^0 and TC_R^H , that is,

$$TC_R = \frac{mxD_F H_0}{2fnP} + \frac{nK_0 D_F}{mx} \tag{1}$$

The associated cost with finished goods (TC_F) includes TC_F^S and TC_F^H , that is [8],

$$TC_F = \frac{1}{2}\left[mx\left(1 - \frac{D_F}{P}\right) + x\right] \times H_F + \frac{D_F K_S}{mx} \tag{2}$$

The production cost (TC_P) is given by

$$TC_P = D_F f(P) \tag{3}$$

Consider the following unit production cost function:

$$f(P) = r + g / P + bP^{\beta}\ , \tag{4}$$

where $r,\ g,\ b$ and β are non-negative real numbers to be chosen to provide the best fit for the estimated unit production cost function. In the above function, r is a cost component independent on production rate and includes a unit raw-material cost. g / P represents a unit cost component that decreases with increased production rate. Such a cost component would include labor cost [14]. bP^{β} means a unit cost component that increases with increased production rate, and includes tool cost and rework cost that might result from increased tool wear out at higher production rates.

Let $\omega = 1/(1+\beta)$. The production rate that minimizes unit production cost given by $f(P)$ in Eq.(4) can be obtained:

$$P_m = [g/(b\cdot\beta)]^{\omega} \tag{5}$$

And the corresponding minimum unit production cost is

$$c_m = r + g^{\beta\omega}\cdot b^{\omega}[\beta^{\omega} + (1/\beta)^{\beta\omega}] \tag{6}$$

By substituting the expression for $f(P)$ from Eq.(4) into Eq.(3), direct production cost (TC_P) is obtained as

$$TC_P = D_F\ \left(r + g / P + bP^{\beta}\right) \tag{7}$$

Combining Eqs.(1), (2) with Eq.(7), therefore, we obtain a general expression for the total production and inventory cost function:

$$TC(m,n,P) = \frac{mxD_F H_0}{2fnP} ++ \frac{1}{2}\left[mx\left(1 - \frac{D_F}{P}\right) + x\right]H_F + \frac{D_F K_S}{mx} + D_F\ \left(r + g / P + bP^{\beta}\right) \tag{8}$$

In the above expression, the total cost $TC(m,n,P)$ is a function of decision variables: *m, n* and *P*. In order to obtain a solution with a minimum total cost $TC(m^0, n^0, P^0)$, the following conditions are necessary:

$\partial TC(m,n,P)/\partial m = 0$, $\partial TC(m,n,P)/\partial n = 0$, and $\partial TC(m,n,P)/\partial P = 0$.

Thus, we have:

$$m^0 = \frac{1}{x}\left(\frac{2D_F K_S}{(1 - D_F / P)H_F}\right)^{1/2} \tag{9}$$

$$n^0 = \left(\frac{D_F K_S H_0}{K_0 P f (1 - D_F / P) H_F} \right)^{1/2} \tag{10}$$

$$\frac{\partial TC(m,n,P)}{\partial P} = \frac{D_F}{P^2} \times \left(b \cdot \beta \cdot P^{\beta+1} - g + \sqrt{\frac{D_F K_S H_F}{2(1 - D_F/P)}} - \sqrt{\frac{K_0 H_0 P}{2f}} \right) = 0 \tag{11}$$

By using Eqs.(9) and (10) in Eq.(8), $TC(m,n,P)$ can be written as

$$TC(P) = \sqrt{2D_F^2 K_0 \, H_0/(f\,P)} + \sqrt{2D_F K_S H_F (1 - D_F / P)} + \frac{1}{2} x H_F + D_F \, \left(r + g / P + bP^{\beta} \right)$$

An efficient method to find its minimum is provided by the following proposition.

Proposition: Equation(11) has at most two solutions: (1) If Eq.(11) has no solution or has one solution at $P = P_{\min}$, then for $P \ge D_F$, the function $TC(P)$ has no stationary point, and $P = D_F$ is optimal; (2) If Eq.(11) has two solutions at $P = P_1$ and $P = P_2$ where $P_1 < P_2$, then for $P \ge D_F$, P_1 is only a local maximum point and P_2 is only a local minimum point. And the optimal solution is the one that minimizes the total cost over P_2 and D_F.

Proof:
Let:

$$G(P) = b \cdot \beta \cdot P^{\beta+1} + \sqrt{\frac{D_F K_S H_F}{2(1 - D_F / P)}} - \sqrt{\frac{K_0 H_0 P}{2f}} - g \tag{12}$$

Then, $G(P) = 0$ has identical solution with Eq.(11).

For $P > D_F$, we have $G''(P) > 0$, hence, $G(P)$ is convex. So we can conclude that Eq.(11) has at most two solutions: $P = P_1$ and $P = P_2$.

For $P > D_F$, we have

$$TC'(P) = \frac{D_F}{P^2} \; G(P) \tag{13}$$

$$TC''(P) = - \frac{2D_F}{P^3} \; G(P) + \frac{D_F}{P^2} \; G'(P) \tag{14}$$

(1) If Eq.(11) has no solution, then $G(P) > 0$. From Eq.(13), we have $TC'(P) > 0$. So we can conclude that $TC(P)$ increases with P. Thus, for $P \ge D_F$, function $TC(P)$ has no stationary point, and $P = D_F$ is optimal.

(2) If Eq.(11) has only one solution at $P = P_{\min}$, then $G(P_{\min}) = G'(P_{\min}) = 0$, From Eqs.(13) and (14), we have $TC'(P_{\min}) = TC''(P_{\min}) = 0$. For $P > D_F$ and $P \ne P_{\min}$, we have $G(P) > 0$ since $G''(P) > 0$, thus, $TC'(P) > 0$. So we can conclude that $TC(P)$ is increasing in P. Therefore, for $P \ge D_F$, function $TC(P)$ has no stationary point, and $P = D_F$ is optimal.

(3) If Eq.(11) has two solutions at $P = P_1$ and $P = P_2$ where $P_1 < P_2$, then $G(P_1) = 0$, $G(P_2) = 0$ and $G'(P_1) < 0$, $G'(P_2) > 0$. From Eq.(13) and (14), we have $TC'(P_1) = 0$, $TC'(P_2) = 0$ and $TC''(P_1) < 0$, $TC''(P_2) > 0$. Therefore, for $P \geq D_F$, P_1 can only be a local maximum point and P_2 can only be a local minimum point, and thus the optimal solution is the one which minimizes the total cost over P_2 and D_F.

4 Pragmatic Computation Approach

After getting P^0, m^0 and n^0 can be calculated from Eqs.(9) and (10), respectively. However, it is worth noting that the global optimal solution (m^0, n^0) with the minimum cost $TC(P^0, m^0, n^0)$ are usually real numbers. Thus, m^0 and n^0 have to be integerized for practical use. In such a case, an integer approximation scheme is applied.

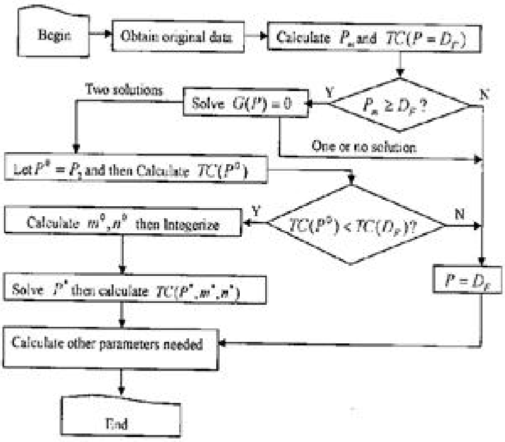

Begin
Obtain original data
Calculate P_m and $TC(P = D_F)$
Two solutions
Solve $G(P) = 0$
Y
$P_m \geq D_F$?
N
One or no solution
Let $P^0 = P_2$ and then Calculate $TC(P^0)$
Calculate m^0, n^0 then Integerize
Y
$TC(P^0) < TC(D_F)$?
N
$P = D_F$
Solve P^* then calculate $TC(P^*, m^*, n^*)$
Calculate other parameters needed
End

Fig. 2. Flow chart of an algorithm for calculating feasible optimal solution

Suppose that (m^*, n^*) is an optimal integer solution. Let $\lfloor m^0 \rfloor \leq m^* \leq \lceil m^0 \rceil$ and $\lfloor n^0 \rfloor \leq n^* \leq \lceil n^0 \rceil$, then it can be claimed that the optimal integer solution (m^*, n^*) belongs to the following set:

$$(\lfloor m^0 \rfloor, \lfloor n^0 \rfloor), (\lfloor m^0 \rfloor, \lceil n^0 \rceil), (\lceil m^0 \rceil, \lfloor n^0 \rfloor), (\lceil m^0 \rceil, \lceil n^0 \rceil)\} \tag{15}$$

Further, total cost function $TC(P^*, m^*, n^*)$ may be empirically observed to have its minimum value either at $m^* = \lceil m^0 \rceil$ and $n^* = \lceil n^0 \rceil$ or at $m^* = \lfloor m^0 \rfloor$ and $n^* = \lfloor n^0 \rfloor$.

Once m^0 and n^0 are modified into m^* and n^* respectively, the corresponding production rate P^* should be recalculated by using the integerized m^0 and n^0 in Eq.(11) so as to ensure that the modified total cost $TC(P^*, m^*, n^*)$ can be closest to the minimum cost $TC(P^0, m^0, n^0)$.

Combining the integer approximation scheme with the proposition in Section 3 for finding out optimal production rate P^0, an algorithm is developed to calculate feasible optimal solution (i.e., integerized optimal solution). Figure 2 describes the flow chart of such an algorithm through which the feasible optimal solution is obtained step by step.

5 Numerical Study

In order to illustrate the proposed pragmatic computation approach for obtaining a feasible optimal solution closest to a minimum total cost, we conducted a numerical study.

For unit cost function $f(P)$, we consider two specific cases with different degrees of flexibility. In case 1, β=1, that is,

Case 1: $f(P) = r + g/P + bP.$

By using Eqs.(5) and (6), the production rate that minimizes unit production cost is $P_m = \sqrt{g/b}$ and the minimum unit production cost is $c_m = r + 2\sqrt{gb}$. In case 2, β=2, that is,

Case 2: $f(P) = r + g/P + bP^2.$

Similarly, the production rate that minimizes unit production cost is $P_m = (g/2b)^{1/3}$ and the minimum unit production cost is $c_m = r + 1.89(g^2b)^{1/3}$. For identical P_m and c_m, the unit production cost function in case 1 describes a more flexible production system (or facility) than case 2 does. Consider, for instance, $f(P)$ =15+18000/P+0.002P for case 1 and $f(P)$ =15+24000/P+4.4444$\times 10^{-7} P^2$ for case 2. Case 1 represents a more flexible production system (facility) in the sense that smaller cost increases are associated with departures from P_m . For both cases of the unit production cost functions, the production rate that minimizes unit production cost is P_m =3000 units/year and the minimum unit production cost is c_m =\$27/unit.

For other original parameters, we consider two examples in which original data are taken from Ref[8], to illustrate the possible cases for the solutions. In example A, $D_F = 2400$ units/year K_0 =\$100/order, K_S =\$300/setup, $H_0 = \$1$ /unit/year, $H_F = \$2$ /unit/year, $x = 100$ units/shipment, and $f = 0.5$. In example B, $D_F = 2400$ units/year, $K_0 = \$100$ /order, $K_S = \$200$ /setup, $H_0 = \$5$ /unit/year, $H_F = \$10$ /unit/year, $x = 100$ units/shipment, and $f = 0.5$. By using the algorithm proposed in Section 4, computational results are shown in Tables 1 and 2.

Table 1. Computational results for example A

	P^*	(m^*, n^*)	TC_R^0	TC_R^H	TC_F^S	TC_F^H	TC_P	TC^*
$P = P_m$ =3000	/	/	400	480	400	460	64800	66540
Case 1	2859	(22, 4)	436	462	327	453	64833	66512
Case 2	2936	(20, 4)	480	409	360	465	64813	66527

Table 2. Computational results for example B

	P^*	(m^*,n^*)	TC_R^0	TC_R^H	TC_F^S	TC_F^H	TC_P	TC^*
$P=P_m$ =3000	/	/	1029	933	686	1200	64800	68648
Case 1	D_F =2400	/	1095	1095	0	500	65520	68211
Case 2	2899	(7, 3)	1029	966	686	1102	64833	68616

As shown in the two tables, at first, the examples are solved under a special selection of $P=P_m$. It is worth to note that such total costs are always larger than those for case 1 or case 2 where production rate is a decision variable, since they, including indirect production costs ($TC_R + TC_F$), are not wholly optimized although the direct production costs are minimal.

For example A in Table.1, when production rate is treated as a decision variable, P^*=2859, m^*=22 and n^*=4 for case 1, and P^*=2936, m^*=20 and n^*=4 for case 2. These results can be explained as follows: when production rate is treated as a decision variable, a tradeoff arises between indirect and direct production costs. As production rate decreases, unit production cost increases. On the other hand, finished goods inventory cost (TC_F) decreases more sharply than raw materials inventory cost (TC_R) increases.

For example B in Table.2, when production rate is treated as a decision variable, P^*=2400 for case 1, and P^*=2899, m^*=7 and n^*=3 for case 2. It is worthy of pointing out that for case 1, there is an optimal solution with P^0=2680, m^0=9.58 and n^0=4.14. However, this solution is only a local optimum (the corresponding TC^0=\$68558) since $P^*=D_F$=2400 results in a smaller cost (TC^*=68211)

To illustrate further numerical results where production rate is treated as a decision variable, a set of ten problems with the input data given in Table 3 is tested. These data are taken from Ref.[8] where production rate is predetermined. As two different functions of unit production cost, case 1 and case 2 are still applied to theses problems. Using the algorithm proposed above, relative computational results are given in Table 4 for case 1 and in Table 5 for case 2. In these two tables, δ=[(TC^* - TC^0)÷TC^0]×1000, TC^1 is obtained by the predetermined P and TC^* given in Ref.[8].

From both of tables it can be seen that the total cost at the optimal integerization solution is always higher than the cost at the optimal solution, but the respective deviation of the cost from the minimal cost at optimal solution is extremely small. It shows that optimal integerization solution is feasible for practical industry activities.

Tables 4 and 5 show that the total cost at the optimal integerization solution, where production rate is treated as a decision variable, is smaller than the cost where production rate is predetermined. This result is due to such a fact that indirect production costs are only optimized at the latter situation, while total production costs including indirect production costs are wholly optimized at the former situation. For production system operation, therefore, the proposed model and its solution are more advantageous in total cost control.

Table 3. Input data for examining different solutions

problem	D_F	K_0	K_S	H_0	H_F	x	f
1	1200	50	150	2	3	50	0.8
2	1500	100	200	4	8	75	0.6
3*	2400	100	300	1	2	100	0.5
4*	2400	100	200	5	10	100	0.5
5	3600	300	400	2	2	50	0.8
6	2000	100	200	10	15	100	1.5
7	1000	400	300	2	10	50	0.2
8	1200	50	200	2	3	50	1.0
9	2500	100	50	5	15	100	0.5
10	4000	50	200	1	10	100	1.0

Table 4. Computational results for Case 1

problem	P^0	m^0	n^0	TC^0	P^*	(m^*, n^*)	TC^*	δ(‰)	TC^1
1	2980	8.96	1.30	33626	2992	(8,1)	33635	0.27	33966
2	2952	5.21	1.31	43347	2979	(5,1)	43375	0.65	45386
3*	2866	21.04	3.93	66511	2859	(22,4)	66512	0.02	67170
4*	2680	9.58	4.14	68558	2400	/	68211	/	76300
5	/	/	/	/	3600	/	100294	/	118273
6	2805	4.31	1.49	58039	2746	(5,2)	58070	0.53	58183
7	3082	5.96	0.60	30879	3000	(6,1)	31083	6.61	34000
8	2967	10.37	1.35	33714	2979	(10,1)	33725	0.33	34035
9	2873	3.58	1.49	71062	2500	/	70986	/	71125
10	/	/	/	/	4000	/	111133	/	117270

Table 5. Computational results for Case 2

problem	P^0	m^0	n^0	TC^0	P^*	(m^*, n^*)	TC^*	δ(‰)	TC^1
1	2990	8.95	1.29	33626	2996	(8, 1)	33635	0.27	34278
2	2976	5.18	1.30	43348	2990	(5,1)	43375	0.62	47146
3*	2939	19.82	3.66	66524	2936	(20,4)	66527	0.05	67714
4*	2889	7.53	3.13	68612	2899	(7,3)	68616	0.06	88300
5	/	/	/	/	3600	/	101110	/	152377
6	2909	4.13	1.40	58063	2873	(5,2)	58114	0.88	58539
7	3040	5.98	0.61	30881	3000	(6,1)	31083	6.54	36000
8	2984	10.35	1.34	33714	2990	(10,1)	33725	0.33	34347
9	2946	3.32	1.37	71073	2968	(3,1)	71122	0.55	71125
10	/	/	/	/	4000	/	113577	/	126514

6 Conclusions

In this research, we extend the classical EPL model to cases where production rate is a decision variable. Unit production cost becomes a function of production rate. The proposed model also takes raw materials inventory cost into account. Further, this model is based on a volume flexible production system operating under a fixed-quantity, periodic delivery policy. Such a demand policy complies better with practical industry activities than continuous demand policy that is used in the classical EPL model. The numerical study we conduct shows that the proposed model is more advantageous in total cost control than other models where, although indirect production costs are optimized, production rate is predetermined and inflexible. Using the developed model and its solution, one can determine an optimal multiple ordering policy for the procurement of raw materials, an optimal batch size for finished goods, and an appropriate production rate to meet the flexible demand of customers.

References

1. Feng, D.Z., Yamashiro, M.: A Pragmatic Approach for Optimal Selection of Plant-specific Process Plans in a Virtual Enterprise. Production Planning & Control. 14 (2003) 562-570
2. Hax, A.C., Candea., D.: Production and Inventory Management. Prentice-Hall, Englewood Cliffs. NJ (1984)
3. Pan, A.C., Liao, C.J.: An Inventory Model under Just-in-time Purchasing Agreements. Production and Inventory Management. 30/1 (1989) 49-52
4. Ramasesh, R.V.: Recasting the Traditional Inventory Model to Implement Just-in-time Purchasing. Production and Inventory Management. 31/1 (1990) 71-75
5. Sarker, B.R., Parija, G.R.: An Optimal Batch Size for a Production System Operating under a Fixed-Quantity, Periodic Delivery Policy. J. Operational Res. Society. 45/8 (1994) 891-900
6. Cachon, G.P., Zipkin, P.H.: Competitive and Cooperative Inventory Policies in a Two-stage Supply Chain. Management Science. 45/7 (1999) 936-953
7. Cheng, T.C.E.: An Economic Order Quantity Model with Demand-Dependent Unit Production Cost and Imperfect Production Processes. IIE Transactions. 23/1 (1991) 23-28
8. Sarker, B.R., Parija, G.R.: Optimal Batch Size and Raw Material Ordering Policy for a Production System with a Fixed-interval, Lumpy Demand Delivery System. European Journal of Operational Research. 89 (1996) 593-608
9. Schweitzer, P.J., Seidmann, A.: Optimizing Processing Rates for Flexible Manufacturing Systems. Management Science. 37 (1991) 454-466
10. Sethi, A.K., Sethi, P.S.: Flexibility in Manufacturing: a Survey. International Journal of Flexible Manufacturing Systems. 2 (1990) 289-328
11. Ramasesh, R.V., Jayakumar, M.D.: Measurement of Manufacturing Flexibility. Journal of Operations Management. 10/4 (1991) 446-467
12. Khouja, M.: The Economic Production Lot Size Model under Volume Flexibility. Computer & Operations Research. 22 (1995) 515–523
13. Feng, D.Z., Yamashiro, M.: Optimal Production Policy for a Manufacturing System with Volume Flexibility in a Supply Chain under Lumpy Demand. International Journal of Advanced Manufacturing Technology. 25 (2005) 777-784
14. Petropoulos, P.G.: Optimal Selection of Machining Rate Variables by Geometric Programming. International Journal of Production Research. 11(1973) 305-314

Flame Image of Pint-Sized Power Plant's Boiler Denoising Using Wavelet-Domain HMT Models*

Chunguang Ji[1], Ru Zhang[1], Shitao Wen[1], and Shiyong Li[2]

[1] School of Computer Science and Technology, Harbin Institute of Technology,
Harbin 150001, P.R. China
jcg@hope.hit.edu.cn
[2] Department of Control Science and Engineering, Harbin Institute of Technology,
Harbin 15000, P.R. China

Abstract. Wavelet-domain hidden Markov Tree (HMT) was recently pro-posed and often applied to image processing. In this paper, HMT is app-lied to denoise the flame image of boiler and has gotten a good result. Having compared with other denoise methods such as wavelet, Wiener filter and median filter. HMT can get better denoise result and the content of flame image edges can be kept better. With the development of HMT research, it will be extended to the fields of signal processing, detection of edge and classification.

1 Introduction

In the north China, there are some medium-sized power plants and pint-size power plants. The boilers in these plants are industrial chain boilers. The structure, principle of working and flaming characteristic of the industrial chain boilers is quite distinct from the boiler of power plants. Thereby, the power plants control system are improper for the industrial chain boilers, especially for the problems of pre-combustion, burning ashes are enough big to keep red flame, flame stopped and coal blocked. At present, we have to use the manual work to control the loss for air leak and vent smoke. Accordingly it is quite import-ant to design the boiler computer-control system for supervising the state of flame in the boilers [1].

Flame images of the boiler different from the general image are collected by the industrial-vidicon that have some dynamic noises and the details of edge of image are important for the boiler computer-control system. Through the practical experiment, the noise is confirmed to be Gaussian white noise. We use some other methods to wipe off the noise, such as: Wiener filter, median filter and wavelet denoise, but we don't obtain good results. In this paper, we used wavelet-domain hidden Markov tree (W-HMT) to denoise and proved that this algorithm has some predominance compared with other methods.

The characteristics of wavelet transform are presented in section 2. Section3 introduced W-HMT algorithm. In section 4, the denoise experiments are made and give some analyses for the result. The conclusion is presented in section 5.

* Aid by Harbin Institute of Technology Cross Subject Fund (HIT.MD2001.35).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 910 – 919, 2005.

2 Characteristics of Wavelet Transformation

The wavelet transform algorithm for the1-D signal divides the signal into two parts including high (H) sequence and low (L) sequence in the frequency-domain but the total lengths of signal don't change. We can find that the wavelet forms a bin-tree and its subbands have the corresponding relationship between different layers(see Fig 1(a)). It is easy to extend this principle from1-D signal to the2-D signal. The2-D image wavelet coefficients include low-pass (LL) coefficients and three band-pass coefficients in the horizontal (LH), vertical (HL) and diagonal directions (HH). The2-D wavelet coefficients form a quad-tree, (see Fig 1(b)).

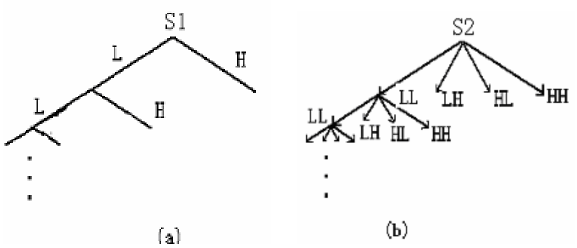


Fig. 1. (a) 1-D Wavelet transformation formed bin-tree; (b) 2-D Wavelet transformation formed quad-tree

Actually, some useful characteristics of the wavelet transform have been known [2]:

1. Each wavelet coefficient represents the local time and local frequency character of the signal.

2. The wavelet transform represents the signal at a nested set of scales.

3. The wavelet coefficients become larger when the signals change dramatic-ally at corresponding locations. Consequently it acts as local edge detector for image.

4. Decrease of the wavelet transform coefficients accompanied with the scales increased. The wavelet is large only if its parent coefficients are large, vice versa.

5. Each subband of wavelet tree tends to be disrelated to the others.

Accordingly, the statistical model of image in wavelet-domain is more efficient than the model constructed only in time-domain or frequency-domain. Since the wavelet transform coefficients of image possess the identical Gaussian property and independent nonGaussian property at same time, the Gaussian mixture model in this paper is used to captures the statistical characters of wavelet coefficients.

3 Wavelet-Domain HMT Models

3.1 HMT Model

HMT model as one of HMM has some characteristics of HMM [3], except that its configuration likes a tree (see Fig2). The parameters of HMT have three parts.

W=(W1, W2, W3.... Wn) is denoted the sequence of observation. The root node is W1 and the height of tree is J. s= (s1, s2, s3.... sn) represents the sequence of hidden state corresponding W and also has the configuration of tree. The root node is s1. To the each hidden state s∈k, k={1,2,3...K}, we connect the hidden states at different layer with Markov-1 chain. The hidden s-ate of the current node can only be influenced by its parent node and children node, as the equation (1), where $\forall u \in \{1,......,n\}$.

$$p(s_u \mid \{s_{u'} \mid u' \neq u\}) = p(s_u \mid s_{p(u)}, s^u_{c1}, ..., s^u_{cn_u}) \tag{1}$$

The observation variable (black node) is controlled by the hidden state variable (w hite node):

$$p(W_u \mid s_1, ..., s_n) = p(W_u \mid s_u) \tag{2}$$

The parameters need to describe the HMT model:

1. The root distribution probability matrix: $\pi = (\pi_k)_{k \in \{1.....k\}}$;
2. The state transition probability matrix: $A = (a^{rm}_{p(u),u})_{u \in (2,......,n), r \in (1,......,k), m \in (1,.....,k)}$;
3. The mixture parameters: $(\theta_1, ...\theta_K)$ $p(W_u = w \mid s_u = k) = p_{\theta_k}(w)$;

p_θ is the probability density function of the observation vector. We model the HMT by the parameters $\lambda = (\pi, A, \theta_1, ..., \theta_K)$ and use the expectation maximization (EM) algorithm maximizing the $p_\lambda(W)$. Get the model parameters thro-ugh training data by the up-down algorithm that is derived from backward-for-ward algorithm [4].

3.2 Wavelet-Domain HMT Model

In section 2, we have known some advantages of wavelet transformation. The W-HMT (see Fig 2, the black node denote wavelet coefficients; the white node denote hidden state) is constructed to model the joint statistics of wavelet coefficients by capturing the key joint dependencies of wavelet coefficients. In Fig 1(b), we can find that each node from a parent wavelet coefficient has an arrow to its four children at next finer scale. In order to capture the parent-children dependencies of wavelet coefficients, HMT models the transition of hid-den state by Markov chain. In paper [2], the marginal density of each wavelet coefficient controlled by hidden state is modeled as a Gaussain mixture model. A two – density Gaussian mixture model has been used to approximate the relationship between the observation and hidden state. Each wavelet coefficient ω_i has a discrete hidden state S_i of node i, and the probability distribution f-unction of W_i:

$$f(W_i) = \sum_{m \in \{0,1\}} p_{S_i}(m) f(W_i \mid S_i = m) \tag{3}$$

Where m={0,1}, $f(W_i \mid S_i = m) \sim N(\mu_{i,m}, \sigma^2_{i,m})$ and $p_{S_i}(0) + p_{S_i}(1) = 1$.

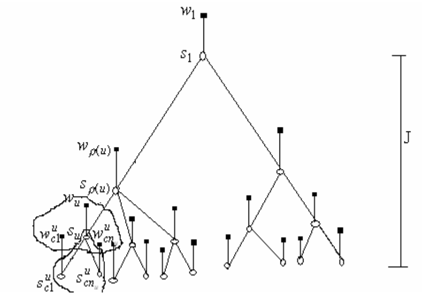


Fig. 2. The quad-tree configuration of HMT, black nodes denote observation; white no-des den ote hidden state

According to characteristics of wavelet transformation, we can see that its parent influences the wavelet coefficients large or small. The wavelet transform coefficients representing the edge of image with large coefficients tend to propagate across scale in wavelet quad-tree. Dependencies between the wavelet coefficients can be captured by the joint probability mass function of the hidden states of the HMT model. In all scales, we assume $\sigma_1^2 > \sigma_0^2$ in order to keep t-he edge content of the real-world image. In general, the wavelet-domain HMT model of an image has the parameters more than the wavelet transform coefficients of the image. The large number of parameters can't be acceptable. In t-his paper, a simple statistical model called Independent Mixture Model (IMM) was discussed which assumes each subband of wavelet coefficients in same sc-ale are independent and the wavelet coefficients in each scale follow a GMM. Therefore the total number of HMT parameters is greatly reduced. All parameters of J-scale can be represented as Θ. The wavelet coefficients of image are represented by w. We apply the EM algorithm to compute Θ and make $f(w \mid \Theta)$ maximizing.

3.3 Parameters of W-HMT

3.3.1 Initialized the Parameters of W-HMT

EM algorithm is a usual and effective method for the deficient data used to estimate the parameters of model. Each iterative step computational complexity of algorithm is simple, but the rate of convergence is exponential. If the parameters of W-HMT have been proper initialized, less time is take for the algorithm convergence. Through the analyzing, the proper initialized parameters can be found.

First, the characteristic of the image wavelet coefficients is that the scale increased accompanies with the image wavelet coefficients decreased according to the logarithm rule, which is derived from the assumption that the image consists of some smooth areas and it is discontinuous between the various areas. Because the Gaussian mixture model (GMM) is used to describe the relationship between the wavelet coefficients and the hidden state, the mixture variances reflect the changing rule of wavelet coefficients. Thus, the variances also decreased according to the logarithm rule inter scales. If the GMM has two states, t-he 'big' and 'small' variance in the J scale is $\sigma_{J,S}^2$, $\sigma_{J,L}^2$:

$$\sigma_{J,S}^{2} = e^{Cs-J}, \sigma_{J,L}^{2} = e^{C_L - J} \tag{4}$$

Where C_S and C_L are constants, $\sigma_{J,S}^2 < \sigma_{J,L}^2$ in each scale and $C_S < C_L$. According to formula (4), initialized variances have the same characteristics with wavelet coefficients that decreased according to the logarithm rule accompany with the scale increasing.

Secondly, the persistence of wavelet coefficients is that large/small values of wavelet coefficients tend to be propagating through the scales and it becomes stronger at finer scales. Wavelet function acts as a local edge detector. If there are edges in the wavelet function support region, the wavelet transform coefficients in that region is bigger and the edges can be exist at the finer scale wavelet function support region. In some scale, if there is no edge in the wavelet function support region, the edge can be found at finer scale, so the small wavelet coefficients must have small children. This is the persistence of wavelet coefficients. In transition probability matrix $\xi_{i,\rho(i)}^{S,L}$ $\xi_{i,\rho(i)}^{L,S}$ reflect the property of state across scales.

Last, if the scales become finer, the probability of small wavelets coefficients having small children increase. Until the edge in a certain scale is absolutely detached, there is no edge in next scales, when $J \to \infty$, $\xi_{i,\rho(i)}^{S,S} \to 1$. However the probability of big wavelet coefficients having big children coefficients is r-educed when the scale becomes finer, till there is no edge exit in next scale.

Accordingly, the state transition matrix initialized as:

$$\xi_{i,\rho(i)}^{S,S} = 1 - e^{-J}, \xi_{i,\rho(i)}^{L,L} = 1/2 + e^{-J} \tag{5}$$

Above all, through analyzing the characteristics of wavelet transform and state transition, the equation (4) and (5) were applied to initialize the parameters of W-HMT. The parameters of W-HMT λ have been initialized λ^0.

3.3.2 Predigest the Parameters of W-HMT

In W-HMT, each wavelet coefficient W_i has six parameters: means $\mu_{i,m}$, variance $\sigma_{i,m}^2$ of GMM, hidden state transition probability $\varepsilon_{i,\rho(i)}^{m,n}$, $m, n = 0{,}1$ and hidden state of root P_{S1}^S. All these parameters of W-HMT denoted by vector θ .A great deal training data are used to estimate the parameters of W-HMT. To predigesting the parameters of W-HMT, a method called 'tying' is used. Since wavelet coefficients in different position of image at same scale have the same statistic characteristics; wavelet coefficients in different sub-band at same layer have the uniform parameters. Through the 'tying', we can decrease a great deal parameters of HMT and make the estimation steadily. The parameters of three quad-trees HH, HL, LH of 2-dimension W-HMT are denoted by θ_{HH} , θ_{HL} , θ_{LH} .If the parameters are confirmed, W-HMT model is realized.

3.3.3 Estimate the Parameters of W-HMT

There are some similarities in estimating the parameters of the hidden Markov tree and hidden Markov chains. The up-down algorithm of HMT derived from backward –

forward algorithm of hidden Markov chain is used to assess the model parameters by training data. To make the description clearly, some concepts are mentioned: T_v is the subtree with root at node v and T_1 denotes the entire tree. T_t is a subtree of T_v, $T_{v\backslash t}$ is the set of nodes in T_v which are not in T_t. W is the observed data of HMT. S denotes the hidden states of HMT. The variables have been defined as follows:

$$\beta_v(m) = P(T_v \mid S_v = m) \tag{6}$$

$$\beta_{v,\rho(v)}(m) = P(T_v \mid S_{\rho(v)} = m) \tag{7}$$

$$\beta_{\rho(v)\backslash v}(m) = P(T_{\rho(v)\backslash v} \mid S_{\rho(v)} = m) \tag{8}$$

$$\alpha_v(m) = P(S_v = m, T_{1\backslash v}) \tag{9}$$

The observation data likelihood given by the formula (10), where $\forall v \in \{1,....,n\}$:

$$P(w) = P(w_1,....w_n) = \sum_{k=1}^{K} \beta_v(m)\alpha_v(m) \tag{10}$$

Up-downward algorithm is same as E-sep in the EM algorithm. M-step in t-he EM algorithm controls the iterative loop of convergence.

E-step algorithm: initialized the parameters of HMT λ^0.

Up-step: k is scale. In the finest scale make $k = 1$, the hidden state $S_v^{(1)} = m$ and compute $\beta_v^{(k)}(m) = P_{\theta_k}(W_v) = f(W_v^{(k)}, \mu_{v,m}^{(k)}, \delta_{v,m}^2)$ $m = 0,1$.

(1) In scale k, $S_v^{(k)} = m$, compute $\beta_{v,p(v)}^{(k+1)}(m) = \sum_{n=0}^{1} a_{nm}\beta_v^k(n)$:

(2) k =k+1;

$$\beta_v^{(k,k+1)}(m) = f(W_{p(v)}^{(k+1)}; \mu_{p(v),m}^{(k+1)}, \delta_{k+1,p(v),m}^2) \prod_{v \in (p(v))} \beta_{p(v),v}^{(k+1,k)}(m) \tag{11}$$

$$\beta_{p(v)\backslash v}^{(k+1\backslash k)}(m) = \beta_{p(v)}^{(k+1)}(m) / \beta_{v,p(v)}^{(k+1,k)}(m) \tag{12}$$

(3) If k=K, the computation stop. Otherwise, turn to step (1).

Down-step:

(4)In the coarsest scale make k=K, $S_v^{(K)} = m$ and compute $\alpha_1^{(k)}(m) = \pi_m = P_{S_k}(m)$, $m = 0,1$;

(5) Make the scale k=k-1, $S_v^{(K)} = m$ and compute

$$\alpha_v^{(k)}(m) = \sum_{n=0}^{1} \alpha_{p(v)}^{(k+1)}(n)\alpha_{v,p(v)}^{m,n}\beta_{p(v)\backslash v}^{(k+1\backslash k)}(n) \tag{13}$$

(6) If k=1, the computation stop. Otherwise, turn to step (4);

M-step algorithm: Through the E-step algorithm, the state likelihood $P(S_v^{(k)} = m \mid W^{(k)} = w, \lambda^{(k)})$ and the state transition probability

$P(S_v^{(k)} = m, S_{p(v)}^{(k+1)} = n \mid W^{(k)} = w, \lambda^{(k)})$ has known. Thus the parameter renew by MAP algorithm (where $v = (i, j)$)as follows(see the equation (14), (15), (16),(17)). Repeat E-step and M-step continuously, till the parameters are convergence. The training the HMT model is over. A realistic model that approximates the probability distribution of hidden state of flame image wavelet coefficient is constructed through the equation (18).

$$\delta_v(m) = \beta_v(k) \tag{14}$$

$$\delta_{v,\rho(v)}(m) = \max_{n=0,1}[\delta_v(n)a_{n,m} / p(s_v = n)]p(s_{\rho(v)} = m) \tag{15}$$

$$\varphi_v(m) = \arg\max_{n=0,1}[\delta_v(n)a_{n,m} / p(s_v = n)] \tag{16}$$

$$\hat{p} = \max_{n=0,1} \delta_1(n) \tag{17}$$

$$\hat{S}_1 = \arg\max_{n=0.1} \delta_1(n) \tag{18}$$

4 Experiment and Results

4.1 The Principle of Experiment

In this paper our research mainly focuses on the flame image polluted by the white Gaussian noise. We use the algorithm described in section 3 to train the noise flame image and obtain the parameters of HMT. We model the flame image wavelet coefficients as a Gaussian mixture model. In wavelet-domain, we assume that the noise that is transformed to wavelet-domain can be approximated as a Gaussian distribution and its parameters are computed by HMT model[5]. The wavelet coefficients influenced by noise express as follows:

$$y_i = c_i + n_i \tag{19}$$

Therefore the variance of noise flame image wavelet coefficients is the sum of variance of noise wavelet coefficients and variance of pure flame image wavelet coefficients.

$$\gamma_{q,i}^2 = \sigma_n^2 + \sigma_{q,i}^2 \tag{20}$$

Where $\gamma_{q,i}^2$ variance of noise image wavelet coefficients, σ_n^2 variance of noise wavelet coefficients, σ_{qi}^2 variance of pure flame image wavelet coefficients.

See Fig3, the wavelet coefficient represented by the hidden state S=0 is noise, and the wavelet coefficient that is represented by hidden state S=1 is signal. Use Bayesian estimation algorithm to obtain the maximum likelihood. The variance of noise σ_n^2 and

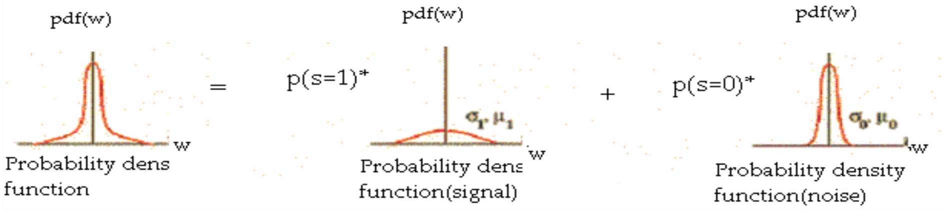

pdf(w)
w
Probability dens function
=
pdf(w)
p(s=1)*
σ_1, μ_1
w
Probability dens function(signal)
+
pdf(w)
p(s=0)*
σ_0, μ_0
w
Probability density function(noise)

Fig. 3. Noise and Signal Independent model

noise flame image $\gamma^2_{q,i}$ is confirmed. To compute the variance of pure flame image $\sigma^2_{q,i} = (\gamma^2_{q,i} - \sigma^2_n)$ and the noise is divided from image.

4.2 Results Analyze

In our paper, we also use other three methods to denoise, median filter denoise, wavelet denoise and Wiener filter denoise. Median filter is effective in smoothing the signal, but it is not accomplished in denoising the Gaussian white noise. The method of Wiener filter regards the signal and noise as random sequences. The optimum filter can be designed by these sequences,. But it is not an effective method for the wiping off the white noise, and this method usually makes the edge faintness. In wavelet-domain, we use some characters of image wavelet coefficients. The traditional method of wavelet transform is used some value δ to decide the transform coefficients is signal or not and don't c-are about the dependency between the various scales. In our experiment, we use the HMT to capture the statistical properties of the coefficients of wavelet transform, which is better than the method of simple using the wavelet coefficients. It is well suited to denoise the image containing singularities (edges and ridges)especially for flame image, and it provides a good denoise result and keeps the more details of the edge of flame than other methods. The denoised results are obtained by experiments (see Fig4).

Instead of comparing the denoise results directly from Fig 4, we used ratio of signal and noise I/MSE [6] to evaluate the denoise results of boiler image:

$$I/MSE = -10\log[\sum_i (I_i - \hat{I}_i)^2 / \sum_i I_i^2] \tag{21}$$

Where I_i is denoted original noise image, $\hat{I}_i$ is denoted denoise image and i is denoted pixel of image. If the value of I/MSE is bigger, the denoise result is better. Furthermore, what details of edge of image can be reserved is also a criterion to evaluating the algorithm of denoising. A parameters FOM [7] computed by equation(22)is used. If the value of FOM is bigger, the result is better.

$$FOM = 1/\max(N_1, N_2) * \sum_{i=1}^{N_1} 1/(1+\eta * d^2) \tag{22}$$

Where N_1 is number of the edge pixel in denoise results, N_2 is number of the edge pixel in ideal image, η is a constant and d is the vertical distance between the two flame image edge.

From the table 1, the result does suggest: (1) Since the image wavelet transform coefficients are sparse and the energy is concentrated in a few coefficients, the energy of white noise is dispersed to more wide scope and it can be model easily. Image denoising in the wavelet-domain can obtain the better result than traditional denoise method; (2) The principle of W-HMT denoise and traditional wavelet denoise is different. In traditional wavelet deonise algorithm, the threshold δ decides the wavelet coefficients signal or noise; In W-HMT, the statistical probability model is constructed to approximate the dependency between wavelet coefficients. After training the image W-HMT parameters, we compute the wavelet coefficients likelihood in the model parameters of W-HMT by the Bayesian algorithm and signal-noise independent model (see Fig4) and obtain the denoise results.

Table 1. Compare the performance of various denoise means

Criterion	Median filter denoise	Wiener filter denoise	Wavelet denoise	W-HMT denoise
FOM	0.1443	0.2073	0.2546	0.3385
I / MSE	22.0034	22.4836	24.4490	27.2495
FOM	0.1516	0.3092	0.3799	0.3950
I / MSE	14.9601	15.9490	16.2822	16.3393
FOM	0.1005	0.2001	0.2342	0.3270
I / MSE	11.2833	12.4175	13.0650	13.7078

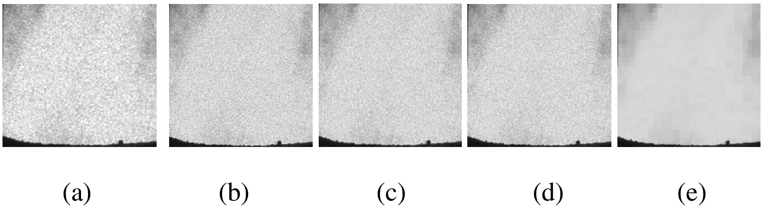


Fig. 4. The experiment results of flame image denoising; (a) the original noise image; (b) denoised by median filter; (c) denoised by wavelet; (d) denoised by Winner filter; (e) denoised by HMT

5 Conclusion

In this paper, we have introduced a new algorithm for boiler flame image denoising, base on the W-HMT model. By concisely modeling the statistical behavior of signals and noises at multiple scales, W-HMT denoising algorithm produces good denoise results and keeps the details of edge in flame images better. It is important for the boiler computer-control system to analyze the flame and supervise the boiler.

We must point out that we do not examine the other noise, since the white noise is usual to the boiler flame images. The limitation of the W-HMT model is that the data training consumes much time. Despite its preliminary character, the HMT model research can also be extended to other fields such as the 1-D signal denoises, detection of edge and image segment and so on.

References

1. Ji Chunguang, Wang peng and Li Shiyong.: Intelligent Control Method for Industrial Chain Boiler Based Information Fusion. Chinese Journal of Scientific Instrument, Vol 24, No.2, 131-134
2. Justin K.Romberg., Hyeokho Choi and Richard G.Baraniuk.: Bayesian tree-structured image modeling using wavelet domain hidden markov models. Submitted to IEEE Transactions on Image Processing ,maren (2000).
3. Yao Tianren.: Digital Speech Processing. HuaZhong University of Science and Technology Press(1992) 312-355 (In Chinese)
4. Jean-Baptiste Durand Paulo Goncalve's.: Statistical inference for hidden Markov tree models and Application to wavelet trees. Rapport de recherché
5. Micael Rochat.: Noise reduction in holographic microscopy images. Semester project .summer (2001). http://bigwww.epf1.ch/teaching /students/rochat/
6. Gagnon, L., Jouan, A.: Speck le filtering of SAR images-A comparative study between complex-wavelet based and standard filters [A]. In: SPIEProc [C]. Washington, U SA (1997) 3169: 80-91
7. Sattar, F., F lo reby, L., Salomonsson, G *et al.*: Image enhancement based on a nonlinear multiscale method[J]. IEEE Transactions on Image Processing (1997) 6 (1): 888-895.

Fast and Robust Portrait Segmentation Using QEA and Histogram Peak Distribution Methods

Heng Liu[1,2], David Zhang[3], Jingqi Yan[1], and Zushu Li[4]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University,
Shanghai, 200030, P.R.China
{hengliu, jqyan}@sjtu.edu.cn
[2] Southwest University of Scientist and Technology, Sichuan, 621000, P.R.China
[3] Department of Computing, The Hong Kong Polytechnic University,
Hong Kong, P.R.China
csdzhang@comp.polyu.edu.hk
[4] Institute of Intelligent Automation, Chongqing University,
Chongqing, 400044, P.R.China
zushuli@vip.sina.com

Abstract. Image segmentation is kernel part of image analysis and processing. Fast and robust portrait segmentation and fusion is still a challenging problem by far. In this paper, we present two fast and robust methods to segment portrait in blue background. One method is threshold obtaining based on histogram peak distribution, the other is Quantum Evolution Algorithm (QEA) threshold searching based on group variance. Detailed experiments and comparison analyses are presented to demonstrate the performance of our methods.

1 Introduction

Since color image segmentation plays a central role in image analysis and machine vision, many techniques have been developed for color image segmentation. Edge detection, seed filling and model match technologies are used usually for image segmentation [1]. As we know, edge detection, sometimes, will loose some faint edges information; seed filling and template matching methods are some time-consuming. Lim and Lee [2] proposed a fuzzy C-means algorithm to cluster color image pixels. Although this algorithm is unsupervised, it will lead to over segmentation when image is noisy or textured. The work in [3] gives a Markov random field (MRF) approach. In this approach, the components of the color pixel duplets are modeled as independent random variables. This reduces modeling complexity, but may be unrealistic for many real world images.

Face detection can be regarded as an essential step of face image segmentation. Chengjun Liu [4] uses the Bayesian statistical method to detect face. Dario Maio [5] takes directional image to find face features. Stan Z. Li [6] combines the FloatBoost learning and statistical methods to detect face. Juan Jose de Dios [7] manages to segment face in component color space. Although these methods are effective, most of

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 920 – 928, 2005.

them focus on face location and are not enough for exact color face segmentation due to varied image size and orientation.

Face indeed is an important feature in portrait; however, there exists a gap between face and portrait after all. Obviously, face detection method are not suited enough for portrait segmentation. As special shape segmentation, due to asymmetrical lighting and diverse dressing color, portrait segmentation is a harder task than ordinary color image segmentation, especially in complex background.

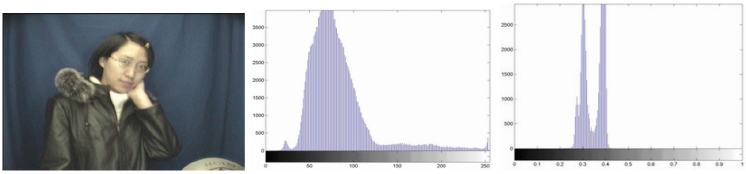
In this work, two methods for fast and robust threshold portrait segmentation were provided. Before that, blue tone space is defined first. One method is threshold obtaining based on histogram peak distribution. The other method introduces QEA [8] for searching threshold adaptively.

The rest of the paper is organized as follows. The blue tone clustering technology is discussed in Section 2. Two methods to segment portrait are discussed in detail in Section 3 and Section 4, respectively. Our experiment results and comparison analyses are presented in Section 5. Finally, a short conclusion is given in Section 6.

2 Blue Tone Clustering

Portrait segmentation can be regarded as a process that dividing portrait and background into their coherent regions. In our works, people are taken photos under blue background. According to the coherence of the pixels color, portrait and background should belong to respective color clustering theoretically. Nevertheless, due to the lighting source complexity, clothes color diversity and white balance of CCD effect, the portrait and the blue background are often mixed-up. Despite of that, blue tone histogram can be used to detach the portrait.

To eliminate the lighting influence, we can map the original color image into a blue tone space, in which we can extract portrait easily. If the blue background is high saturation, blue tone will not change much even if the environment lighting varies greatly. Then we can cluster the original image into two distinct clusters: portrait and background, as Fig. 1(c) shows. Blue tone is defined as Eq. (1).



(a) (b) (c)

Fig. 1. Original image histogram (b) and corresponding blue tone histogram (c)

$$P_{bt} = {Pixel.B}\Big/\left(Pixel.R + Pixel.B + Pixel.G\right) \tag{1}$$

where $Pixel[\cdot]$ is pixel component value of RGB. In most practical cases, due to the complexity of the environment, there are always over two peaks as the Fig. 2 Shows.

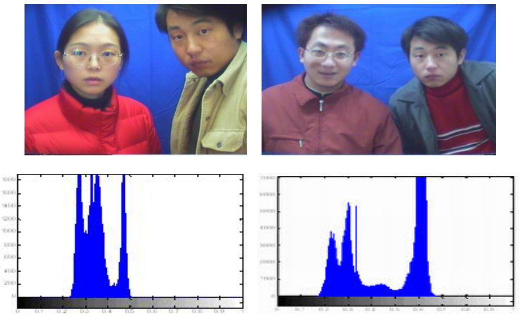


Fig. 2. Multi-peaks blue ton histogram

3 Threshold Obtaining Based on Histogram Peak Distribution

In multi-peaks blue tone histogram, two main peaks of the histogram are always the peak of portrait in the left and the peak of background in the right (for blue tone of the background is greater than that of the portrait). In addition, the lowest valley between two main peaks can be taken as threshold. Therefore, we can use this threshold to segment the portrait from the background. However, when we treat of the valley as threshold to segment the portrait, proper valley should be chosen skillful to exclude the noise valley.

Assume the histogram image is f(x). f(x) arrives at maximum when x satisfies $df/dx = 0$. Let x1 and x2 are the coordinates of two main peaks: portrait peak and background peak. Then, finding the suitable valley in [x1, x2] is the crucial problem.

Assuming the range of x value is [a, b], the idea peaks distribution is that portrait stands in left and background stands in right and the valley is just at the middle point $(a+b)/2$. Inspired by this assuming, we can choose the middle point coordinate as a coordinate conference point. Then most of valley choosing ways can be classified as six typical cases according to the coordinates' distribution of two main peaks to the middle point coordinate $(a+b)/2$. The sketch map is shown in Fig. 3.

Threshold obtaining method can be described in the flow:

(1)Search the histogram from left to right to find the coordinates of two main peaks: x1, x2.

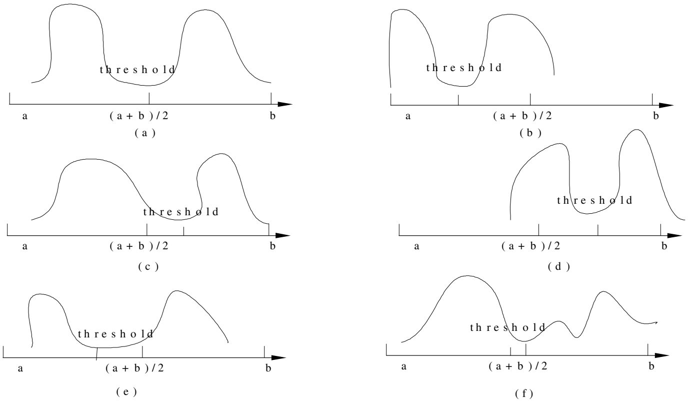


Fig. 3. Threshold choosing sketch map

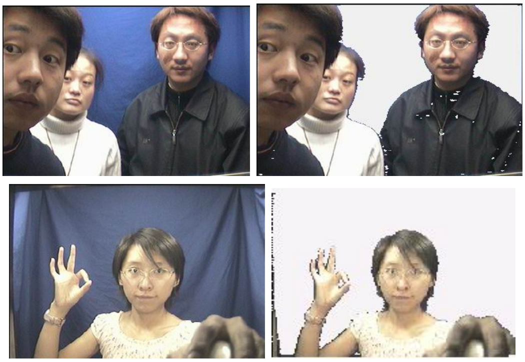


Fig. 4. The segmentation effect of threshold obtaining based on histogram peaks distribution

(2)If $x1 > (a+b)/2$, then find x in[x1, x2] or in[x2, x1] to satisfy $f(x) = \min(f(x))$; Else turn to (4).

(3)Treat of the x as the threshold and use it to extract portrait from original image. Then turn to (5).

(4)Search the histogram from right to left to find the coordinate of first peak: x3; then find x in[x1, x3] or in [x3, x1] to let f(x) be the minimum $f(x) = \min(f(x))$; then turn to (3).

(5)End.

When we apply the bidirectional global searching for threshold, the searching range is [0,255]. This algorithm can acquire acceptable effect while the segmentation speed is less than 15 frames/second(PIII 1.0G , 256M Memory). If we have some prior knowledge on searching range, the segmentation efficiency can be improved. In our indoor lighting environment, the range always lies in [30,150] .The segmented result are shown in Fig. 4.

4 Quantum Evolution Algorithm Threshold Searching Based on Group Variance

Analyzing the statistics attribute of histogram, we can get two useful characteristics. First, once the group variance [9] is defined as object function, the process for searching threshold is the procedure for maximizing the group variance. Secondly, if the histogram is regarded as a probability density function, the boundary of the threshold can be determined adaptively.

Suppose the image gray grade is from one to m. N_i is the number of pixels with gray grade i. Note: $N = \sum_{i=1}^{m} N_i$ is the total number of pixels; $P_i = N_i / N$ is probability of every gray grade; $u = \sum_{i=1}^{m} i \times P_i$ is gray mean of the whole image; $u(k) = \sum_{i=1}^{k} i \times Pi$ is gray mean when threshold is the k; $w(k) = \sum_{i=1}^{k} Pi$ is the total probability between [1, k]. Then the group variance as the object function is shown as Eq. (2).

$$\sigma^2(k) = \frac{[u \times w(k) - w(k)]^2}{w(k) \times [1 - w(k)]} \tag{2}$$

From the above notation, the density function $w(i), i \in (1, m)$ can be gotten. Then the range [a, b] of the threshold can be obtained adaptively: fix lower boundary a when w (i) arrives at zero at the last time, and fix upper boundary b when w (i) arrives one at the first time.

To assure the real time portrait segmentation in a real time way, quantum evolution algorithm is applied for threshold searching. Quantum evolution algorithm is a probabilistic algorithm inspired by quantum computing. In quantum computing, the basic unit is called Q-bit.

A Q-bit can stay in two basic states ‘0’ and ‘1’ or any superposition of the two basic states. A Q-bit can be represented as Eq. (3).

$$|\Psi\rangle = \alpha |0\rangle + \beta |1\rangle \tag{3}$$

where α, β is the probability amplitude of the corresponding state.

Normalization of the state unity is:

$$|\alpha|^2 + |\beta|^2 = 1 \tag{4}$$

An m Q-bits system, which is called Q-bit individual, can be defined in Eq. (5).

$$\begin{bmatrix} \alpha_1 & \alpha_2 \ldots \alpha_m \\ \beta_1 & \beta_2 \ldots \beta_m \end{bmatrix} \tag{5}$$

where $|\alpha_i|^2 + |\beta_i|^2 = 1, i = 1...m.$ This format can easily represent the superposition of states. That also means the m Q-bits system can contain 2^m states probability information. QEA maintains population as $Q(t) = \{q_1^t, q_2^t, \cdots q_n^t\}$at generation t. In QEA, a rotation operation Q-gate can be defined in Eq. (6) to update the Q-bit individuals.

$$\begin{bmatrix} \alpha_i^{'} \\ \beta_i^{'} \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \tag{6}$$

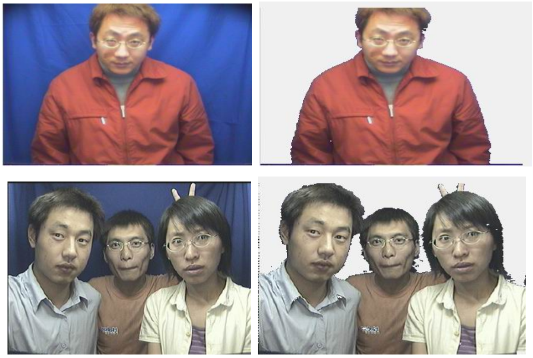


Fig. 5. The segmentation effect of QEA threshold searching based on group variance

where $\Delta\theta_i, i = 1,2\cdots m$, is the rotation angle of each Q-bit toward '0' or '1' state depending on its sign.

With applying QEA, the segmentation effect is satisfactory. And, segmentation speed is about 20 fame/seconds (PIII 1.0G , 256M Memory). The QEA segmentation results are shown in Fig. 5.

5 Experiments and Comparisons

The effects of two segmentation methods are contrasted in Fig. 6 and the comparison results are listed in Tab. 1. According to the comparisons, we see the method of peak distribution is finer than QEA method. Nevertheless, the method of QEA threshold searching is more efficient as for segmentation adaptability and time.

To show the efficiency of QEA threshold searching method, the comparisons between threshold searching of QEA and GA (genetic algorithm) are shown in Fig. 7 and Tab. 2.

Table 1. Segmentation comparison

methods	Accuracy	Adaptability (light varying)	Invalidation (per 50 images)	Speed(f/s)
Threshold obtaining	finer	No	7%	< 15
QEA searching	fine	Yes	3%	20
GA searching	fine	Yes	5%	16

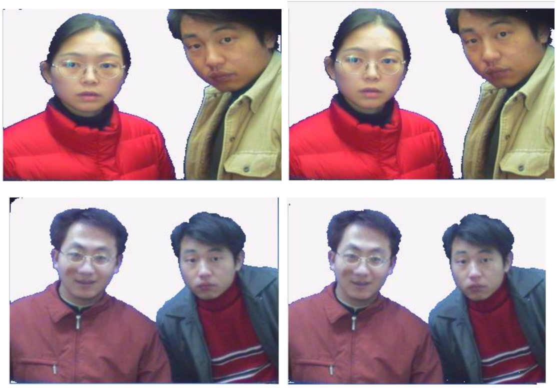

(a) (b)

Fig. 6. Segmentation contrast effect. (a) Peak distribution threshold obtaining. Hair and edge contained more finely; (b) Quantum evolution algorithm threshold searching. Some information of hair and edge was lost.

From comparisons, we know, though GA also has fair performance but it needs more generations to get the solution and is easy to surge, while QEA only needs fewer generations and is more robust.

Finally, segmentation comparisons with traditional methods including edge detection and seed filling are shown in Fig. 8. From these results, we see our segmentation methods are very competitive.

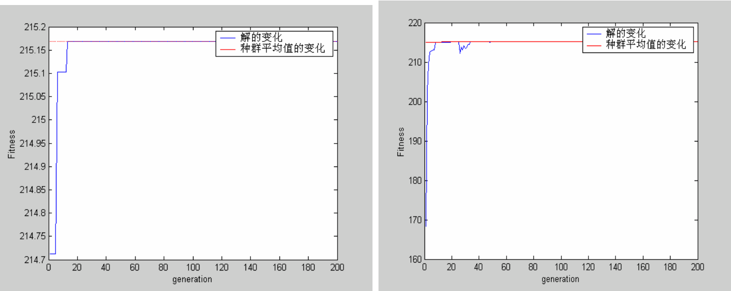
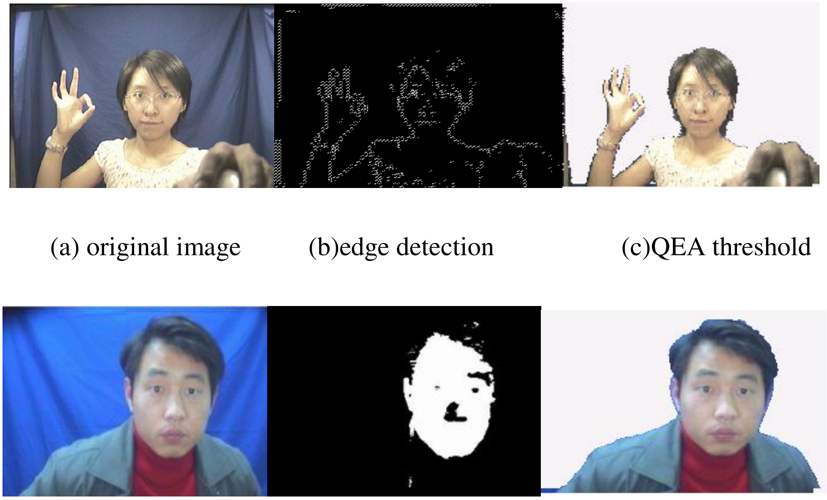


Fig. 7. Threshold solution comparison between QEA (left) and GA (right)

Table 2. Performance comparison between QEA and GA in threshold searching

Parameters	QEA segmentation	GA segmentation
Population size	5,10,8...	5,10,20...
Individual length	4,8,6...	4,8,8...
Generations for threshold	8,7,7...	38,35,20...
Speed (f/s)	20	16



(a) original image (b)edge detection (c)QEA threshold

(d) original image (e) seeds filling (f)QEA threshold

Fig. 8. Segmentation comparisons with edge detection and seeds filling methods

6 Conclusion

This paper studies on fast and robust approaches for portrait segmentation. Two threshold-acquiring methods are presented to segment portrait, one of which introduces quantum evolution algorithm for searching threshold adaptively based on group variance. In future, we focus our work on portrait segmentation and portrait beautification under complex background.

Acknowledgment

We would like to thank for all those people who have contributed to this work by providing their data and comments, especially thanks for my team partners at Chongqing University. This work was partly supported by National Science Fund of China (No.60402020), and Medium-Small Enterprise Innovation Fund of Chinese Ministry of Science and Technology (No. 01C26225110588).

References

1. Forsyth, D. A., Jean Ponce: Compute Vision: A Modern Approach. Prentice Hall (2003)
2. Lim, Y.W., Lee, S.U.: On the Color Image Segmentation Algorithm Based on the Thresholding and the Fuzzy C-means Techniques. Patten Recognition 23 (9) (1990) 935-952
3. Chang, M.M., Sezan, M.I., Tekalp, A.M.: Adaptive Bayesian Estimation of Color Images. Electron Imaging 3 (1994) 404-414
4. Chengjun Liu: A Bayesian Discriminating Features Method for Face Detection. IEEE Trans. Pattern Analysis and Machine Intelligence 25 (6) (2003) 725-740
5. Dario Maio, Davide Maltoni: Real-time Face Location on Gray-scale Static Images. Pattern Recognition 33 (9) (2000) 1525-1539
6. Stan, Z.Li, ZhenQiu Zhang: FloatBoost Learning and Statistical Face Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004)
7. Juan Jose de Dios, Narciso Garcia: Fast face segmentation in component color space. International Conference on Image Processing, ICIP (2004) 191–194.
8. Kuk-Hyun, Han: Genetic Quantum Algorithm and Its Application to Combinatorial Optimization Problem. Evolutionary Computation Proceeding 2 (2000) 1354-1360
9. Otsu, N.A: Threshold Selection Method from Gray-level Histograms. IEEE□SMC 9 (l) (1979) 62-66

Suppressing Chaos in Machine System with Impacts Using Period Pulse

Linze Wang[1], Wenli Zhao[1], and Zhenrui Peng[2]

[1] School of Computer Science, Hangzhou Dianzi University,
Hangzhou Zhejiang 310018, China
aozhwlz@hzcnc.com
[2] Zhejiang University, Hangzhou Zhejiang 310027, China
zrpeng@iipc.zju.edu.c

Abstract. In this paper we have studied the suppression of chaotic vibration in the system with impacts. Period pulse method of chaos suppression in the system with impacts has been presented. We consider the stable fixed point as the control target. Once the state signals derivate from the control target, the system dynamically produces period impulse signals to suppress chaos and bifurcation. The method has been employed to a two-degree-of–freedom reciprocating impact vibration model. Using the stable fixed point of Poincaré map equation as the control target, suppressing bifurcation and chaos under different parameters by numerical simulation. The results show that the method can suppress chaotic motion effectively.

1 Introduction

There inevitably exists clearance between the mechanical components (mechanical parts) or boundaries of different components. In the running of mechanical system, there are reciprocating collisions due to the clearance. In design stage, the reasonable clearance is permitted. With the change of design parameters such as wear, deformation, fracture, drop out and aging of these components, or with the running condition change of the system itself such as the change of motion velocity, the originally stable reciprocating periodic motion will be able to evolve into abnormal chaotic vibration. The abnormal system vibration can influence the stability, safety, and longevity of system. Finding these abnormal chaotic vibration in time and suppressing them can effectively retard the failure.

As for the system with impacts, much more attention has been paid to the nonlinear dynamics mechanism, stability, and the bifurcation and chaos motion of different parameters in recent years[1]~[5]. A number of methods have been developed to suppressing chaotic vibration[6]~[9]. But we haven't find literature reports on the chaos suppressing method of mechanical system with impacts.

In this paper we present a method for suppressing bifurcation or chaos motion of the system with impacts, which can implement real time monitoring, produce periodic impulse signal. By means of the analysis and simulation of a typical reciprocating impact vibration model, the validity of this method is investigated.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 929–938, 2005.

2 Basic Idea of Periodic Impulse Suppressing Chaos

Consider a general nonlinear mapping system

$$X_{n+1} = F(\mathbf{X}_n) \tag{1}$$

where $X \in R^N$.

Starting from an initial point X_0, after pth iteration, the system (1) can be written as

$$X_P = F^{(P)}(X_0) \tag{2}$$

where $F^{(P)} \equiv \underbrace{F(F(F(\cdots F}_{p\text{ times}}(X)\cdots)))$.

The convergence property after pth iteration from the initial points can be known from the Jacobi matrix formulation of mapping (2) .The stability condition of periodic P orbit is: All moduli of Jacobi matrix eigenvalues are less than 1. Assuming the system exits stable P(P=1,2,3...)periodic orbit, from Eq.(2), we know $X_P = X_0$. If the variation of system parameters causes chaos motion of the system, mapping result will deviate from original stable periodic points. This is $X_P \neq X_0$.

The idea of our control method is: impulse signal Q is presented to the system at the P period point.

$$Q = X_P - X_0 \tag{3}$$

Thus, the orbit, which starts from an initial point X_0, after p-th iteration combined with impulse signal Q, will be

$$F^{(P)}(X_0) - Q = X_0 \tag{4}$$

When $X_P = X_0$, $Q = 0$, there is no control.

When $X_P \neq X_0$, $Q \neq 0$, Q will counteract departure. Then the system maps continuously according to Eq. (2). Repeating the above process, it is equal to present a period pulse to system .The period pulse can be written as follow

$$Q_n = \sum_{m=0} Q\delta(n - mP) \tag{5}$$

$m = 0$ is the moment when the control is applied.

Impulse signal Q makes the orbit end to end, thus forms disturbed periodic orbit $X_0, X_1 + \Delta X_1, X_2 + \Delta X_2, \cdots, X_{P-1} + \Delta X_{P-1}$ and suppresses the chaos eventually. The purpose of this method is not to make the control system return to the original stable periodic orbit but to make it be in periodic motion all the time so to suppress the chaos.

As for reciprocating impact vibration system with clearance, take one reciprocating impaction as a period, P=1. When the system is stable, it satisfies

$$X_0 = F(X_0) \tag{6}$$

Fig.1 illustrates the block diagram of periodic impulse suppressing chaos. X_i denotes the input at i moment and X_0 denotes the stable output vector (control target). *E* denotes noise from measurement error, time delay error, and other factors.

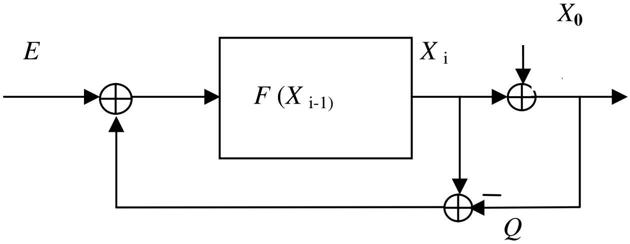


Fig. 1. The Block diagram of periodic impulse suppressing chaos

X_i is detected in real time. If $X_i = X_0$, then $Q = 0$, there is no control; if $X_i \neq X_0$, then $Q = \Delta X_i$. Where $\Delta X_i$is the derivation of X_i from X_0 and the error ΔX_i is used impulse control signal.

3 Two-Degree-of–Freedom Reciprocating Impact Vibration Model

In order to validate the above method for suppressing the chaos of reciprocating impact system with clearance, take the model (Fig.2) as the object.

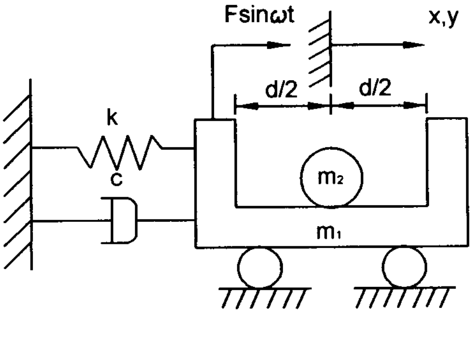


Fig. 2. Model of the system

The system consists of a primary rigid body with mass m_1, a rigid sphere with mass m_2, a linear spring with stiffness *k, and* viscous damper with damping constant *c*. Assuming that no friction exists between these two masses. The clearance between two masses is denoted by d and an external harmonic force, $F\sin\omega t$, is applied to the primary mass. The absolute motion of the primary mass is x .The absolute motion of the rigid sphere is y . Between two consecutive impacts, the differential equation of this dynamical system is obtained.

$$\left.\begin{aligned} &m_1\ddot{x}(T)+c\dot{x}(T)+kx(T)=F\sin(\Omega T+\alpha) \\ &m_2\ddot{y}(T)=0 \end{aligned}\right\} \tag{7}$$

From the condition of conservation of momentum and the coefficient of restitution during the impact, the impact equation can be obtained.

$$\left.\begin{aligned} \dot{x}_+ &= \dot{x}_- + \mu(\dot{y}_- - \dot{y}_+) \\ \dot{x}_+ &= \dot{y}_x + R(\dot{y}_- - \dot{x}_-) \end{aligned}\right\} \tag{8}$$

where $\mu = \dfrac{m_2}{m_1}$ is the mass ratio, *R* is the coefficient of restitution. $\dot{x}_-$ and $\dot{y}_-$ denote the instant velocity of the two masses before impact, and $\dot{x}_+$ and $\dot{y}_+$ denote the instant velocity of the two masses after impact respectively.

In the above two equations, let

$$x_0 = \frac{F}{k},\ x_1 = \frac{x}{x_0},\ x_2 = \frac{y}{x_0},\ h = \frac{c}{2\sqrt{m_1 k}},\ \omega_n = \sqrt{\frac{k}{m_1}},\ \omega = \frac{\Omega}{\omega_n},\ t = \omega_n T\ ,$$

then the equation (7)~(8) can reformulated as the following non dimensional form:

$$\left.\begin{aligned} &\ddot{x}_1(t)+2h\dot{x}_1(t)+x_1(t)=\sin(\omega t+\alpha) \\ &\ddot{x}_2(t)=0 \end{aligned}\right\} \tag{9}$$

$$\left.\begin{aligned} \dot{x}_{1+} &= \dot{x}_{1-} + \mu(\dot{x}_{2-} - \dot{x}_{2+}) \\ \dot{x}_{1+} &= \dot{x}_{2+} + R(\dot{x}_{2-} - \dot{x}_{1-}) \end{aligned}\right\} \tag{10}$$

When the external harmonic force is applied to the primary rigid body and the rigid sphere impact with the primary body, then according to kinematics relation, impacts occur at $x_2 - x_1 = \pm d_0$. Where $d_0 = d/(2x_0)$.

Take the impact plane after right impact as the Poincaré section and take the moment after impact as the moment of the period τ ,

$$\sigma = \left\{x_1, \dot{x}_1, \dot{x}_2, \tau\right) \in R^3 \times S\ , x_2 = x_1 + d_0, \tau = \tau_+ \right\}.$$

According to boundary constraints and equations(9)~(10), the mapping $F:\sigma \to \sigma$ can be built as

$$X_{n+1} = F(v, X_n) \tag{11}$$

where v is a real parameter, $v \in R^1$, $X = (x_1, \dot{x}_1, \dot{x}_2, \tau)^T$,the stable state is corresponding to stable fixed point, then the formulation(11)should satisfy :

$$X^* = F(v^*, X^*) \tag{12}$$

where $X^* = (x_1^*, \dot{x}_1^*, x_2^*, \tau^*)^T$ is the coordinate of stable fixed point in the Poincaré section. v^* is parameter at the fixed point of system.

If $\Delta X = (\Delta x_1, \Delta \dot{x}_1, \Delta \dot{x}_2, \Delta \tau)^T$ is the perturbations of the steady state at $\tau = 0_+$. Then the next perturbations of steady states just after the next impact at Poincaré section can be written as follows

$$\left.\begin{aligned} \Delta x_1' &= f_1(\Delta x_1, \Delta \dot{x}_1, \Delta \dot{x}_2, \Delta \tau) \\ \Delta \dot{x}_1' &= f_2(\Delta x_1, \Delta \dot{x}, \Delta \dot{x}_2, \Delta \tau) \\ \Delta \dot{x}_2' &= f_3(\Delta x_1, \Delta \dot{x}_1, \Delta \dot{x}_2, \Delta \tau) \\ \Delta \tau' &= f_4(\Delta x_1, \Delta \dot{x}_1, \Delta \dot{x}_2, \Delta \tau) \end{aligned}\right\} \tag{13}$$

where $\Delta X' = (\Delta x_1', \Delta \dot{x}_1', \Delta \dot{x}_2', \Delta \tau')^T$ is the next derivative from stable fixed point. Here the Jacobi matrix will be employed to analyze the stability of the periodic motion. It can be written in the form

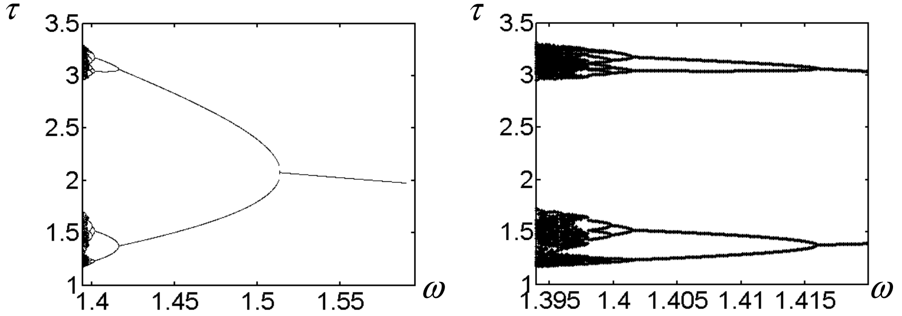
$$DF(v,0) = \begin{bmatrix} \frac{\partial f_1}{\partial \Delta x_1} & \frac{\partial f_1}{\partial \Delta \dot{x}_1} & \frac{\partial f_1}{\partial \Delta \dot{x}_2} & \frac{\partial f_1}{\partial \Delta \tau} \\ \frac{\partial f_2}{\partial \Delta x_1} & \frac{\partial f_2}{\partial \Delta \dot{x}_1} & \frac{\partial f_2}{\partial \Delta \dot{x}_2} & \frac{\partial f_2}{\partial \Delta \tau} \\ \frac{\partial f_3}{\partial \Delta x_1} & \frac{\partial f_3}{\partial \Delta \dot{x}_1} & \frac{\partial f_3}{\partial \Delta \dot{x}_2} & \frac{\partial f_3}{\partial \Delta \tau} \\ \frac{\partial f_4}{\partial \Delta x_1} & \frac{\partial f_4}{\partial \Delta \dot{x}_1} & \frac{\partial f_4}{\partial \Delta \dot{x}_2} & \frac{\partial f_4}{\partial \Delta \tau} \end{bmatrix}_{(v,0.0.0.0)} \tag{14}$$

If the moduli of all the eigenvalues of $DF(v, 0)$ are less than unity, then the periodic motion is stable; otherwise, it is unstable. As the system parameters vary, the modulus of one of the eigenvalues of $DF(v, 0)$ may pass through the unit circle and the bifurcation occurs.

Refs.[2-3] have been concerned with the stability, bifurcation and chaos motion of the system .In the present work we focus attention on the suppression of chaos in the system.

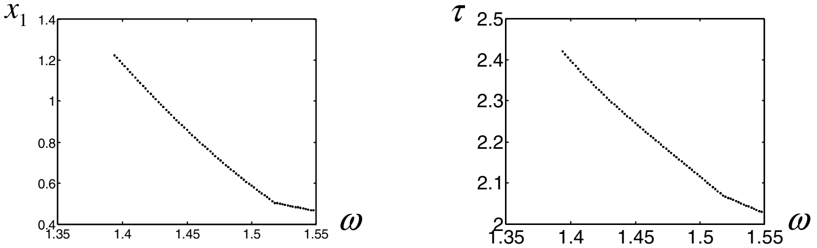
4 Numerical Simulation

The next step is to consider the effect on periodic impulse signal suppressing the chaos of system when V derivates from V^* and the bifurcation or chaos occurs in the system. Consider the ideal case of $E = 0$ in Fig.1 and the stable fixed point as the control target. Once the state signals derivate from the control target, the system dynamically produces period impulse signals to suppress chaos and bifurcation.



(a) Period-doubling bifurcation diagram of ω (b) Local zooming-in bifurcation diagram

Fig. 3. Period-doubling bifurcation diagram of ω varying from 1.58 to 1.398



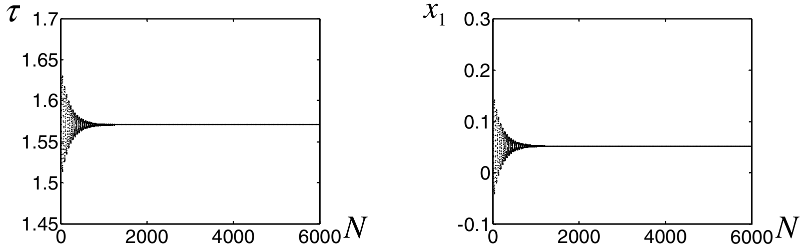
(a) The variation of x_1 with ω (b) The variation of impaction period τ with ω

Fig. 4. Control effect from ω =1.55 to ω =1.398

From Fig.3, in the region $1.4 < \omega < 1.58$, when no control is applied, the process of system enters chaos through period-doubling sequence with the decreasing of ω. $V = \omega$ is the bifurcation parameter. For $\omega > 1.515$, the left and right impact of system is completely symmetrical. That is, the system of two-degree of freedom retains stable 1-1-1 period motions. With the parameter ω reducing, symmetric impacts become asymmetric one and period-doubling bifurcation occurs. For $\omega < 1.398$, the chaos occurs in the system. Fig.3(a) shows the system period-doubling bifurcation diagram with the variation of ω. Fig.3 (b) is the local zooming-in bifurcation diagram.

Fig.4 illustrates the control effect after presenting the control to the system with the same parameters as that in the above system. The simulation starts from $\omega = 1.55$. In order to examine the long-term stability, take every ω iteration times as 2000.

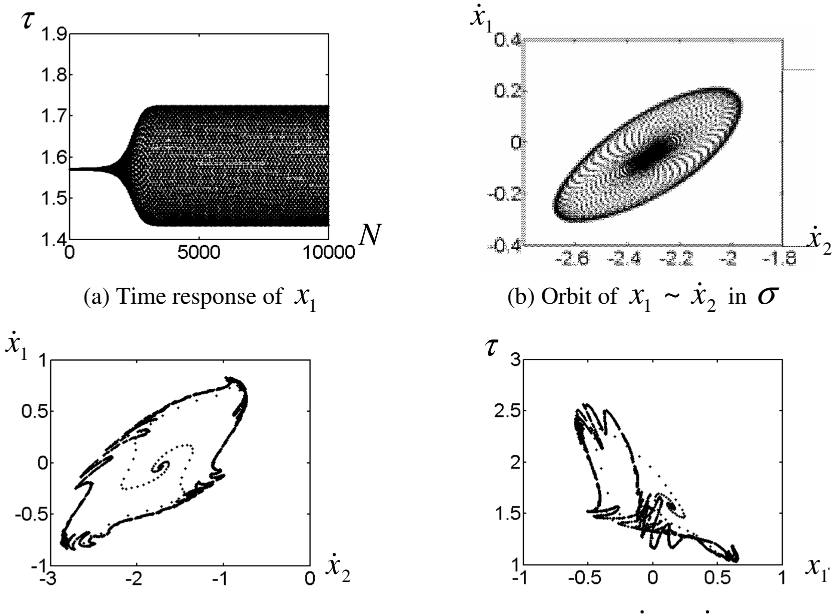
Compare Fig.3 with Fig.4, owing to the control, the original bifurcation and chaos motion are suppressed.



(a) $N \sim \tau$

(b) $N \sim x_1$

Fig. 5. Time response of τ and x_1 when the system in stability



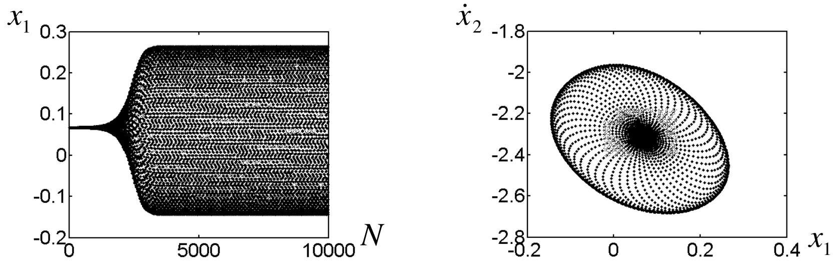
(a) Time response of x_1

(b) Orbit of $x_1 \sim \dot{x}_2$ in σ

(c) Time response of τ

(d) Orbit of $\dot{x}_2 \sim \dot{x}_1$ in σ

Fig. 6. Hopf bifurcation diagram when k=4.9



(a) $\dot{x}_2 \sim \dot{x}_1$ (b) $x_1 \sim \tau$

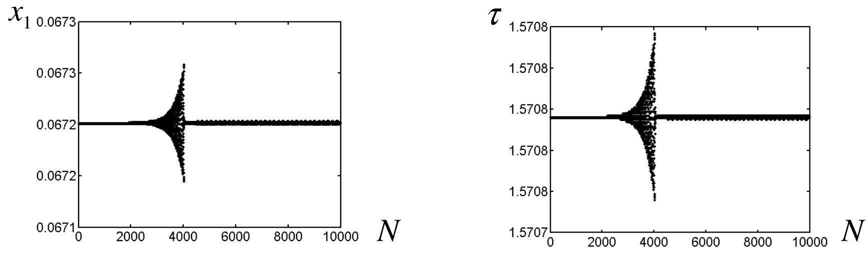
Fig. 7. Chaos attractor in σ when k=3.4

Fig.5~Fig.7 show the simulation result that the system varies from stable fixed point to HOPE bifurcation and even chaos. Take spring coefficient k as bifurcation coefficient. Take the system parameter $(\omega, F, R, d, c, \mu) = (2,7,0.8,5,0.1,0.1)$; for $k > 5$, the system is in stable 1-1-1 period symmetric impaction motions. Hence Poincaré diagram is a fixed point. Fig.5 (a) and Fig.5 (b) show the trend that mapping value τ and x_1 varies with time *N* respectively. This figure is called time response for short. The variations of other variables are similar to this, so not described here. Obviously, if abandon the transition process caused by initial disturbance, the Poincaré section diagram from *N*>1000 is a fixed point.

When *k* decreases gradually, the system will appear Hopf bifurcation. When *k*=4.9 and $\Delta x_1 = 0.001$, the numerical result is shown in Fig. 6. From Fig.6, we know that even there is a little Δx_1, the system will tend to be in immovability circle.

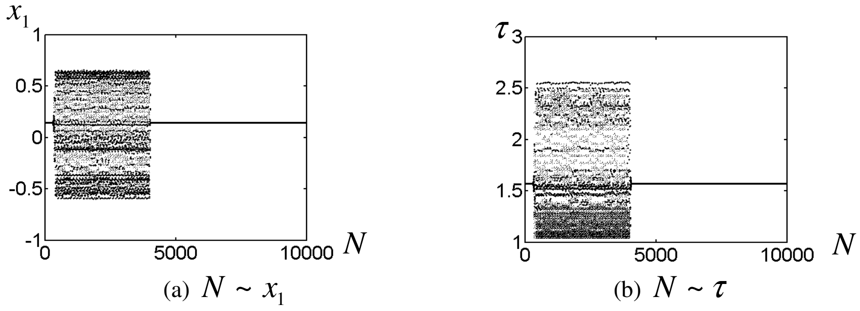
If *k* decreases continuously, the immovability circle will begin distort, and eventually chaos motion will take place. When *k*=3.4, the orbit in σ starting from stable fixed points is shown in Fig.7 and chaos motion happens.

Fig. 8 and Fig. 9 shows the simulation results after acting control on the above bifurcation and chaos motion. Fig. 8 shows the control result while *k*=4.9. Fig. 9 shows the control result while *k*=3.4. In order to observe the evolvement process clearly, the control acts on when N=4000.



(a) $N \sim x_1$ (b) $N \sim \tau$

Fig. 8. Time response after control when k=4.9



(a) $N \sim x_1$ (b) $N \sim \tau$

Fig. 9. Time response after control when k=3.4

From Fig.8 and Fig.9, the control can suppress chaos effectively. The method needs little time from control action to chaos suppressing and the time response is very fast within one-decade impact periods. Owing to length limitation of this article, simulation results of other variables are not listed here. But the characteristics of other variables have the same as that discussed here and have agreement with the conclusion here.

5 Conclusion

Through the simulation of a typical reciprocating impact model, this method was validated. The result shows that this method has a good suppressing effect to bifurcation and chaos motion in machine system with impacts. For reciprocating impact machine system with clearance, its parameters change gradually in long-term running. Our work is based on the several kinds of bifurcation and chaos motion, which nearly approximates engineering practice.

This method is not limited to reciprocating impact machine system with clearance. It can also be applied to other similar nonlinear system.

Acknowledgement

This work is granted by natural science foundation of Zhejiang Province (NO.502144) and education scientific planning project of Zhejiang education bureau (NO. 20040459).

References

1. Peterka, F., Vacik,J.: Transition to Chaotic Motion in Mechanical Systems with Impacts. Journal of Sound and Vibration 1 (1992) 95-115
2. Zhao Wenli, Wang Linze: Hopf Bifurcation of an impact Damper. Proceeding of the 3rd International conference on nonlinear Mechanics (1998) 437-440
3. Sung, C.K, Yu ,W. S.: Dynamics of a Harmonically Excited Impact Damper: Bifurcations and Chaotic Motion. Journal of Sound and Vibration 2 (1992) 317-329

4. Lin, S.Q., Bapat, C.N.: Estimation of Clearances and Impact Forces Using Vibroimpact Response. Journal of Sound and Vibration 3 (1993) 407-421
5. Guanwei Luo, Jianhua Xie: Bifurcation and Chaos in a System with Impacts. Physica D. 148 (2001) 183-200
6. Lima, R, Pettini, M.: Suppression of Chaos by Resonant Parametric Perturbations. Phys Rev A. 41 (1990) 726-733
7. Braiman, Y., Goldhirsch, I.: Taming Chaotic Dynamics with Weak Periodic Perturbations. Phys Rev Lett. 66 (1991) 2545-2549
8. Kapitaniak, T.: Controlling Chaotic Oscillators without Feedback. Chaos, Soliton, Fractals. 2 (1992) 519-530
9. Yong Xu, Gamal, M., Mahmoud, Wei Xu, Youming Lei: Suppressing Chaos of a Complex Duffing's System Using a Random Phase . Chaos, Solitons and Fractals 23 (2005) 265-273

The Cognitive Behaviors of a Spiking-Neuron Based Classical Conditioning Model

Guoyu Zuo, Beibei Yang, and Xiaogang Ruan[1]

Institute of Artificial Intelligence and Robotics,
Beijing University of Technology,
Beijing 100022, China
{gyzuo, annyoung}@emails.bjut.edu.cn
adrxg@bjut.edu.cn

Abstract. A spiking-neuron based cognitive model with classical conditioning behaviors is proposed. With a reflex arc structure and a reinforcement learning method based on the Hebb rule, the cognitive model possesses the property of 'stimulate-response-reinforcement' and can simulate the learning process of classical conditioning. An experiment on the inverted pendulum validated that this model can learn the balance control strategy by classical conditioning.

1 Introduction

Classical conditioning is the basic form of learning which was firstly reported by Ivan Pavlov, a Russian biologist, in 1927. From then on, many researchers have devoted to this field and proposed several learning models. Roscorla and Wagner proposed the first computational model of classical conditioning in 1972 [1]. Sutton and Barto [2] proposed both the Time-Derivative model and the Temporal-Difference model in 1981 and 1987 respectively. The former is an early reinforcement model in conditioning, as an extension of the former, the goal of conditioning in the latter model is to predict the temporally discounted value of all future rewards, and the model can successfully model the inter-stimulus interval (ISI) dependency. Klopf introduced the Drive Reinforcement (DR) model, which separates inhibitory and excitatory learning and can simulate the secondary conditioning and the reacquisition effect, in 1982 [4]. Schmajuk and DiCarlo introduced a model in 1992 [5], which was able to model a number of classical conditioning phenomena and especially to model the effect of various types of configurational stimuli and the effects of hippocampal lesion on conditioning. Balkenius presented a model based on a neural interpretation of the conditioning mechanism in 1995 [6], which can handle secondary conditioning and blocking very well. However, it can neither model reacquisition effects nor distinguish between delay and trace conditioning. Johansson and Lansner presented a classical conditioning model that is composed of some interconnected Bayesian confidence propagating neural networks (BCPNNs) in 2002 [7], by implementing Hebbian learning, which is able to make a closer tie between the output and the underlying neural activity.

[1] This research is supported by National Natural Science Foundation of China (60375017).

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 939 – 948, 2005.

Aiming at simulating the spiking transfer and process mechanism of information in biological nervous system, this paper proposes a spiking-neuron based cognitive model with classical conditioning behaviors. With a reflex arc structure and the implement of Hebbian reinforcement learning, the cognitive model possesses the property of ‘stimulate-response-reinforcement’ and is able to simulate the learning process of classical conditioning. An experiment on the inverted pendulum validated that this model can learn the balance control strategy by classical conditioning.

2 Architecture of the Classical Conditioning Model

The Architecture of the Classical Conditioning Model (CCM) is showed in Figure 1, where

- Reflex Agent (RA) is constructed according to the reflex arc of the biological nervous system and is used to realize the reflex function, i.e., generate the response for the stimulus.
- Learning Agent (LA) implements the Hebbian learning and modifies the connections weights of RA.
- Evaluate Agent (EA) evaluates the learning results and instructs the modifications in LA.

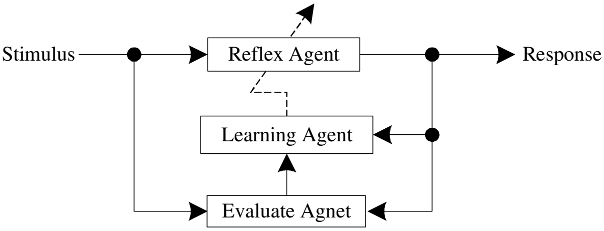


Fig. 1. Architecture of the classical conditioning model

Therefore, the Classical Conditioning Model can be described as a triple:

$$\text{CCM=<RA, EA, LA>.} \tag{1}$$

where

- $\text{RA}:\mathbf{S}\rightarrow R$ is the mapping from stimulus (S) to responses (R).
  $S=\{US,CS_1,\ CS_2,\cdots,CS_n\}$ is the set of stimulus, in which
  - US is the unconditioned stimulus
  - $CS_i(i=1,\ 2,\cdots,\ n)$ is the conditioned stimulus.
- $\text{EA}:\{\mathbf{S},R\}\rightarrow v$ is the mapping from the 2-tuple {S,R} to the evaluation v.
- $\text{LA}:\mathbf{W}\rightarrow\mathbf{W}$ will modify the weight matrix W to achieve learning.

Classical conditioning is a cognitive process. CCM can only respond to US in the beginning, but it can gradually exhibit the cognitive behavior and set up the classical conditioning during the iteration process of stimulus- response-reinforcement.

3 Reflex Agent and Spiking Neuron

3.1 The Reflex Agent

The structure of the reflex agent is showed in Figure 2, where

- Stimulus Neuronal Group (SNG) acts as the receptor and the afferent nerve of the biological reflex arc. They receive the outside stimulus and transform this stimulus to the nerve impulses.
- Central Processing Neuronal Group (CPNG) acts as nerve center, they process the received nerve information.
- Response Unit (RU) acts as efferent nerve and effectors, they transform the nerve information to responses.

No inner connections exist in each group, and connections exist only between the neuronal group of SNG and CPNG and between the neuronal group of CPNG and RU.

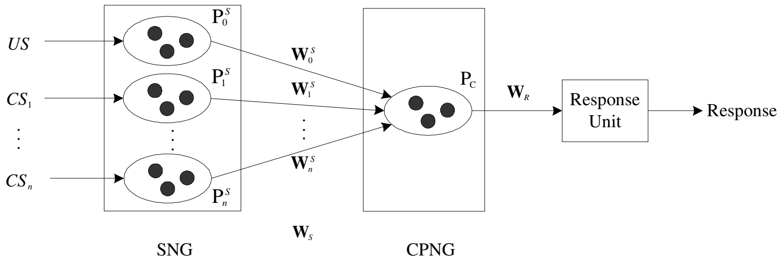


Fig. 2. Structure of RA

Also, the reflex agent can be described as a 4-tuple:

$$\text{RA}=<\text{SNG, CPNG, RU, W}>. \tag{2}$$

where

- SNG=$\{P_k^S \mid k=0,1,\ldots,n\}$, where
  - $P_0^S=\{v_i^{S(0)} \mid i=1,2,\ldots,n_0\}$: the unconditioned stimulus neuronal group (USNG), where $v_i^{S(0)}$ denotes the ith neuron of P_0^S.
  - $P_k^S=\{ v_i^{S(k)} \mid i=1,2,\ldots,n\}(k=1,2,\ldots,n_k)$: the conditioned stimulus neuronal group (CSNG), where $v_i^{S(k)}$ denotes the ith neuron of P_k^S
- CPNG=$\{P_C\}$, where
  - $P_C=\{ v_i^C \mid i=1,2,\ldots,n_C\}$: the set of central process neurons where n_C is the number of neurons
- RU=$\{v^R\}$, composed of a response neuron
- $\mathbf{W}=\{\mathbf{W}_S,\mathbf{W}_R\}$: weight matrix，where
  - $\mathbf{W}_S=\{\mathbf{W}_k^S=(w_{ij}^{S(k)})_{n_C\times n_C} \mid k=0,1,\ldots,n\}$: the matrix between P_k^S and P_C.
  - $\mathbf{W}_R=(w_{ij}^R)_{n_R\times n_C}$: the weight between P_C and RU;
  - $\mathbf{W}_k^S(k=1,2,\ldots,n)$ can be modified, $\mathbf{W}_0^S$ and $\mathbf{W}_R$ is constant matrix.

3.2 The Spiking Neurons

All neurons in this module adopt the Spiking Response Model developed by Gerstner [8], which can simulate the electrical pulse of biological nerve information and transmit information in the form of spiking.

In the Spike Response Model (SRM), the state of a spiking neuron is described by the membrane potential u_i of neuron i. u_i will remain at the resting potential u_{rest}. If there is neither a spike arrives nor an input. Once there is an incoming excitatory or a spike emitted by a pre-synaptic neuron making u_i exceed the firing threshold θ, the neuron will emits a spike. This process is described as follows:

$$t = t_i^{(f)} \Leftrightarrow u_i(t) = \theta \text{ and } \frac{\mathrm{d}u_i(t)}{dt} > 0 \tag{3}$$

where, $t_i^{(f)}$ is the spike time train of neuron i.

The evolution of u_i is decided by three parts, the potential caused by the last spike, the potential caused by other neurons and the potential caused by external inputs:

$$u_i(t) = \eta(t - \hat{t}_i) + \sum_j w_{ij} \sum_f \varepsilon_{ij}(t - \hat{t}_i, t - t_j^{(f)}) + \int_0^\infty \kappa(t - \hat{t}_i, s) I^{ext}(t - s) ds \tag{4}$$

where

- $\hat{t}_i$ is the last spiking time of neuron i
- $\eta(t - \hat{t}_i) = \delta(t - \hat{t}_i) - \eta_0 \exp(-\frac{t - \hat{t}_i}{\tau_{rec}})$, in which
  - $\delta(s)$ is the pulse function
  - $\eta_0 > 0$ is a coefficient
  - τ_{rec} is the refractory period or the recharging period
- w_{ij} is the weight between neuron j and i
- $\varepsilon(s,t) = \frac{1}{C}\int_0^s \exp(-\frac{t'}{\tau_m})\alpha(t - t')dt'$ is the effect of an incoming spike on the neuron membrane, in which
  - τ_m is a membrane potential time constant
  - $\alpha(s) = \frac{q}{\tau_s}\exp(-\frac{s}{\tau_s})\Theta(s)$ is the response function of postsynaptic current, in which, τ_s is a membrane current time constant, and $\Theta(s)$ is a step function
- $\kappa(s,t) = \frac{1}{C}\exp(-\frac{t}{\tau_m})\Theta(s - t)\Theta(t)$ is the effect of an input on the neuron membrane potential

In the real biological nerve system, the electrical properties of neurons are different even though they belong to the same kind of neuron. SRM simulates this dynamic property of neurons by introducing some noises. This paper introduce noises to both the time constant and the threshold, where the time constant are decided in a certain range according to the uniform distribution, and the threshold are decided according to the probability function

$$P = 1 - \exp\{-\exp[\beta(u - \theta)]\} \tag{5}$$

where, β is a firing coefficient, u is the membrane potential, and θ is the threshold.

The response of spiking neurons are shown in Figure 3, where u_1 is the membrane potential of neuron v_1, u_2 is the membrane potential of neuron v_2, v_1 only receive the input, and v_2 receive both the input and the spike emitted by v_1. In Figure 3, real line denotes the membrane potential, dashed line denotes the input, and dash-dotted line denotes the threshold. Once the membrane potential exceeds the threshold, the neuron will emit a spike with 3 as its membrane potential.

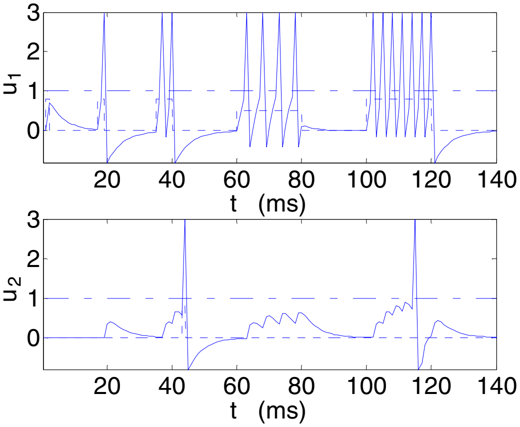


Fig. 3. Response curves of spiking neurons

According to (4), the membrane potential of neurons in the reflex arc should be

(1) If the neurons in SP receive input from outside only, then

$$\begin{aligned} &u_i^{S(k)}(t) = \eta_i^{S(k)}(t - \hat{t}_i^{S(k)}) + \int_0^\infty \kappa_i^{S(k)}(t - t_i^{S(k)}, s) I^{ext(k)}(t - s) ds \\ &k = 0,1,\ldots,n;\ i = 1,2,\ldots,n_k \end{aligned} \tag{6}$$

(2) If the neurons in CPP receive the spike of SP only, then

$$u_i^C(t) = \eta_i^C(t - \hat{t}_i^C) + \sum_{k=0}^{n}\sum_{j=1}^{n_k} w_{ij}^{S(k)} \sum_f \varepsilon_{ij}^C(t - \hat{t}_i^C, t - t_j^{S(k)(f)}) \quad i = 1,2,\cdots,n_C \tag{7}$$

(3) If the neurons in RU receive the spike of CPP only, then

$$u_i^R(t) = \eta_i^R(t - \hat{t}_i^R) + \sum_{j=1}^{n_C} w_{ij}^R \sum_f \varepsilon_{ij}^R(t - \hat{t}_i^R, t - t_j^{C(f)}) \quad i = 1,2,\cdots,n_R \tag{8}$$

4 Reinforcement Learning Mechanism

4.1 Evaluate Agent

Menzel and Giurfa [9] studied the complexity of cognitive functions in the honeybee brain. They found a particularly striking neuron in the bee brain named VUMmx1 (ventral unpaired median neuron of the maxillary neuromere 1), which serves the function of a value system. Referring to the different response to the unconditioned and conditioned stimulus before and after the setup of conditioning, this paper presents an evaluate agent that is shown in Figure 4, where

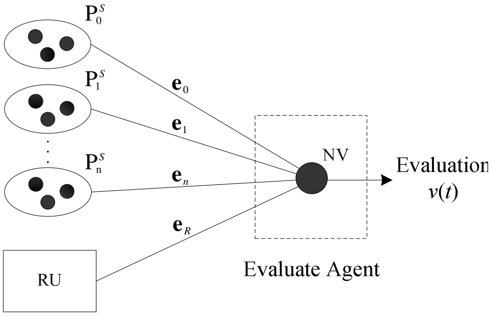


Fig. 4. Evaluate agent

- NV: Evaluate neuron;
- $\mathbf{e}_k = (e_k)_{1\times n_k}$ $(k = 0,1,\cdots,n)$: the linkage weight matrix between P_k^S and NV; e_k vary as the setup and forgetting of conditioning.
- $\mathbf{e}_R$: the linkage weight between P_R and NV. They are used to estimate the setup and forgetting of conditioning.
- NV adopts the Spiking Response Model; its membrane potential is the evaluation value.

Only e_0 and e_1 are used in the experiment, where

$$e_0(t) = \begin{cases} 1 & \text{case 1} \\ a_0 & \text{case 2} \\ e_0(t-1) + c_0(1 - e_0(t-1)) & \text{otherwise} \end{cases} \tag{9}$$

$$e_1(t) = \begin{cases} 0 & \text{case 1} \\ a_1 & \text{case 2} \\ c_1 e_1(t-1) & \text{otherwise} \end{cases} \tag{10}$$

in which,

- $-1 < a_0 < 0$ and $0 < a_1 < 1$ are the value of e_0 and e_1 at the time when conditioning set up respectively;
- c_0 and c_1 are coefficients, and $c_0, c_1 \in [0,1]$;
- Case 1 means that no conditioning has been set up at time *t* or the conditioning acquired has been forgotten at time *t*.

When no conditioning has been set up, the value of e_0 is big and thus NV has a strong response on *US*, e_1 is zero and NV has not response to CS_1. When a conditioning has been set up, i.e., when CS_1 can generate a *CR*, e_0 will be a negative and *US* will inhibit NV, and e_1 will increase so that CS_1 can respond to NV. Once a conditioning been forgotten, e_0 and e_1 will resume their initial states.

4.2 Learning Algorithm

Bi and Poo [10] studied the synaptic modifications in cultured hippocampal neurons, and they found that the synaptic change depended greatly on relative timing of pre- and postsynaptic activity. Referring to their results, a reinforcement learning method based on the Hebbian learning with a decay item is introduced. This algorithm is characterized with the property of 'stimulate-response-reinforcement' and it is detailed as follows:

$$\begin{aligned}\Delta w_{ij}^{S(k)}(t) = -\gamma w_{ij}^{S(k)} + v(t)\int_0^t S_j^{S(k)}(t)W(s)S_i^C(t-s) + S_i^C(t)W(-s)S_j^{S(k)}(t-s)ds \\ i = 1,2\cdots,n_C,\ \ j = 1,2,\cdots,n_k,\ \ k = 1,2,\cdots,n\end{aligned} \tag{11}$$

where

- The first term is decay term that prevent the weights high up to the maximum value, the decay rate $\gamma > 0$;
- The second term is the Hebbian learning term that modifies the weights according to the timing of pre- and postsynaptic spiking, where
  - *v*(t) is the evaluation.
  - $S_j^{S(k)}(t)$ and $S_i^C(t)$ is the spike trains of pre-synaptic neuron j and postsynaptic neuron i respectively.
  - $W(s)$ is the critical window and it is defined as follows:

$$W(s) = \begin{cases} \alpha_+(1 - w_{ij})\exp(s/\tau_+) & \text{s} < 0 \\ \alpha_-(w_{ij} - 1)\exp(s/\tau_-) & \text{s} \geq 0 \end{cases} \tag{12}$$

in which, $\alpha_+ > 0$ and $\alpha_- < 0$ are the amplitudes of the critical window; $\tau_+ > 0$ and $\tau_- > 0$ are the time constants of the critical window; $s = t_j^{(f)} - t_i^{(f)}$ is the interval between the time of pre- and postsynaptic spiking. A curve of the critical window is shown in Figure 5, where $\alpha_+ = 0.2$, $\alpha_+ = 0.2$, $\tau_+ = 5$ and $\tau_- = 10$.

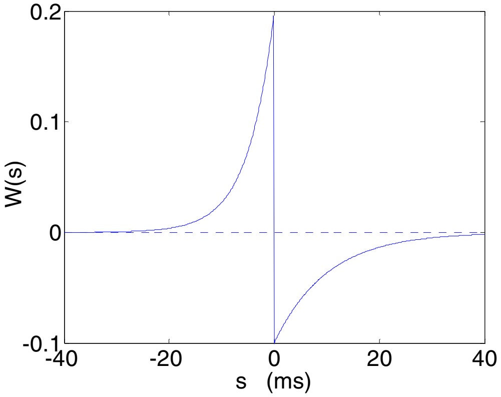


Fig. 5. Critical window

5 Experiment on Inverted Pendulum Control System

The proposed spiking neuron based cognitive model is applied to the inverted pendulum system to learn the balance control technique. The state of the system at time t is specified by four variables, i.e., $(\theta, \dot{\theta}, x, \dot{x})$, where θ is the angle between the pendulum and vertical, $\dot{\theta}$ is the angular velocity of the pendulum, x is the horizontal position and $\dot{x}$ is the velocity of the cart. In this experiment, the inverted pendulum is the environment and the states of the inverted pendulum are outside stimulus, the proposed spiking neuron based cognitive model is expected to set up conditioning and learn the balance control technique of the pendulum in the process of stimulate-response-reinforcement. The experiment system is shown in Figure 6.

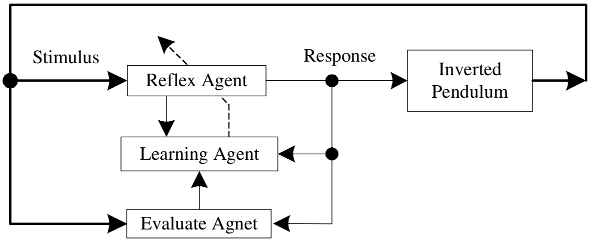


Fig. 6. Structure of the inverted pendulum control experiment system

In the experiment, once θ increase and is larger than ±20°, a failure will occur and a new trial will start. Once θ exceeds ±10°, US, as an input signal to USNG, will produce an unconditioned response to prevent the pendulum from falling down. CSNG in RA consists of four groups and each group receives a state variable as CS.

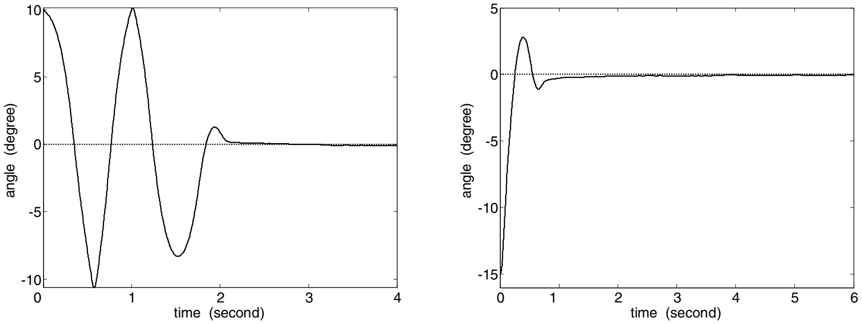


Fig. 7. Results of the inverted pendulum experiment

The experiment results are shown in Figure 7. The left plot of Figure 7 shows the angle curve in the learning process, from which we can learn: the learning process begins with an initial angle of 10°; in the beginning of the experiment, as no conditioning is set up, RA does not respond to CS and it has output only if US presented, i.e., θ exceeds ±10°, thus the initial amplitude is comparatively large; during the establishing process of conditioning , RA gradually learns to make proper responses before the angle exceeds ±10°, i.e., CS can produce conditioned response to balance the pendulum, even without US. The right plot of Figure 7 shows the curve of the control process with initial degree at -15°, it proves that once conditioning been set up by the proposed cognitive model, it can balance the pendulum in a large range of initial angle.

6 Discussion

A spiking-neuron based cognitive model with classical conditioning behaviors is introduced. With a reflex arc structure and the Hebbian learning, the cognitive model possesses the property of 'stimulate-response-reinforcement' and can simulate the learning process of classical conditioning. An experiment on the inverted pendulum validated that this model can learn the balance control strategy by classical conditioning.

In the real-world situation, the model of the environment is always imprecise even unknown. A controller based on the proposed cognitive model of classical conditioning can learn control strategy from the interaction with imprecise even unknown environment without prior knowledge. After conditioning has been set up, the cognitive model based controller can even successfully balance the pendulum in a larger range than it has learnt. This capacity of the spiking-neuron based cognitive model proposed in this paper may suggest its great potential in a lot of applications.

References

1. Rescorla, R.A., Wagner, A.R.: A theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Non-Reinforcement. In: Black, A.H., Prokasy, W.F. (eds.): Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, New York (1972) 64-99

2. Sutton, R.S., Barto, A.G.: Toward a Modern Theory of Adaptive Networks: Expectation and Prediction. Psychological Review 88 (1981) 135-170
3. Sutton, R.S., Barto, A.G.: Time-Derivative Models of Pavlovian Reinforcement. In: Gabriel, M., Moore, J. (Eds.): Learning and Computational Neuronscience: Foundations of Adaptive Networks. MIT Press, Cambridge, MA (1990) 497-537
4. Klopf, A.H.: The Hedonistic Neuron: a Theory of Memory, Learning and Intelligence. Hemisphere, Washington, D.C. (1982)
5. Schmajuk, N.A., DiCarlo, J.J.: Stimulus Configuration, Classical Conditioning, and Hippocampal Function. Psychological Review 99 (1992) 268-305
6. Balkenius, C.: Natural Intelligence in Artificial Creatures. Lund University Cognitive Studies 37. Lund University Cognitive Science, Lund, Sweden (1995)
7. Johansson, C., Lansner, A.: An Associative Neural Network Model of Classical Conditioning. TRITA-NA-P0217. Stockholm, SANS (2002)
8. Gerstner, W., van Hemmen, J.L.: Associative Memory in a Network of ‘Spiking’ Neurons. Network 3 (1992) 139-164
9. Randolf, M., Martin, G.: Cognitive Architecture of a Mini-Brain: the Honeybee. Trends in Cognitive Sciences 2 (2001) 62-71
10. Bi, G., Poo, M.: Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type. J. Neurosciences 18 (1998) 10464-10472

Probabilistic Tangent Subspace Method for M-QAM Signal Equalization in Time-Varying Multipath Channels

Jing Yang, Yunpeng Xu, and Hongxing Zou

Department of Automation, Tsinghua University,
Beijing, 100084, P. R. China
{yang-jing03, xuyp03}@mails.tsinghua.edu.cn, hongxing_zou@tsinghua.edu.cn

Abstract. A new machine learning method called probabilistic tangent subspace is introduced to improve the performance of the equalization for the M-QAM modulation signals in wireless communication systems. Due to the mobility of communicator, wireless communication channels are time variant. The uncertainties in the time-varying channel's coefficients cause the amplitude distortion as well as the phase distortion of the M-QAM modulation signals. On the other hand, the Probabilistic Tangent Subspace method is designed to encode the pattern variations. Therefore, we are motivated to adopt this method to develop a classifier as an equalizer for time-varying channels. Simulation results show that this equalizer performs better than those based on nearest neighbor method and support vector machine method for Rayleigh fading channels.

1 Introduction

For wireless mobile communication, the channels are time-varying because of the mobility of communicator. For such channels, amplitude and phase distortions typically occur simultaneously. For a transmitted sequence of symbols, such distortions manifest as dispersions so that any one symbol in the received demodulated sequence is not well defined. Otherwise, the overlap or smearing, known as *intersymbol interference* (ISI) in multipath channels, and additive noises arise in most modulation systems. Therefore, equalization is required.

Equalization is partitioned into two categories: *maximum likelihood sequence estimation* (MLSE) [2] and *symbol detection*. MLSE method is an optimum solution for equalizing an information sequence corrupted by ISI and additive Gaussian noise. Its optimality is based on the assumption that the channel is known. However, in the actual applications, the assumption can not be hold. Thus, channel estimation is needed, and such channel estimation process causes the computation complexity very high. To solve this problem, symbol detection method is considered. Symbol detection method treat the equalization process as a classification problem, in which channel estimation is not needed. This method is a mapping of received signals on an expected label set of symbols. Efforts on symbol detection method have been made in recent years. For example, In [3]-[5], several receiver structures based on neural networks are proposed; in [6], a

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 949–957, 2005.

nearest neighbor rule is used; and a nonlinear equalization scheme using support vector machine technology is described in [7].

In this paper, We consider the symbol detection approach for *Quadrature Amplitude Modulation* (QAM) signals [9], which are most suitable when bandwidth efficiency and high data rates are desired. QAM scheme is most used in many communication systems. For example, CDMA 2000 uses 16-QAM and the highest speed wireless lan standard (802.11a) uses 64-QAM. M-QAM signal equalization can be viewed as a supervised M-class classification problem. First, we use some prior symbols with labels to train the classifier, then, we can use the classifier as an equalizer in the receiver.

For time-varying channels, neural network methods often require long convergence time and a large number of neurons, which make the implementation expensive. While in pattern recognition field, Nearest Neighbor (NN) and Support Vector Machine (SVM) methods regard the patterns as invariant and tolerate only small variations of input patterns while keeping the class label unchanged. However, it is more desirable to take into account the effects of variations of the channels in the classification methods. This calls for the adoption of invariant classification techniques.

There has been much research addressing to this problem in machine learning and pattern recognition fields. These methods consider the pattern variations and can give the invariant classification results while the variations are large. We interpret the channel's time variations as the uncertainties in its coefficients, which correspond to the uncertainties of pattern variations. This interpretation motivates us to use these invariant algorithms as an equalizer for time-varying channels.

There have been some invariant algorithms for pattern variations, such as Tangent Distance (TD) [10] [11] and Virtual-Support Vector Machine (V-SVM). But these two methods need the prior knowledge of the variations, which is difficult for time-varying channel because of its uncertainties of the channel confidences. To cope with these problems, in this paper we introduce Probabilistic Tangent Subspace (PTS) method [1] to the equalization problem for the M-QAM signals in time-varying multipath channel. PTS method is a novel and practical way which can encode the variations and need no prior knowledge of the channel. The simulation results demonstrate that this method can well solve the problem of the amplitude and phase distortions, which are caused by the time variations of the channel.

The rest of the paper is organized as follows. In Section 2, we introduce the pattern recognition concept and the model of the equalization. Section 3 gives the details about the PTS algorithm. Simulation results for M-QAM modulation signals equalization in the time-varying multipath channels will be given in Section 4. Finally, we conclude in Section 5.

2 Equalization Viewed as Pattern Recognition

2.1 Notations

For convenience, notations which will be used in the following sections are shown in Table 1.

Table 1. Notations

$s(n)$:	The nth transmitted symbol after M-QAM modulation
$x(n)$:	The nth output symbol of the time-varying multipath channels
$y(n)$:	The nth received symbol with white Gaussian noise
$e(n)$:	The nth additive white Gaussian noise with $\mathrm{E}\left[\boldsymbol{e}^{\mathrm{H}}\boldsymbol{e}\right] = \sigma^2\boldsymbol{I}$
L:	The number of time-varying multipath channel's taps
h_l:	The coefficient of time-varying multipath channel's $(l+1)$th tap
$\boldsymbol{y}_{\mathcal{F}}$:	The feature vector as the input of the PTS-I equalizer

2.2 QAM Modulation Signals

Quadrature Amplitude Modulation (QAM) [9] is a modulation scheme that is most suitable when bandwidth efficiency and high data rates are desired. It is a combination of Amplitude Shift Keying (ASK) and Phase Shift Keying (PSK). Each symbol is mapped to a certain phase and certain amplitude to be transmitted. A common way of representing a QAM signal is through the use of a graphical tool called a constellation. A QAM constellation consists of a set of axis and points, or phasors, which represent each possible symbol. The angle the phasor makes with the horizontal axis is the phase of the transmitted symbol. The magnitude of the phasor is the amplitude of the transmitted symbol.

A square M-QAM signal can be represented in the following form

$$s_m(t) = I_m \cos(t) - Q_m \sin(t) \tag{1}$$

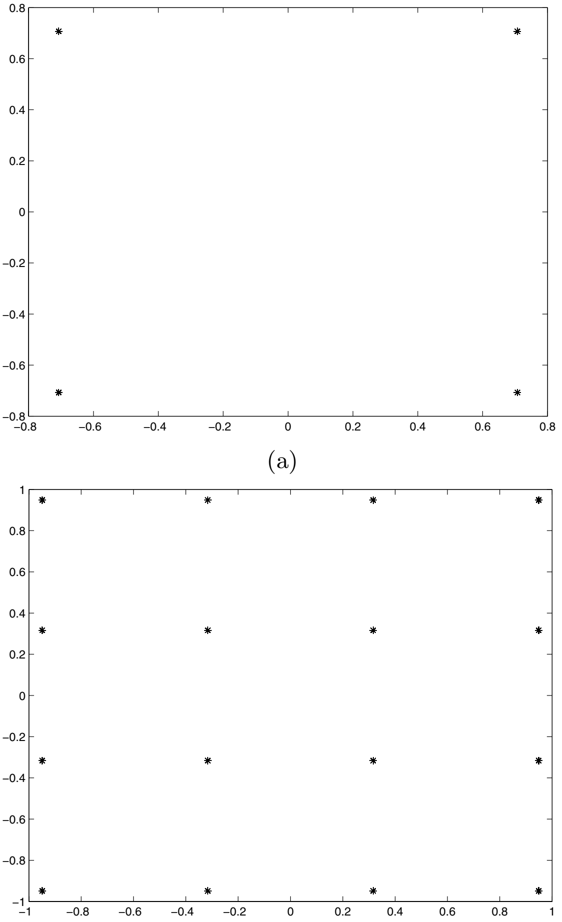
where $m = 1, \ldots, M$, and $0 \leq t \leq T$.

T is the time duration of a symbol, M is the number of used phase. I_m and Q_m are the levels outputted by the level generator to the in-phase and quadrature components of the modulator, respectively.

In order to be a square constellation these levels must follow a certain formula. For M-QAM, The values of I_m and Q_m are chosen from a $\sqrt{M} \times \sqrt{M}$ matrix in which each matrix element (I_m, Q_m) maps directly to a point on the constellation. When we consider two special cases: $M = 4$ and $M = 16$. The 4-QAM and 16-QAM square constellations are shown in Fig. 1. Therefore, the M-QAM signals equalization can be viewed as an M-class classification problem.

2.3 Model of the Equalization

M-QAM signal equalization can be regarded as an M-class classification problem. Based on this point, the received symbols are mapped onto a feature space, i.e., they are grouped as a feather vector $\boldsymbol{y}_{\mathcal{F}}$. The input of the classifier is the feature vector and the output should match as better as the original signal entering the channel. This method only attempts to map the output onto a finite set, which equals to the label set in the classification problem. Therefore, it can overcome the amplitude and phase distortions and the intersymbol interference (ISI).



(a)

(b)

Fig. 1. M-QAM modulation signal constellations. (a). 4-QAM, (b). 16-QAM.

The Model which is subject to intersymbol interference and additive Gaussian noise is illustrated in Fig. 2, where the number of time-varying channels taps is L and the multipath channel is $\mathcal{H}_L = [h_1, \cdots, h_L]$. The nth received symbol $y(n)$ can be represented as

$$y(n) = \sum_{l=1}^{L} h_l(n)s(n-L) + e(n) \tag{2}$$

Here, we define the dimension of the feather vector according to the expected amount of intersymbol interference (ISI), which is the number of channels taps L. In our scheme, we concatenate the real parts and the imaginary parts of L adjacent symbols as the feature vector. Then, the input of the equalizer $\boldsymbol{y}_{\mathcal{F}} \in \mathbb{R}^{2L}$ can be defined as

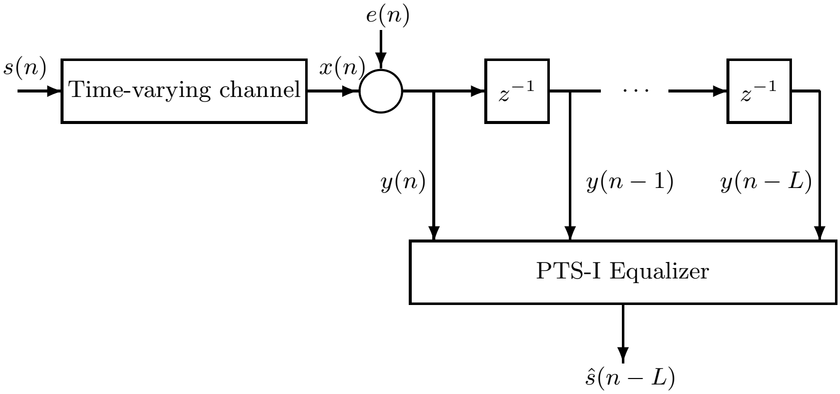


Fig. 2. Model of the equalization for time-varying multipath channel

$$\boldsymbol{y}_{\mathcal{F}} = [\mathcal{R}\{y(n)\}, \mathcal{I}\{y(n)\}, \mathcal{R}\{y(n-1)\}, \mathcal{I}\{y(n)\}, \cdots, \mathcal{R}\{y(n-L)\}, \mathcal{I}\{y(n-L)\}]$$

3 Probabilistic Tangent Subspace Method

Probabilistic Tangent Subspace (PTS) [1] is based on Tangent Distance algorithm. Its basic assumption is that tangent vectors can be approximately represented by the pattern variations. In [1], three subspace models are proposed, including the linear subspace, nonlinear subspace, and manifold subspace models. In this paper, we apply the linear subspace method called PTS-I to design the equalizer.

3.1 Linear Subspace: PTS-I

Let the training sequence $\boldsymbol{y}_T = \{y_i\}_{i=1}^{m} \subset \mathbb{R}^{2L}$. First we form the tangent vector set S according to

$$S = \{z | z = y - y_r, \quad \text{if} \quad c(y) = c(y_r) \quad \text{and} \quad y \in \in \mathcal{L}(y_r)\} \tag{3}$$

where $c(y)$ denotes the class label of sample y. $\mathcal{N}(y_r)$ indicates the neighbor set of prototype y_r.

While S is characterized by a linear subspace, one Gaussian density function in this space can be estimated as

$$G(z|S) = \frac{1}{(2\pi)^L |\Sigma|^{1/2} exp\{-\frac{1}{2}(z-\mu)^{\mathrm{T}} \Sigma^{-1} (z-\mu)\}} \tag{4}$$

where μ is the mean vector, and Σ is the covariance matrix.

The exponent term of equation (4) is a Mahalanobis distance

$$d(z) = (z-\mu)^{\mathrm{T}} \Sigma^{-1} (z-\mu) = \boldsymbol{u}^{\mathrm{T}} \Lambda^{-1} \boldsymbol{u} = \sum_{i=1}^{2L} \frac{\boldsymbol{u}_i^2}{\lambda_i} \tag{5}$$

where $\Lambda = \mathrm{diag}\{\lambda_1, \cdots, \lambda_{2L}\}$.

PTS-I represents S as a linear space. The principal subspace of S is spanned by the first p components, which is principal component analysis (PCA) on S. Then, the Mahalanobis distance $d(z)$ can be approximated by

$$\hat{d}(z) = \sum_{i=1}^{p} \frac{\boldsymbol{u}_i^2}{\lambda_i} + \sum_{i=p+1}^{2L} \frac{\boldsymbol{u}_i^2}{\lambda_i} = \sum_{i=1}^{p} \frac{\boldsymbol{u}_i^2}{\lambda_i} + \frac{1}{\rho} \sum_{i=p+1}^{2L} \boldsymbol{u}_i^2 = \sum_{i=1}^{p} \frac{\boldsymbol{u}_i^2}{\lambda_i} + \frac{1}{\rho} \varepsilon^2(z) \tag{6}$$

where ρ is a weight coefficient defined by

$$\rho = \frac{1}{2L-p} \sum_{i=p+1}^{2L} \lambda_i \tag{7}$$

Finally, The estimated independent Gaussian density function in S can be wrote as

$$\hat{G}(z|S) = \left[\frac{\exp\left\{ -\frac{1}{2} \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i} \right\}}{(2\pi)^{p/2} \prod_{i=1}^{p} \lambda_i^{1/2}} \right] \left[\frac{\exp\left\{ -\frac{\varepsilon^2(z)}{2\rho} \right\}}{(2\pi\rho)^{(2L-p)/2}} \right] \tag{8}$$

Through the above analysis, the estimated Gaussian density $\hat{G}(z|S)$ can be computed only from the principal tangent subspace. Therefore, while we use the information from the principle tangent subspace information, the full tangent space's information is utilized.

It is assumed that the tangent subspace is a uniform subspace. Such subspace is invariant while the class label varies. Therefore, except the estimated Mahalanobis distance $\hat{d}(z)$, all the other parameters in $\hat{G}(z|S)$ are constant after training.

While classifying a sample y_t, we project the linear variation $z_r = y_t - y_r$ into the principal subspace. If the Mahalanobis distance $\hat{d}(z_r)$ is the shortest, the class label of y_t is the same as that of y_r.

3.2 The Algorithm of the PTS-I

The algorithm of the PTS-I [1] which is used to design the equalizer is presented in Table 2.

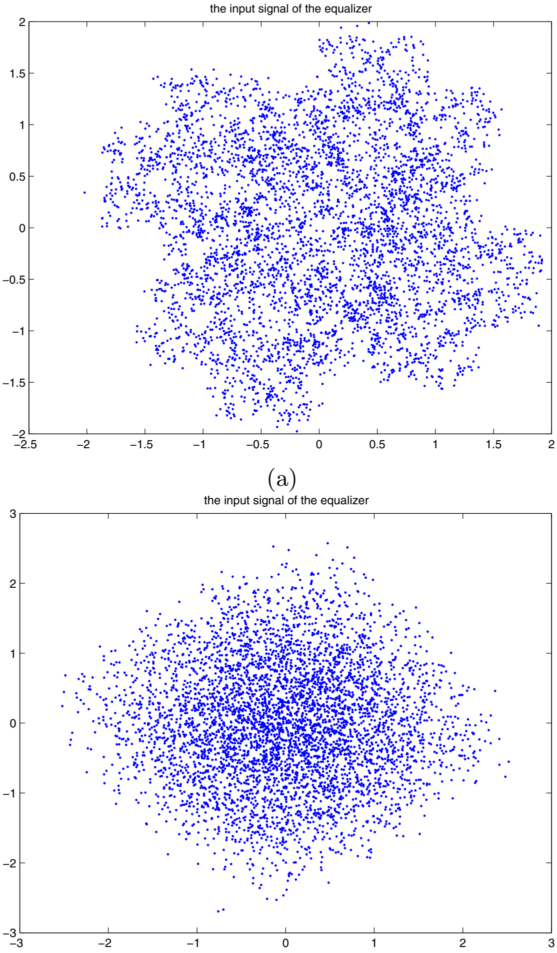
Table 2. Classification algorithm for PTS-I

```
Training:
    Step 1: Obtain $\Sigma$ by the tangent vector set;
    Step 2: Perform eigen-decomposition (PCA).
    Step 3: Estimate the weight coefficient $\rho$.
Classification (Equalization):
    Step4 : Project $z_r = y_t - y_r$;
    Step 5: Compute the error $\varepsilon^2(\cdot)$;
    Step 6: Compute the approximate Mahalanobis distance $\hat{d}(z_r)$;
    Step 7: Repeat Steps 4-6 for each sample $y_t$;
    Step 8: Return the label of $y_t$.
```

4 Simulation Results

In this section, to evaluate the performance of the PTS-equalizer, we conducted simulations on 4-QAM and 16-QAM modulation signals in Rayleigh fading channel. We set the number of taps of channel $L = 4$. The average power ratio of the first channel tap to the fourth tap is -5 dB, i.e., the power ratio is $[0, -5, -10, -15]$ dB. The normalized Doppler frequency spread is set to be 3.4×10^{-3} Hz. Signal-to-noise ratio (SNR) is 20 dB.

After being transmitted in the channels, the received symbols scatter plots are shown in Fig. 3. It is clear from the figure that the received symbols are severely corrupted by the amplitude and phase distortions and ISI.



(a)

(b)

Fig. 3. The received symbol scatter plots. (a). 4-QAM, (b). 16-QAM.

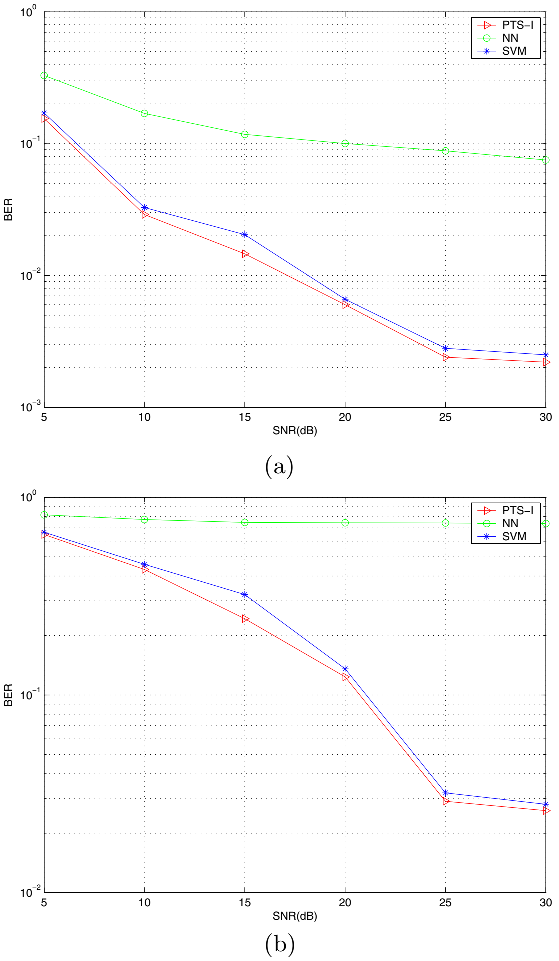


Fig. 4. The comparison of BERs for PTS-I equalizer, NN equalizer and SVM equalizer at different SNR levels. (a). 4-QAM, (b). 16-QAM.

In our simulation, 1000 symbols are used for training. After training, 5000 symbols are equalized using different equalizers, including the PTS-I equalizer, the NN equalizer and the SVM equalizer. Fig. 4 shows the results of Bit Error Rate (BER), which are averaged over 50 runs. By comparing the BERs of different equalizers, it is obvious that the PTS-I equalizer outperforms the SVM equalizer and the NN equalizer at each SNR level, although sometimes the improvements over the SVM equalizer are not significant. The superiority of the PTS-I equalizer is resulted from that PTS-I method can partly encode the variations of the time-varying multipath channel, but NN and SVM can not.

5 Conclusions

In this paper, we apply PTS-I classifier as an equalizer to time-varying multipath channel. The uncertainties of the time-varying multipath channel's coefficients are interpreted as the uncertainties of pattern variations. Therefore, in this scheme, the variations of the time-varying multipath channel can be partly encoded by the PTS-I method. We implement simulations on 4-QAM and 16-QAM signals in Rayleigh fading channel. Simulation results demonstrate that this scheme can provide satisfactory performance and is superior to the NN equalizer and SVM equalizer.

References

1. Lee, J. G., Wang, J. D., Zhang, C. S., Bian, Z. Q.: Probabilistic Tangent Subspace: a Unified View. In: Proc. 21st Intl. Confs. on Machine Learing (ICML 2004) Banff. Alberta. Canada. (2004)
2. Benelli, G., Gastellini, G., Re, E. D., Fantacci, R., Pierucci, L., Pagliani, L.: Design of a Digital MLSE Receiver for Mobile Radio Communications. IEEE Proc. Globecom (1991) 1469–1473
3. Veciana, G. D., Zakhor, A.: Neural Netobased Continous Phase Modulation Receivers. IEEE Trans. Communication 40 (1992) 1392–1408
4. Kechriotix, G., Zervas, E., Manolakos, E. S.: Using Recurrent Neural Network for Adaptive Communication Channel Equalization. IEEE Trans. Neural Networks 5 (1994) 267–278
5. Parisi, R., Di Claudio, E. D., Orlandi, G., Rao, B. D.: Fast Adaptive Digital Equalization by Recurrent Neural Network. IEEE Trans. Signal Processing 45 (1997) 2731–2739
6. Savazzi, P., Favalli, L., Costamagna, E., Mecocci, A.: A Suboptimal Approach to Channel Equalization Based on the Nearest Neighbor Rule. IEEE Journal on Selected Areas in Communications 16 (1998) 1640–1648
7. Sebald, D. J., Bucklew, J. A.: Support Vector Machine Techniques for Nonlinear Equalization. IEEE Trans. Signal Processing 48 (2000) 3217–3226
8. Liang, Q. L., Mendel, J. M.: Equalization of Nonlinear Time-varying Channels Using Type-2 Fuzzy Adaptive Filters. IEEE Trans. Fuzzy Systems 8 (2000) 551–563
9. Proakis, John G.: Digital Comunications. New York. McGraw-Hill. 3rd ed (1995)
10. Hastie, T., Simard, P., Saeckinger, E.: Learning prototype models for Tangent Distance. Advances in Neural Information Processing Systems 7 (NIPS 7)
11. Simard, P., LeCun, Y., Denker, J., Victorri, B.: Transformation Invariance in Pattern Recognition - Tangent Distance and Tangent Propagation. Inernational Journal of Imaging System and Technology 11 (2001) 181–194

Face Recognition Based on Generalized Canonical Correlation Analysis

Quan-Sen Sun[1,2], Pheng-Ann Heng[3], Zhong Jin[4,2], and De-Shen Xia[2]

[1] School of Science, Jinan University, Jinan 250022, China
qssun@126.com
[2] Department of Computer Science, Nanjing University of Science & Technology,
Nanjing 210094, China
deshen_x@263.net
[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong,
Hong Kong
pheng@cse.cuhk.edu.hk
[4] Centre de Visió per Computador, Universitat Autònoma de Barcelona, Spain
zhong.jin@cvc.uab.es

Abstract. We have proposed a new feature extraction method and a new feature fusion strategy based on generalized canonical correlation analysis (GCCA). The proposed method and strategy have been applied to facial feature extraction and recognition. Compared with the face feature extracted by canonical correlation analysis (CCA), as in a process of GCCA, it contains the class information of the training samples, thus, aiming for pattern classification it would improve the classification capability. Experimental results on ORL and Yale face image database have shown that the classification results based on GCCA method are superior to those based on CCA method. Moreover, those two methods are both better than the classical Eigenfaces or Fishierfaces method. In addition, the newly proposed feature fusion strategy is not only helpful for improving the recognition rate, but also useful for enriching the existing combination feature extraction methods.

1 Introduction

Face recognition always has an important status in the area of pattern recognition. As face images are influenced by illumination, expression, pose… *etc*, the recognition technique for them has long become a challenge in pattern recognition. Among face recognition methods, there are two most famous techniques: Eigenfaces and Fisherfaces. Eigenfaces is based on the technique of Principal Component Analysis (PCA). It was originally proposed by Kirbyand Sirovich[1] and popularized by Turk and Pentland[2,3]. Fisherfaces is based on Fisher Linear Discriminant Analysis (LDA) and was put forward by Swets[4] and Belhumeur[5], respectively. Now linear feature extraction based methods have been widely used in face recognition[6-8]. Recently, the methods based on kernel PCA(KPCA)[9] and kernel LDA(KLDA)[10] have extended the problem of face recognition into nonlinear methods, this promotes the research on face recognition.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 958 – 967, 2005.

Most of the researches above are all based on single feature, and their recognition results usually have some limitations in certain way. In Ref.[11,12], based on the idea of feature fusion, by combining two groups of features into one, the authors proposed a theory of complex linear projective discriminant analysis, with an application to face recognition. This algorithm has achieved a good effect, which is superior to the method with single feature. According to the same idea, a framework of canonical correlation analysis (CCA)[13,14], which is applied in image recognition, was built [15], and it has been used in the region of image recognition that enriches the methods of feature fusion. In light of this, from the angle of favoring pattern classification, the theory of CCA was extended and a new method — generalized CCA(GCCA) was established[16], the method has acquired better classification effects than that of CCA. This paper further improves the theory of GCCA, and proposes a new feature fusion strategy. The advantages of this proposed method have been validated on the ORL and Yale face image database.

The rest of this paper is organized as follows. In Section 2 the theory and method of GCCA are presented. A new feature fusion strategy and the classification methods are proposed in Section 3 and Section 4, respectively. In Section 5, we give experiment results on the ORL and Yeal face image database. Finally, conclusions are drawn in Section 6.

2 Generalized Canonical Correlation Analysis (GCCA)

Let A and B be two groups of feature sets on a pattern sample space Ω, any pattern sample $\zeta \in \Omega \subset \mathrm{R}^n$, of which the two corresponding feature vectors are $x \in A \subset \mathrm{R}^p$ and $y \in B \subset \mathrm{R}^q$, respectively. Let S_{Wx} and S_{Wy} denote the within-class scatter matrices of training sample space A and B, respectively, i.e.

$$S_{Wx} = \sum_{i=1}^{c} P(\omega_i)[\sum_{j=1}^{l_i} \frac{1}{l_i}(x_{ij} - m_i^x)(x_{ij} - m_i^x)^{\mathrm{T}}],$$

$$S_{Wy} = \sum_{i=1}^{c} P(\omega_i)[\sum_{j=1}^{l_i} \frac{1}{l_i}(y_{ij} - m_i^y)(y_{ij} - m_i^y)^{\mathrm{T}}],$$

where $x_{ij} \in A$ and $y_{ij} \in B$ denote the j^{th} training sample in class i ; $P(\omega_i)$ is the prior probability of class i; l_i is the number of training samples in class i; m_i^x and m_i^y are the mean vectors of training samples in class i , respectively. Let L_{xy} denote between-set covariance matrix of A and B, and $L_{xy}^{\mathrm{T}} = L_{yx}$ and $r = rank(L_{xy})$. i.e.

$$L_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - m^x)(y_i - m^y)^{\mathrm{T}}$$

where $x_i \in A$ and $y_i \in B$; m^x and m^y denote the mean vectors of training sample spaces A and B , respectively. Assuming that S_{Wx} and S_{Wy} are both positive matrices, we build a criterion function as follow.

$$J_g(\xi,\eta) = \frac{\xi^{\mathrm{T}} L_{xy} \eta}{(\xi^{\mathrm{T}} S_{Wx} \xi \cdot \eta^{\mathrm{T}} S_{Wy} \eta)^{1/2}} \tag{1}$$

Criterion (1) is called generalized canonical correlation discriminant criterion. The pair vectors that maximize criterion function $J_g(\xi,\eta)$ are regarded as the projective directions, of which the physical meaning are that two sets of the feature vectors have the maximum correlation when the projected samples minimize the within-class scatter.

The focus of GCCA is to seek a pair of unitary projective directions ξ_k and η_k such that

$\{\xi_k, \eta_k\} = \arg \max\limits_{|\xi|=|\eta|=1} J_g(\xi,\eta)$ for $k = 1,2,\cdots,r$

is subjected to the following constraints

$$\xi_k^{\mathrm{T}} S_{Wx} \xi_i = \eta_k^{\mathrm{T}} S_W \eta_i = 0 \tag{2}$$

for all $1 \le i < k$.

Through the method above, we can obtain two groups of projective vectors $\{\xi_i\}$ and $\{\eta_i\}$ called generalized canonical projective vectors (GCPV)[16]. The feature extracted by this means is called generalized canonical correlation discriminant feature(GCCDF).

Here, we will discuss the solution of GCPV.

In order to ensure the solution be exclusive, we first assume

$$\xi^{\mathrm{T}} S_{Wx} \xi = \eta^{\mathrm{T}} S_{Wy} \eta = 1 \tag{3}$$

So the problem will be transformed to the finding of a pair of GCPV ξ and η that maximize criterion function (1) under the constraint (3). Via Lagrange multiplier method, the problem will be further transformed to the solving of two generalized feature equations as follows:

$$L_{xy} S_{Wy}^{-1} L_{yx} \xi = \lambda^2 S_{Wx} \xi \tag{4}$$

$$L_{yx} S_{Wx}^{-1} L_{xy} \eta = \lambda^2 S_{Wy} \eta \tag{5}$$

So we can obtain the following Theorem 1[16]:

Theorem 1. Based on the generalized canonical correlation criterion function (1), GCPV can be *d* pair eigenvectors corresponding to the first *d* maximum nonzero eigenvalues ($\lambda_1^2 > \lambda_2^2 > \cdots > \lambda_d^2$) of the two generalized eigenequations (4) and (5), and these GCPV must satisfy:

$$\begin{cases} \xi_i^{\mathrm{T}} S_{Wx} \xi_j = \eta_i^{\mathrm{T}} S_{Wy} \eta_j = \delta_{ij} \\ \xi_i^{\mathrm{T}} L_{xy} \eta_j = \lambda_i \delta_{ij} \end{cases} \quad (i, j = 1,2,\cdots,d) \tag{6}$$

where $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

Following the Proof in Ref. [16], We Can Obtain the Following Corollary:

Corollary. If all eigenvalues of generalized eigenequations (4) and (5) satisfy $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2 > \lambda_{r+1}^2 = \cdots = \lambda_p^2 = 0$ ($p \leq q$), the generalized canonical correlation criterion function is $J_g(\xi_i, \eta_i) = \lambda_i$ $(i = 1,2,\cdots,p)$.

Theorem 2. Under the generalized canonical correlation criterion function (1), the number of the GCPV is more than *r* pairs, where $r = rank(S_{xy})$. These $d(\leq r)$ pair GCPV can be composed of the eigenvectors which satisfy (6) and correspond to the first *d* maximum eigenvalues of the two generalized eigenequations (4) and (5).

Proof: From the Corollary, $J_g(\xi_i, \eta_i) = \lambda_i$ $(i = 1,2,\cdots,p)$. So when $i = r+1,\cdots,p$, $J_g(\xi_i, \eta_i) = 0$, and no effective generalized canonical correlation discriminant feature can be extracted, thus the number of effective GCPV is no more than *r* pairs. From Theorem 1, the $d(\leq r)$ pair GCPV can be composed of the eigenvectors that satisfy constraint (2) and correspond to the first *d* maximum eigenvalues of the generalized eigenequations (4) and (5). □

3 A New Feature Fusion Strategy

We have proposed two feature fusion strategies (FFS1 and FFS2) in Ref. [15]. In this Section, we will present another new feature fusion strategy called FFS3. In Section 2, we have acquired GCPV $\{\xi_i\}$ and $\{\eta_i\}$, and built the projective matrixes:

$$W_x = (\xi_1, \xi_2, \cdots, \xi_d)\,;\ W_y = (\eta_1, \eta_2, \cdots, \eta_d)\,.$$

Therefore, for all $x \in A$ and $y \in B$, then the generalized canonical correlation discriminant features can be extracted as follows:

$$z_1 = W_x^{\mathrm{T}} x,\ z_2 = W_y^{\mathrm{T}} y$$

The canonical correlation discriminant features constitute the following matrix:

$$M = [z_1,\ z_2] = [W_x^{\mathrm{T}} x,\ W_y^{\mathrm{T}} y] \in \mathrm{R}^{d \times 2}$$

where matrix *M* is called the correlation feature matrix of feature vectors *x* and *y* (or pattern sample ζ).

4 Classification Based on Correlation Feature Matrix

Let $\omega_1, \omega_2, \cdots, \omega_c$ be the c known pattern classes and the number of training samples be $n = n_1 + \cdots + n_c$, where n_i represents the sample number of class *i*. From the analysis above, we can know that every pattern sample corresponds to a *d*×2 correlation feature matrix. The distance between any two correlation feature matrices $M_i = [z_1^{(i)},\ z_2^{(i)}]$ □and $M_j = [z_1^{(j)},\ z_2^{(j)}]$ is defined as

$$d(M_i, M_j) = \sum_{k=1}^{2} \| z_k^{(i)} - z_k^{(j)} \|_2 \ ,$$

where $\| \cdot \|_2$ represents the vector's Euclidean distance.

4.1 The Minimum-Distance Classifier

Assume that the mean matrix of the correlation feature matrixes of class *i*'s training samples is

$$m^{(i)} = \frac{1}{n_i} \sum_{k=1}^{n_i} M_k^{(i)} = [m_1^{(i)}, m_2^{(i)}],\ i = 1, 2, \cdots, c\ .$$

Given any testing sample ζ , its correlation feature matrices is $M = [z_1,\ z_2]$. If $d(M, m^{(l)}) = \min_j\ d(M, m^{(j)})$, then $M \in \omega_l$ ($\zeta \in \omega_l$).

4.2 The Nearest-Neighbor Classifier

Assume that $\zeta_1, \zeta_2, \cdots, \zeta_n$ are the all training samples and their corresponding correlation feature matrixes are $M_1, M_2, \cdots, M_n$. For any testing sample ζ , its correlation feature matrix is $M = [z_1, z_2]$.
If $d(M, M_l) = \min_j\ d(M, M_j)$ and $M_l \in \omega_k$ ($\zeta_l \in \omega_k$), then $M \in \omega_k$ ($\zeta \in \omega_k$).

5 Experiments

5.1 Pretreatment of Face Image

As mentioned above, the within-class scatter matrices of training sample space *A* and *B* must be positive in the GCCA algorithm. Face recognition are typical small sample size and high-dimensional problems. Because the dimension of the image vector to be recognized is high, it is very difficult or even impossible to find enough training sample so as to ensure the reversibility of the within-class scatter matrices. In Ref. [15], we have proposed a method which can deal with small sample size and high-dimensional problems, which is PCA plus CCA. In this strategy, PCA is used first for dimensional reduction, and then CCA is used for feature extraction in the PCA transformed space. Since the dimension of the PCA transformed space is usually much lower than that of the original feature space, the difficulty resulting from the singularity of the covariance matrix is avoided. In the same way, the above strategy also can be applied to GCCA. After two groups of feature sets on face images are extracted, firstly Using K-L transform(PCA) to reduce the dimension of two groups of feature sets, and assuring two within-class scatter matrices must be nonsingular; then use the proposed GCCA algorithm to extract the GCCDF in the transformed low-dimensional feature spaces.

5.2 Experiment on ORL Face Image Database

Our proposed method is tested on the ORL face image database (http://www.cam-orl.co.uk), which contains images of 40 individuals, where each person has 10 different images. For some people, images were taken at different times, so the facial expression (open/closed eyes, smiling/ non-smiling) and facial details (glasses/no glasses) are different. The images were taken at a dark homogeneous background and the people are in upright, frontal position with a tolerance for some tilting and rotation of up to 20º. Moreover, there is some variation in a scale of up to about 10%. All images are grayscale and normalized with a resolution of 112×92. Some images in ORL are shown in Fig.1.

In this experiment, we use the first five images of each person for training and the remaining five for testing. Thus, the total amount of training samples and testing samples are both 200.

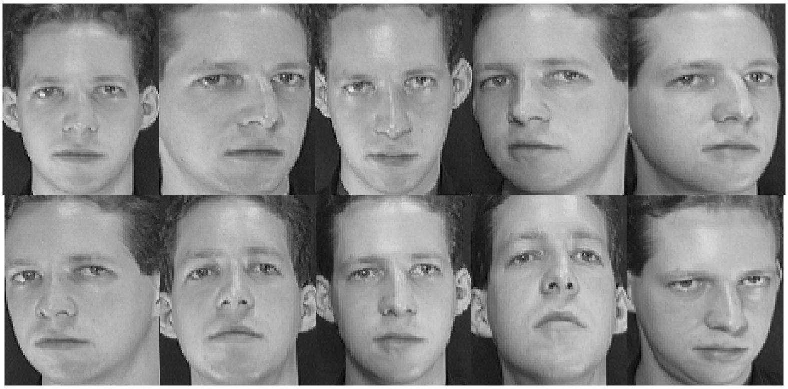


Fig. 1. Ten images of one person in ORL face database

First, we apply the K-L transformation, 112×92=10304-dimensional vectors of original image will be reduced into 42-dimensional feature vectors which make up of the first class training sample space $A=\{x \mid x\in \mathrm{R}^{42}\}$; Then, we perform a quartic wavelet transformation on the original images using Daubechies orthonormal wavelet in order to obtain the low-frequency images with scale of $7\times 6=42$, thus the second class training sample space $B=\{y \mid y\in \mathrm{R}^{42}\}$ can be constructed; Finally, we combine those two groups of features above, and according to the GCCA proposed in this paper, solve the GCPV. By applying the new FFS3 in extracting the canonical correlation feature matrices of face images, we proceed to classify by the Nearest-Neighbor Classifier based on the correlation feature matrix proposed in Section 4. The classification results are shown in Table 1.

In order to compare our method with that proposed in Ref. [15] and [16], in Table 1 we also give the results based on CCA and GCCA under the three kinds of feature fusion strategies. In addition, we also provide the classification results by Eigenfaces (PCA) and Fisherfaces (FLDA) methods[2, 5] based on the single feature on original face images.

Table 1. Classification error numbers and optimal recognition rates based on CCA, GCCA under the three kinds of feature fusion strategies, PCA and FLDA in the Nearest-Neighbor Classifier

Number of axes (d)	CCA			GCCA			PCA	F LDA
	FFS1	FFS2	FFS3	FFS1	FFS2	FFS3		
16	22	23	22	14	14	14	33	39
18	17	17	16	14	17	13	30	38
20	13	14	12	12	15	11	28	36
22	14	14	13	12	16	10	26	34
24	14	14	14	10	13	8	25	28
26	13	13	13	9	12	8	26	31
28	13	13	13	9	13	9	25	32
30	13	13	13	5	10	7	25	32
32	13	13	12	9	8	8	23	28
34	12	12	13	10	10	6	22	28
36	13	13	13	9	9	9	22	28
38	13	13	14	6	10	7	22	25
40	13	13	15	9	10	8	22	
42	14	14	14	8	11	8	22	
Optimal results	94.0%	94.0%	94.0%	97.5%	96.0%	97.0%	91.0%	88.5%

From Table 2, we know that under the three kinds of feature fusion strategies, GCCA's classification results are all superior to CCA's while GCCA's and CCA's classification results are all superior to PCA's or FLDA's. In addition, Table 1 also shows that the classification results under the three kinds of feature fusion strategies are almost equivalent. CCA's optimal recognition rates are all 94% under the three kinds of feature fusion strategies. GCCA's optimal recognition rates are 97.5%, 96% and 97%, respectively. However, if we consider the time of classification, GCCA, for example, spends 25.750s, 25.125s and 27.688s respectively under the three kinds of feature fusion strategies (The time of wavelet transformation not included).

5.3 Experiment on Yale Face Image Database

The Yale face database is adopted in this Experiment. There are 15 persons, of 11 facial images each, i.e. 165 images in total. The size of each image is 120x91 with

256 gray levels per pixel. These images are taken at different angle of view, different expression and illumination, and parts of the images are not integral. Fig.2 shows a typical example of images for one person.

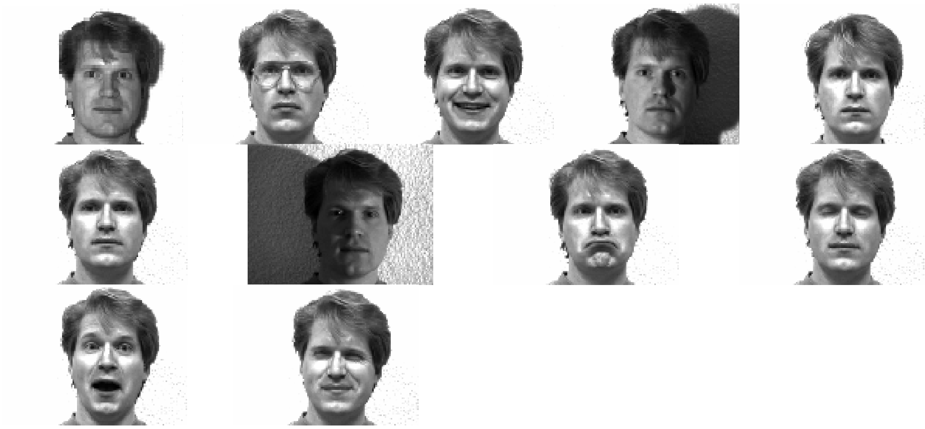


Fig. 2. Typical example with 11 face images for one person in the Yale face image database

In this Experiment, we use the first five images of each person for training and the remaining six for testing. Thus the total number of training samples and testing sample are 75 and 90 respectively.

Table 2. Classification error numbers and optimal recognition rates based on CCA and GCCA under the three kinds of feature fusion strategies in the Minimum-Distance Classifier

Number of axes (*d*)	CCA			GCCA		
	FFS1	FFS2	FFS3	FFS1	FFS2	FFS3
1	72	72	73	40	65	40
5	27	29	27	21	22	19
10	17	17	15	8	9	7
15	14	17	16	2	4	2
20	8	9	8	2	4	2
25	9	9	5	2	4	2
30	7	7	5	2	4	2
35	5	5	4	2	4	2
40	6	6	3	3	3	2
45				2	4	2
Optimal results	95.56%	95.56%	97.78%	97.78%	96.67%	97.78%

By performing a cubic wavelet transformation on the original image using Daubechies orthonormal wavelet, four sub-images with 12×15 resolution can be obtained. We make use of the low-frequency and the level high-frequency sub-images for fusion. By applying the K-L transformation, two groups of 180-dimensional sub-images' vectors will be compressed into 45- dimensional feature spaces. In this sense, two classes of sample feature spaces $A = \{x \mid x \in \mathrm{R}^{45}\}$ and $B = \{y \mid y \in \mathrm{R}^{45}\}$ are constructed. We perform the classification by the Minimum-Distance Classifier based on the correlation feature matrix; results are given in Table 2.

In Table 2, we also present the GCCA's classification results based on FFS1 and FFS2. In addition, the CCA's recognition results under the three kinds of feature fusion strategies are also provided (here the dimension of the projective vectors taken in the step of PCA is 40).

From Table 2, we know that under the three kinds of feature fusion strategies, GCCA's classification results are all superior to CCA's. However, GCCA's classification results under FFS3 are a little bit better than those under the other two feature fusion strategies. When CPV is set to 14 pair, the recognition rate will be stable at 97.78%.

6 Conclusion

In the paper, we have constructed a framework of GCCA applied in face recognition. The proposed new feature fusion strategy (FFS3) has acquired good classification results when it is performed on the two face image database. Compared with the face feature extracted by CCA, as in a process of GCCA, it contains the class information of the training samples, thus, aiming for pattern classification it would improve the classification capability. In addition, experimental results also show that the extracted correlation matrix of face images can reflect the intrinsic feature of images.

Acknowledgements

We would like to thank National Science Foundation of China under Grant No. 60473039. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region under Earmarked Research Grant (project no. CUHK 4223/04E).

References

1. Kirby, M., Sirovich, L.: Application of the KL Procedure for the Characterization of Human Faces. IEEE Trans. Pattern Anal. Machine Intell. 12 (1) (1990) 103–108
2. Turk, M., Pentland, A.: Face Recognition Using Eigenfaces. Proceeding of IEEE Conference on Computer Vision and Pattern Recognition (1991) 586–591
3. Pentland, A.: Looking at People: Sensing for Ubiquitous and Wearable Computing. IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 107–119
4. Swets, D.L., Weng, J.: Using Discriminant Eigenfeatures for Image Retrieval. IEEE Trans. Pattern Anal. Mach. Intell. 18 (8) (1996) 831–836
5. Belhumeur, P.N., Hespanha, J., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection, IEEE Trans. Pattern Anal. Machine Intell. 19 (7) (1997) 711–720
6. Jin, Z., Yang, J.Y., Hu, Z.S., Lou, Z.: Face Recognition Based on Uncorrelated Discriminant Transformation. Pattern Recognition 33 (7) (2001) 1405–1416

7. Chen, L.F., Liao, M., Ko, M.T., Lin, J.C., Yu, G.J.: A new LDA-based Face Recognition System Which can Solve the small Sample Size Problem. Pattern Recognition 33 (10) (2000) 1713–1726
8. Yang, J., Yang, J.Y.: Why can LDA be Performed in PCA Transformed Space? Pattern Recognition 36 (2) (2003) 563–566
9. Yang, M., Ahuja, N., Kriegman, D.: Face Recognition Using Kernel Eigenfaces. Proceeding of IEEE, ICIP (2000) 37–40
10. Liu, Q., Huang, R., Lu, H., Ma, S.: Kernel-based Optimized Feature Vectors Selection and Discriminant Analysis for Face Recognition. Proceeding of ICPR (2002) 362–365
11. Yang, J., Yang, J.Y., Zhang, D., Lu, J.F.: Feature Fusion: Parallel Strategy vs. Serial Strategy. Pattern Recognition 36 (6) (2003) 1369–1381
12. Yang, J., Yang, J.Y., Frangi, A.F.: Combined Fisherfaces Framework. Image and Vision Computing 21 (2003) 1037–1044
13. Hotelling, H.: Relations between two sets of variates. Biometrika 8 (1936) 321–377
14. Borga, M.: Learning Multidimensional Signal Processing. Linköping Studies in Science and Technology, issertations, No.531, Department of Electrical Engineering, Linköping University, Linköping, Sweden (1998)
15. Sun, Q.S., Zeng, S.G., Liu, Y., Heng, P.A., Xia, D.S.: A new Method of Feature Fusion and its Application in Image Recognition. Pattern Recognition, to Appear
16. Sun, Q.S., Liu, Z.D., Heng, P.A., Xia, D.S.: A Theorem on the Generalized Canonical Projective Vectors. Pattern Recognition 38 (3) (2005) 449–452

Clustering Algorithm Based on Genetic Algorithm in Mobile Ad Hoc Network[1]

Yanlei Shang and Shiduan Cheng

The State Key Lab of Networking and Switching,
Beijing University of Posts and Telecommunications,
100876 Beijing, P.R.China
shangyl@bupt.edu.cn

Abstract. In this paper we propose a novel clustering algorithm in mobile ad-hoc network. By selecting the node optimally in both time connectivity and space connectivity as the cluster head with Genetic Algorithm (GA), the resulting clustering algorithm can provide a generic, stable and lower communication overhead cluster structure for the upper-layer protocols. For this clustering scheme, we give analytical model and evaluate the performance by simulation.

1 Introduction

A mobile ad hoc network (MANET) is a collection of wireless mobile nodes that dynamically form a network without the need for any pre-existing network infrastructure or central control. In many circumstances, mobile nodes in an ad hoc network are geographically dispersed and multihop. So the mobile nodes in an ad hoc network will serve both as information sources well as the router to relay packets.

They have potential applications in such scenarios as the absence of network infrastructures, for example, the conference and battle-field environments. In a MANET, the network topology could change rapidly as nodes move, fail, or start up. Network management and routing strategies become big challenges. Many researchers have focused their attention on partitioning the ad hoc network into multiple clusters [1].

The clustering algorithm and the selection criterion of cluster head (CH) are crucial to a clustering ad hoc network. A cluster head should be elected within each cluster to form the upper-layer backbone. With the help of the cluster head, a hierarchical routing protocol can be more easily implemented.

Since the movement is the main cause of uncertainty, we propose a clustering scheme considering the connection duration with its neighbors and the location in the cluster of a specific mobile nodes. In our work, this novel clustering protocol is based on the Genetic Algorithm (GA) [2]. The major difference from the previous research is that we design the clustering operation and selecting the cluster head from node connectivity's point of view, i.e., getting the optimal both in time connectivity and

[1] This work is supported by the NFSC (No.90204003, 60472067, 60402012), the National 973 project (No.2003CB314806), the Fund of DPHE (No.20010013003), RFDR (20010013003) and EYTP.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 968 – 977, 2005.

space connectivity by the GA scheme. In contrast, most prior work focuses on the algorithm design out of regard for routing and management support, lacking an overall cluster stability evaluation. Our goal is to provide a generic and stable cluster structure for the upper-layer protocols. Furthermore, simulation results provide a complete comparison on cluster stability of the GA with other clustering approaches.

The remainder of this paper is organized as follows. Section 2 discusses the related work and our motivation. In section 3, we introduce the two considerations in the proposed clustering scheme. Section 4 depicts the clustering model based on multi-object GA. In section 5, we give the clustering process by GA. Performance evaluations are given in section 6. Section 7 concludes the paper.

2 Related Works and Our Motivation

A multi-hop ad hoc network architecture for wireless systems should be able to dynamically adapt itself with the changing network configurations. Cluster heads are responsible for the formation of clusters and maintenance of the topology of the network. A cluster head does the resource allocation to all the nodes belonging to its cluster. Because the frequent cluster head changes adversely affect the performance of other protocols such as scheduling, routing and resource allocation, Ad hoc network performance metrics such as throughput and delay are tightly coupled with the frequency of cluster reorganization. Choosing cluster heads optimally is an NP-hard problem.

While there are many clustering techniques with CH selection have been proposed, almost all of them fail to guarantee a stable cluster formation. In the previous work, there are two most popular criteria to partition mobile users. One is based on the lowest identifier (Lowest-ID) [1] and the other is based on node maximum connectivity degree (the number of direct links to its neighbors) [3]. There exists an edge between two nodes and one node is called a neighbor if they can communicate with each other directly, i.e. one node lies within the transmission range of another. But these two, along with others, do not provide a quantitative measure of cluster stability. In the former, a highly mobile lowest ID CH will cause severe re-clustering; in addition, if this CH moves into another region it may unnecessarily replace an existing CH, causing transient instability. In the latter, depending on node movement and traffic characteristics, the criterion values used in the election process can keep on varying, and hence also result in instability.

We believe a good clustering scheme should preserve its structure as much as possible when nodes are moving and the topology is slowly changing. Otherwise, recomputation of CH and frequent information exchange among the participating nodes will result in high computation cost overhead. Based on this, we propose a GA-based clustering algorithm which takes into considerations the stability of the cluster.

3 Two Aspects Considered in the Proposed Clustering Algorithm

It is important to distinguish between vital factors and the trivial factors when clustering the Ad hoc network. We propose a novel clustering algorithm which takes into account two metrics: the node connectivity in both time and space in the cluster. On one hand, the precise analysis of the mobile node may include the moving direction, the velocity and the acceleration. These movement modes are difficult and unnecessary for our clustering consideration. We think it is important whether the node can communicate with its neighbors or not. On the other hand, for the transmission power reduction, the selected cluster head should be as possible as near the cluster center.

Before introducing the GA-based clustering criterion, we give a precise definition of the cluster structure to be formed. The ad hoc network is represented by means of an undirected graph G=(V, E), where V is the set of nodes in the graph, and E is the set of edges in the graph (wireless links in the network). Each node would be assigned a unique ID. Because each node belongs to one cluster and the clusters are non-overlapping. Each cluster consists of one CH and zero or more ordinary nodes.

Now, we would give the two considerations used in the proposed clustering algorithm based on GA.

3.1 The Connection Duration with Neighbors

Definition 1. Let $t_i^j(n)$ indicate the connection duration after n_th reconnection between node i and node j. It is a measurement result. As we have known, due to the dynamic nature of the mobile nodes, their association and dissociation to or from each other are unavoidable. If the two nodes could not communicate directly, $t_i^j(n) = 0$.

Definition 2. Let t_i^j denote the smoothed connection duration between node i and node j. Given a new measurement of $t_i^j(n)$, Clustering algorithm updates an estimate of t_i^j by:

$$t_i^j \leftarrow (1-g)t_i^j + g t_i^j(n),\ 0 \le g \le 1\ . \tag{1}$$

Definition 3. Let t_i denote the connection duration of node i with all its neighbor nodes. t_i implies the connection stability of node i and could be regard as one CH selection criterion. We give t_i in the following expression:

$$t_i = \frac{1}{m}\sum_j t_i^j\ . \tag{2}$$

m is the total neighbor number of node i.

3.2 The Weighted Cluster Center in Single Cluster

Before grouping the whole ad hoc network into multiple clusters, we would have to study how to get the cluster center in single cluster. Assumed an ad hoc network topology with m mobile nodes, we compute its center. The coordinates of node i is expressed by (x_i, y_i). We assigned a weight w_i to node i, which presents the computing and power capability. Now we would find the node i (x_i, y_i) in a cluster with the minimum c_i:

$$c_i = \sum_j w_j d_{ij} \cdot \tag{3}$$

where d_{ij} denotes the distance between the node i and any other node j in the same cluster.

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \cdot \tag{4}$$

Replace d_{ij} in (3) with equation (4) and get the partial differential:

$$\frac{\partial c_i}{\partial x_i} = \sum_j w_j \frac{x_i - x_j}{d_{ij}}, \quad \frac{\partial c_i}{\partial y_i} = \sum_j w_j \frac{y_i - y_j}{d_{ij}} \cdot \tag{5}$$

$$\begin{gathered} \frac{\partial^2 c_i}{\partial x_i^2} = \sum_j w_j \frac{(y_i - y_j)^2}{d_{ij}^2}, \quad \frac{\partial^2 c_i}{\partial y_i^2} = \sum_j w_j \frac{(x_i - x_j)^2}{d_{ij}^2}, \\ \frac{\partial^2 c_i}{\partial x_i \partial y_i} = -\sum_j w_j \frac{(x_i - x_j)(y_i - y_j)}{d_{ij}^2} \cdot \end{gathered} \tag{6}$$

The 2-order increment can be expressed with the Taylor series:

$$\begin{aligned} &(\triangle x_i)^2 \frac{\partial^2 c_i}{\partial x_i^2} + 2(\triangle x_i)(\triangle y_i) \frac{\partial^2 c_i}{\partial x_i \partial y_i} + (\triangle y_i)^2 \frac{\partial^2 c_i}{\partial y_i^2} \\ &= \sum_j w_j \frac{\left[(y_i - y_j)\triangle x_i - (x_i - x_j)\triangle y_i\right]^2}{d_{ij}^3} \geq 0 \end{aligned} \tag{7}$$

So c_i is a protruding function and we could derive its minimum when (5) equal to zero, i.e.:

$$x_i = \frac{\sum \frac{w_j x_j}{d_{ij}}}{\sum \frac{w_j}{d_{ij}}}, \quad y_i = \frac{\sum \frac{w_j y_j}{d_{ij}}}{\sum \frac{w_j}{d_{ij}}} \cdot \tag{8}$$

From equation (8), we would have to get d_{ij} before compute the x_i and y_i. But the expression of d_{ij} in (4) still contains x_i and y_i. Of course, we could get x_i and y_i by iterative algorithm with equation (4) and (8). We would like to attempt another method to arrive these x_i and y_i for computing reduction.

We consider the square of d_{ij} in (4):

$$d'_{ij} = d_{ij}^2 = (x_j - x_i)^2 + (y_j - y_i)^2 \cdot \tag{9}$$

From the radio transmission theory we know that the receiving power is proportional inversely to the square of the transmission distance. When c_i in equation (3) arrive the minimum, the needed transmission power is minimized. The node (x_i, y_i) with the minimum c_i is just the weighted cluster center in single cluster.

Now according to (6)(7)(8) and (9), we could get the explicit solution:

$$x_i = \frac{\sum w_j x_j}{\sum w_j}, \quad y_i = \frac{\sum w_j y_j}{\sum w_j} \cdot \tag{10}$$

3.3 The k Cluster Centers in the Global Network

Based on the identification of the cluster center in single cluster, we could begin computing the k cluster centers in the global network topology. Consider an ad hoc network with n mobile nodes (x_j, y_j) $j = 1, 2, \ldots n$. We would group the general network into k clusters and select k cluster centers (x_{qi}, y_{qi}) $i = 1, 2, \ldots k$. The chosen k nodes should minimize the following c:

$$c = \sum_{i,j} c_{ij} d_{ij} \ . \tag{11}$$

$$c_{ij} = \begin{cases} 1 & there\ is\ a\ edge\ between\ node\ i\ and\ j \\ 0 & there\ is\ no\ edge\ between\ node\ i\ and\ j \end{cases}$$

d_{ij} is the distance between the cluster center node i and node j.

We will select the k cluster center nodes (x_{qi}, y_{qi}) $i = 1, 2, \ldots k$ among all n mobile nodes (x_j, y_j) $j = 1, 2, \ldots n$ $(k \le n)$ according to the following two steps:

(1) Select any k nodes as the temporary cluster center (x_{qi}, y_{qi}). Then all the other nodes would be assigned to these k center nodes according to the minimum distance, i.e.:

$$c_{ij} = 1, \ if\ d_{ij} = \min_k (d_{kj})$$

We will have to compute the cost c.

(2) We treat all the nodes which belong to center node (x_{qi}, y_{qi}) as one cluster temporarily. And the cluster center node (x'_{qi}, y'_{qi}) is recomputed by the algorithm mentioned in sub-section 3.2. We compare the two cluster center node (x_{qi}, y_{qi}) and (x'_{qi}, y'_{qi}). If the costs c' equal to c approximately, we think the node with the minimum distance to (x_{qi}, y_{qi}) just is the cluster center node. Otherwise, the step (1) should repeat and attempt other different k nodes.

4 The Clustering Model Based on Multi-object Genetic Algorithm

Due to treating the connection duration and the CH location as the selection criterion of the CH, we would have to optimize these two objects simultaneously. Now we have modeled the CH selection as a multi-object decision-making problem. The primary differences between multi-object and single-object decision-making lie in two aspects. Firstly, multiple objects could not be measured in the same criterion. Secondly, there are maybe conflicts among multiple objects. Due to the non-existence of the common maximum among all object functions, we could not combine multiple objects into single-object decision-making. Consequently, we can only get the non-deterioration solutions or Pareto solutions.

4.1 The Mathematical Model of Multi-object Optimization

A multi-object optimization problem could be expressed by the following:

$$(VOL)\max\left[f_1(x), f_2(x), \ldots f_p(x)\right], \; s.t. \; x \in X \; . \tag{12}$$

where $x = (x_1, x_2, \ldots x_n)^T$ is a n-dimension vector. Its space is called the decision-making space. $f_1(x), f_2(x), \ldots f_p(x)$ are the objective functions. The space of p-dimension vector ($f_1(x), f_2(x), \ldots f_p(x)$) would be called as objective space. X is a executable set in decision-making space. So we could regard the multi-object optimization as a vector optimization.

If x^* is a non-deterioration solution of a multi-object optimization, it is sufficient and necessary for any given $v_j > 0$, $j = 1, 2, \ldots, p$, there exist δ_j ($j = 1, 2, \ldots, p$), and x^* is one optimal solution of the following expression [4]:

$$\max \sum_{i=1}^{p} v_i f_i(x) \; . \tag{13}$$

$s.t. \; f_i(x) \geq \delta_i, \; i = 1, 2, \ldots p, \; x \in X$

Then we could get the approximative non-deterioration solution set by adjusting δ_j ($j = 1, 2, \ldots, p$).

4.2 The Model of Multi-object Clustering

In our clustering algorithm, the decision-making space $x=(x_1,x_2,...x_n)^T$ are corresponding to $t_i^j(n)$ and coordinate of node i. The objective function $f_1(x),f_2(x),...f_p(x)$ are corresponding to t_i (Equation 2) and c_i (Equation 3). We could get the general expression of the selection of the cluster head.

$$(VOL)\min\sum_{i=1}^{2}v_i f_i(x)\ \cdot \tag{14}$$

$s.t.\ f_i(x)\le\delta_i,\ i=1,2,\ x\in X$

where X is the profile set of the nodes. δ_i are a series of thresholds. v_i are the weights and there is the following expression:

$$\sum_{i=1}^{2}v_i=1\ \cdot \tag{15}$$

5 Clustering Algorithm Based on GA

As mentioned above, we would select the node optimally in both time connectivity and space connectivity as the cluster head. But the optimization of the combination of two and more unrelated parameters is NP-complete [4]. We utilize the Genetic Algorithm[5][6] to implement the cluster head selection and cluster formation with multiple objects constrained.

5.1 Encoding

This is also called a string representation of the given data which would be the nodes in the network under consideration. If there is a one-to-one correspondence between the search space and string representation, the design of the genetic operator would be considerably less complex. Each chromosome will be represented as a string of integers form where each node ID is present and appears only once. The genetic operator is encoded by the following:

The length of the chromosome equals to the node number of the whole Ad hoc network. If some a node is regarded as a cluster head, its code is '1' in the chromosome, otherwise the code is '0'. The order of code is according to the node ID, beginning with ID 1, and ending with ID n.

Obviously, the chromosome of encoded nodes is one-to-one correspondence with the clustering formation. Due to the constant length of the chromosome, the following mutation and crossover process could become more facilitated.

5.2 Crossover

The purpose of crossover operation is to have more diverse population. It is random in nature and dependent on the rate specified which is best suited for a given application and can be found experimentally. In this implementation, the X-Order1 method is used [2] and the crossover rate is chosen to be 0.8.

5.3 Mutation

Mutation operation is performed to avoid premature convergence by occasional random alternation of randomly determined bit in the given string with a specified rate. For the mutation operator, we use a swap method with mutation rate of 0.1. In this method, from the parent, we randomly select two genes at position j and i, swap them to create the new child.

6 Simulation and Evaluation

The simulations attempt to compare the performance of our clustering algorithm with the Lowest-ID, maximum connectivity clustering algorithms, in terms of stability of clusters being formed. The cluster stability is measured by determining the number of each mobile node either attempts to become or gives up the cluster head role. The clustering formations have been demonstrated in the Fig. 1.

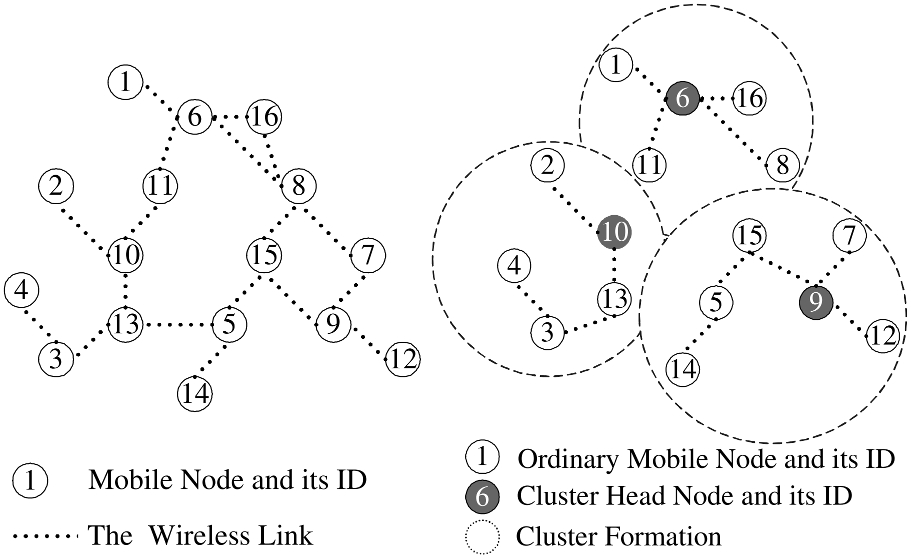


Fig. 1. One ad hoc network topology. The left figure give its un-clustering structure. The right figure is the clustering topology with the proposed scheme based on the Genetic Algorithm.

The y-axis of Fig. 2 (left figure) shows the statistical frequency of cluster head changes by each mobile node, and hence measures the stability associated with each clustering algorithm. As it can be seen from Fig. 3, the GA clustering algorithm leads to more stable cluster formation. Fig. 3 (right figure) depicts the average service time duration of each cluster head in each of the clustering algorithms. The longer the service time of each cluster head, the better its support for cluster stability is. As it can be seen from Fig. 3 (right figure), in the GA model each cluster head has longer average service time than that of any other algorithm. These simulations are performed for up to 200 mobile nodes. The cluster stability is measured by calculating the total number of cluster head changes by all mobile nodes. We performed our simulations using the ns-2 simulator in which we implemented and compared the Lowest-ID, maximum connectivity and our algorithm.

In simulating the GA model, the radius of a cluster, R, is 150m. Lowest-ID and maximum connectivity clustering algorithms form 2-hop clusters. Since it was necessary to ensure that clusters formed by all the schemes approximately cover equal area, the transmission range of each node is set to 80m. A terrain-area of 600m X 600m with nine clusters was considered in our simulations.

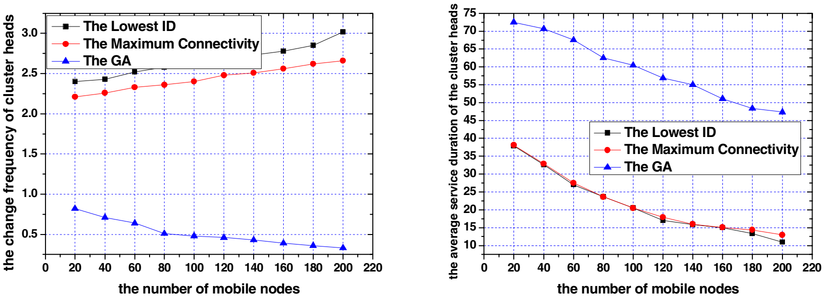


Fig. 2. The simulation results of clustering algorithm based on GA. The left figure shows the statistical frequency of cluster head changes. The right figure depicts the average service time duration of each cluster head. All the simulations are compared with Lowest-ID, maximum connectivity clustering algorithm.

7 Conclusion

In this paper we presented a new clustering approach based on the Genetic Algorithm (GA)[6] in mobile ad hoc. We have demonstrated that this clustering scheme results in more stable clusters than those of other well-known schemes. This stability improvement, however, depends on the selection of an appropriate cluster head. The stable cluster frame is essential for the inter-cluster seamless handover of mobile IP and mobile SCTP in the future work.

References

1. Hou, T.C., Tsai, T.J.: An Access-Based Clustering Protocol for Multihop Wireless Ad Hoc Networks. IEEE Journal on Selected Areas in Communications. Vol. 19. (2001) 1201–1210
2. Man, K.F., Tang, K.S., Kwong, S.: Genetic Algorithms: Concepts and Designs, Springer, (1999)

3. Sivavakeesar, S., Pavlou, G., Liotta, A.: Stable Clustering Through Mobility Prediction for Large-Scale Multihop Intelligent Ad Hoc Network. IEEE Wireless Communications and Networking Conference (WCNC'04), (2004) 1488–1493
4. Yang Y., Xu, Y.H., Li, Q.M.: A Multi-object Genetic Algorithm of QoS Routing, Journal of China Institute of Communications, Vol. 25. (2004) 43-51
5. Mainak, CH., Sajal, K., Damla, T.: An On-demand Weighted Clustering Algorithm (WCA) for Ad hoc Networks, Proceedings of IEEE Globecom (2000) 1697-1701
6. Zheng, Y.X., Tian, J., Dou, W.H.: Vector Constraint Multicast Routing Based on GA, Chinese Journal of Computers, Vol. 26 (2003) 746-752

A Pair-Ant Colony Algorithm for CDMA Multiuser Detector*

Yao-Hua Xu, Yan-Jun Hu, and Yuan-Yuan Zhang

School of Electronic Science and Technology, Anhui University,
Hefei, Anhui 230039, China
sharkahxyh@yahoo.com.cn, yanjunhu@ustc.edu.cn

Abstract. As a novel computational approach from swarm intelligence, an Ant Colony Optimization algorithm attracts more researches, and has been applied in many fields. The paper proposes a pair-ant colony algorithm for multiuser detecting problem based on the Max-Min ant colony algorithm. The optimum multiuser detector has the best performance, but its computation complexity is very high, it is an NP-complete problem. Experiment results show that the proposed method improves the search quality, lower the iteration times, its performances are better than those of the conventional detector and decorrelating detector, and its complexity is lower than that of optimum detector.

1 Introduction

Code-division multiple access (CDMA) is one of several methods of multiplexing that has taken a significant role in cellular and personal communications. Each user's signal is assigned a different signature waveform, and the received signal is the superposition of the signals transmitted by each user. The conventional detection approach is to pass the received signal through a bank of match filters. Each user is detected separately considering the others as interference or noise, which is multiple-access interference (MAI). Thus, multiuser detection (MUD) is necessary to cancellation of MAI, and increase system performance.

The optimum multiuser detector for code-division multiple-access (CDMA) systems [1], which is based on the maximum-likelihood sequence-estimation (MLSE) rule, searches exhaustively for the specific combination of the users' entire transmitted bit sequence that maximizes the so-called log-likelihood function [2]. Its complexity is exponential as a function of user number, and it has already been proved as an NP-complete problem. A great many of researches to find suboptimum detectors have been carried out, like linear and interference cancellation type algorithms.

Ecological system Algorithms which mimic the mechanism of the nature and creatures' behaviors have emerged in the field of evolutionary computation. The typical kinds of this algorithm are genetic algorithms (GA), Immune algorithms, Ant colony optimization algorithms (ACO). Multiuser detection based on GA has been proposed

* This work has been supported by the Research Sustentation Fund to Youth College Teacher of Anhui Province under grant number 2005jq1032zd, China.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 978 – 987, 2005.

by Juntti et al. [3]. The purpose of this paper is to investigate the performance when applied the Max-Min ACO to the MUD problems, to find new methods to solve the problem of MUD in CDMA.

2 Multiuser Detecting Problem of CDMA System

2.1 System Model of CDMA Communication

The system model of multiuser DS-CDMA communication is shown as Fig.1.

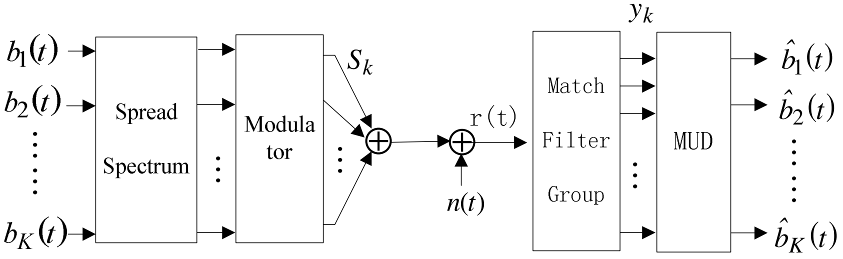


Fig. 1. System model of multiuser DS-CDMA communication

As we known, the Additive White Gaussian Noise channel shared by K users, the receive signal is r(t):

$$r(t) = \sum_{i=-M}^{M} \sum_{k=1}^{K} A_k b_k(i) S_k(t - iT_d - \tau_k) \cos(\omega_c t + \varphi_k) + n(t) \tag{1}$$

Where n(t) is white Gaussian noise with power spectral density N_0, $S_k(t)$ is the normalized signature waveform of user k and is zero outside the interval [0,T], and A_k is the received energy of it. The sampled output of the normalized matched filter for the ith (i =-M;…, M) bit of the *kth* user is

$$y_k(i) = \int_{iT+\tau_i}^{iT+T+\tau_i} r(t) S_k(t - iT - \tau_k) = A_k b_k(i) + \sum_{j=1, j \neq k}^{K} A_j b_j(i) R_{j,k} + n_k \tag{2}$$

Where Define the K*K normalized signal cross-correlation matrices $R(l)$ whose entries are given by

$$R_{k,j}(l) = \int_0^T S_k(t - \tau_k) S_j(t + lT - \tau_j) dt$$

$$n_k = \sigma \int_0^T n(t) S_k(t) dt$$

Note that Equation 2 consists of three terms. The first is the desired information which gives the sign of the information bit b_k. The second term is the result of the multiple access interference, and the last is due to the channel noise. The second term typically dominates the noise so that one would like to remove its influence. Its influence is felt through the cross correlations between the chip sequences and the powers of users.

2.2 Conventional Detector

The conventional detector (CD) is the optimal receiver for the single user system. It detects the bit from user k by correlating the received signal with the chip sequence of user k. Thus it makes its decision at the output of the matched filter bank.

At the output of conventional detector (CD), the decision made on the *ith* bit of the *kth* user is:

$$\hat{b}_{CD,k}(i) = sign(y_k(i)) \tag{3}$$

When MAI terms are significant, the bit error rate is high. A multiuser detector can remedy this problem.

2.3 Multiuser Detector

The multiuser demodulation problem, which needs to be solved at the receiver, is to recover the transmitted sequence b∈L from the sequence y∈L. The decorrelating detector (Decor) applies the inverse of the correlation matrix R^{-1} to the matched filter output to decouple the receive signal. From (3), the result is:

$$\hat{b}_{Decor} = sign(R^{-1}y) \tag{4}$$

Note that the decorrelating detector can completely eliminates MAI in theory, but the performance of it degrades as the cross-correlations between users increase. It isn't the optimal in the sense of minimum bit-error ratio (BER) since additive noise is colored with autocorrelation N_0R^{-1}.

In optimum MUD (OMD), the maximum-likelihood estimation proposed by Verdu in 1986, optimal demodulation of character vector $b = [b_1, b_2, ..., b_s]^T$ has been looked as K-unit estimation problem. The objective of MLSE is to find the input sequence which maximizes the likelihood of the given output sequence. For the CDMA system, the maximum likelihood decision for the vector of bits is equivalent to solving the following formula's maximum problem:

$$\hat{b}_{OMD} = arg \max_{b\in\{-1,+1\}^{2KM}} \left[2y^T b - b^T Hb\right] \tag{5}$$

Where $y = \left[y_1, y_2, ..., y_K\right]$, $H = ARA$, and A is the received energy of users in the one time slot. This equation dictates a search over the 2^K possible combinations of the components of the bit vector, the MLSE detector can be implemented using the Viterbi algorithm. Although the OMD has excellent performance, it is too complex for practical implementation, and suboptimal detector may be applied.

It is a combinatorial optimization problem, and many heuristics have been proposed that attempt to find a suboptimum solution for MUD problem. Such heuristic algorithms include evolution strategies, genetic algorithms, neural networks, tabu search. Our proposal, Pair-ACO detector is also to solve the OMD as a combinatorial optimization problem.

3 Expansion of Ant Colony Optimization Algorithm for MUD

3.1 Basic ACO Algorithm

The ACO, originally introduced in[4] by M. Dorigo, is a new cooperative search algorithm inspired by the behavior of real ants, since ants are able to find good solutions to shortest path problems between a food source and their home colony. Its characteristics include positive feedback, distributed computation and the use of a constructive greedy heuristic. These characteristics can make the ACO outperform other heuristics for some applications. The ACO approach has been applied recently to scheduling problems, job-shop, and single machine tardiness problems.

In ACO, several generations of artificial ants search for good solutions. In general, ants that find a good solution mark their paths through the decision space by putting some amount of pheromone (τ_{ij}) on the edge of the path (i, j). The following ants of the next generations are attracted by the pheromone so that they search in the solution space near previous good solutions. In addition to the pheromone values, the ants will usually be guided by some problem-specific heuristic (η_{ij}) for evaluating the possible decisions.

Firstly, each ant is set on some randomly selected point and begins constructing a valid tour starting from the initial point. A tour is successively built by choosing the next node probabilistically according to a probability distribution proportional to:

$$P_{ij}^{k} = \frac{\tau_{ij}^{\alpha} * \eta_{ij}^{\beta}}{\sum_{k \in s} \tau_{ij}^{\alpha} * \eta_{ij}^{\beta}}, \ if\, j \notin Tabu_{k};\ else\ P_{ij}^{k} = 0 \tag{6}$$

α , β weighs the relative importance of the trail strength and the heuristic information. To keep touring, every ant maintains a tabu list.

Secondly, after all ants have constructed a complete tour, the trail is updated. Every ant lay down a different quantity of pheromone, the trail are updated according to the formula:

$$\tau_{ij}^{new} = \rho \cdot \tau_{ij}^{old} + \sum_{k=1}^{M} \Delta\tau_{ij}^{k} \tag{7}$$

Where ρ is the decline of the pheromone (ρ <1) and M is the number of ants. The amount $\Delta\tau_{ij}^{k}$ is zero if path (i, j) is not used by ant k in its tour. The local and global trail updating is similar to a reinforcement learning scheme in which better solutions get a higher reinforcement.

The performance of ACO can be enhanced by the MAX-MIN Ant system (MMAS) [5]. Based on basic ACO algorithm, MMAS method allow only the best ant to update the trails, and use the explicit maximum and minimum trail strengths ($\tau_{\min}, \tau_{\max}$) on the paths to alleviate the early stagnation problems.

3.2 Pair-ACO Algorithm

The main idea of Pair-ACO [6] algorithm is that there are two ants in a touring instead of one ant in Basic-ACO algorithm. The two ants start a touring from a same point, and choose the next step from different directions.

By sharing a same tabu list, two ants explore the same search space; they cooperate for the whole touring time, so the complexity of the Pair-ACO does not higher than the Basic-ACO. At any point of search, the Pair-ACO has two directions to find the next point in their solution, but the BACO only one way to select. From this point of view, Pair-ACO algorithm has more chance to escape from local optimal to find global optimal than the Basic-ACO. The Pair-ACO touring is slightly shorter than the one-ant touring, and its iteration times are also shorter.

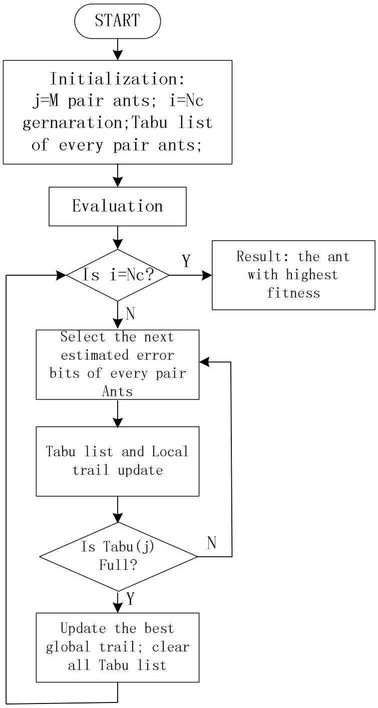


Fig. 2. The structure of the proposed Pair-ACO used to detect the transmitted users' bit

3.3 Multiuser Detector Based on Pair-ACO Algorithm

In this paper, we employed the Pair-ACO in order to detect the transmitted users' bit b(i) vector in the every time slot 'i', where the so-called objective function is defined by the Equation (5). The structure of the proposed Pair-ACO based multiuser detector can be understood with the aid of the flowchart shown in Fig.2.

Make decision on the match filter group output signals in one bit time slot, the dimension of the problem is K, it is also the number of communication system user. The Pair-ACO algorithm search in the K dimensions to find the possible error bits.

Pair-ACO detector commences its search for the optimum solution at the i=0th generation with an initial population of individuals, each consisting of initial bits drives from the match filter outputs. The total iteration times are Nc. The number of individuals in the population is given by the population size M, which is fixed throughout the entire iteration process.

4 Simulation Results

In this section we present several simulation results to compare the proposal with the Basic-ACO detector, the conventional detector (CD) and decorrelating detector (Decor). The performances of these detectors are all in white Gaussian synchronous channels with assumption of ideal power controlled. The length of spread spectrum sequence is 63, and we take the Gold sequence for all users.

The number of communication system user is K, and population size of the basic ACO is M=K, population size of our Pair-ACO is M=0.5K, decline of pheromone ρ=0.8, $\Delta\tau$=0.5, α and β are all equal to 1.

4.1 Bit Error Ratio with SNR

In the simulation system, we transmit 1M×10 bits and receive the signal by Conventional Detector (CD), Decorrelation Detector (Décor), Basic ACO and Pair-ACO separately. Under different channel signal-noise ratio (SNR), the performance of these detectors shows in the Fig.3 and Fig.4. AS single user has no MAI, we take single user performance as the OMD result (Optimal) in Fig.3 and Fig.4.

In Fig.3, the users number of CDMA communication system is 20. It is seen that the Pair-ACO an Basic-ACO receivers yield a performance gain of 1dB with respect to the Decor receiver. The Pair-ACO receiver achieves a performance very close to the bound of optimal receiver than the Basic-ACO.

In Fig.4, instead, the users number of communication system is 40. The results are again obtained through an average over 10 simulation times. Due to the larger number of users, the Decor receiver performance is now quite worse than that of Basic-ACO. The Pair-ACO receiver performance is significantly better than that of the Basic-ACO and Decor receivers. For example, for a BER equal to 10^{-4}, the Pair-ACO achieve a performance gain of more than 2dB with respect to the Basic-ACO and additional more higher than the Decor receiver.

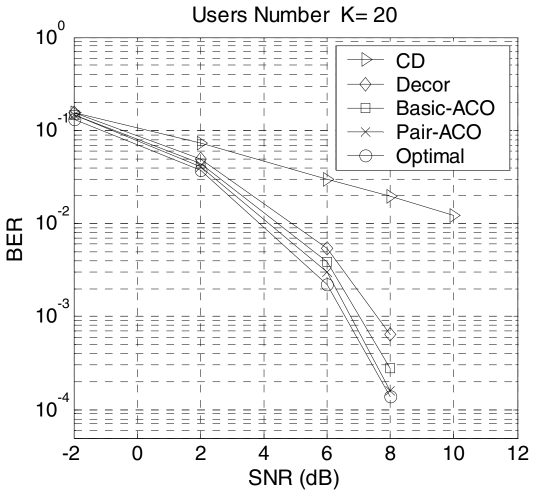


Fig. 3. The bit error ratio of Pair-ACO multiuser detector compare with other detectors with different channel SNR. The system users' number is 20. The length of Gold spread spectrum sequence is 63.

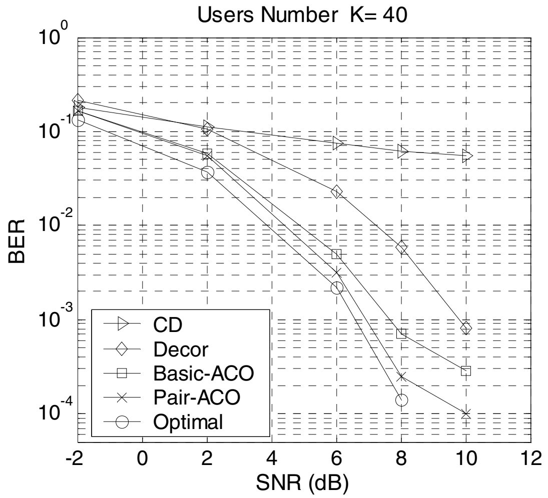


Fig. 4. The bit error ratio of Pair-ACO multiuser detector compare with other detectors with different channel SNR. The system users' number of is 40. The length of Gold spread spectrum sequence is 63.

4.2 Bit Error Ratio with the Number of User

The BER performance comparison with different number of users is shown as Fig.5.

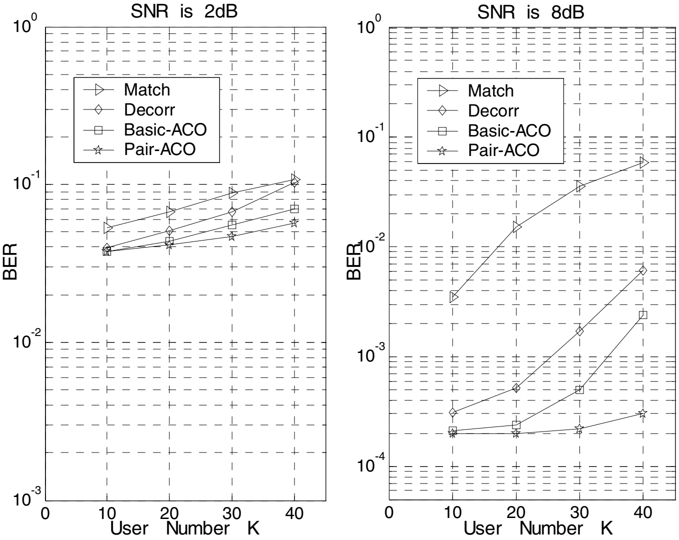


Fig. 5. Bit error ratio for the Pair-ACO multiuser detector compare with Basic ACO, Conventional Detector and decorrelating Detector with different number of users. The channel SNR of left figure is 2 dB, and right one is 8dB. The length of spread spectrum sequence is 63.

From the results, it is obvious that the performance of our method is improved much more than those of conventional match detector and decorrelating detector. Especially, it can keep good performance with a great number of users when signal noise ratio descends. On the other hand, the BER performance of the Pair-ACO detector is better than the Basic-ACO detector. Our proposal is more suit large system user number and a good system circumstance.

4.3 Bit Error Ratio to the Number of Generations

Based on above results, we compare the convergence of our proposal and the Basic-ACO. The SNR of the AWGN channel is 6 dB, and the number of communication system is 40. The population size of the both algorithms is equal to 20. We can see that by using Pair-ACO based detector, a faster convergence rate was attained than the Basic ACO detector. A faster convergence rate implies a reduction in the complexity.

4.4 Complexity Comparison of the Algorithms

From the analysis of our proposal, the computation complexity of Basic-ACO Algorithm is $O(Nc \times K^2 \times M)$, where *Nc* is the total iteration times of the algorithm, *M* is the number of ants, *K* is the dimensions of problem, here *K* is also the number of system users. In the Basic-ACO Algorithm, the number of ants, M is nearly equal to *K*, and the worst complexity is $O(K^4)$. In Pair-ACO, M is the number of pairs of ant, and it is

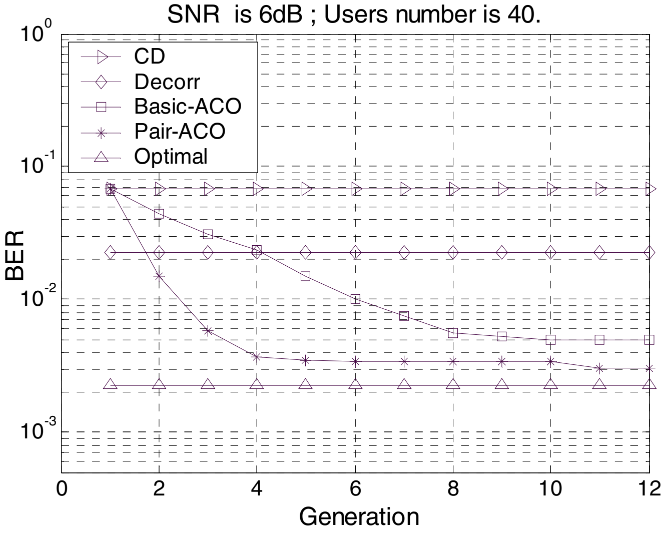


Fig. 6. Bit error ratios performance with respect to the number of generations of the Basic ACO based detector and Pair-ACO based detector at an SNR of 6 dB with number of system users is 40

lower than that of Basic-ACO, and the convergence of it more rapid, so the *Nc* is subsequently lower. The complexity of Pair-ACO is $O(K^3)$. Their complexity are all much lower than that of the optimal multiuser detector, $O(2^K)$, it is a exponential function with number of users, especially when the number *K* is big enough.

5 Conclusions

In this paper, a kind of suboptimal multiuser detector methods, the Pair-ACO Algorithm detector is developed. As expected, the performance of the proposed method is very good, especially when users number increase or signal noise ratio descend the performance of it drops much slower than the conventional receiver and decorrelating detector. Its convergence rate is also faster than basic ant colony algorithm detector. Moreover, the computational complexity is much lower than that of OMD receiver.

References

1. Verdu, S.: Minimum probability of error for asynchronous Gaussian multiple access channels. IEEE Trans. Information Theory, Vol. 32. (1986) 85–96
2. Verdu, S.: Multiuser Detection. Cambridge University Press, (1998)
3. Juntti, M.J., Schlosser, T., Lilleberg, J.O.: Genetic algorithms for multiuser detection in synchronous CDMA. Proc IEEE Information Theory, Germany (1997) 492

4. Dorigo, M., Gambardella, L.M.: Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Trans. Evolutionary. Computation 1(1997) 53–66
5. Stutzle, T., Hoos, H.: MAX-MIN ant system. Future Generation Computer System, Vol.16. (2000) 889–914
6. Wu, Bin., Shi, Zhong-Zhi.: An Ant Colony Algorithm Based Partition Algorithm for TSP. Chinese Journal of Computers, Vol.24. (2001) 1328-1333

An Enhanced Massively Multi-agent System for Discovering HIV Population Dynamics

Shiwu Zhang[1], Jie Yang[1], Yuehua Wu[1], and Jiming Liu[2]

[1] University of Science and Technology of China, Anhui, China
[2] Hong Kong Baptist University, Kowloon Tong, Hong Kong
{swzhang, jieyang}@ustc.edu.cn

Abstract. In this paper, we present an enhanced massively multi-agent system based on the previous MMAS for discovering the unique dynamics of HIV infection [1]. The enhanced MMAS keeps the spacial characteristics of cellular automata (CA), and employs mathematical equations within sites. Furthermore, new features are incorporated into the model, such as the sequence representation of HIV genome, immune memory and agent remote diffusion among sites. The enhanced model is closer to the reality and the simulation captures two extreme time scales in the typical three stages dynamics of HIV infection, which make the model more convincing. The simulation also reveals two phase-transitions in the dynamics of the size of immune memory, and indicates that the high mutation rate of HIV is the fatal factor with which HIV destroys the immune system eventually. The enhanced MMAS provides a good tool to study HIV drug therapy for its characterizing the process of HIV infection.

1 Introduction

1.1 Human Immune System

Human Immune System is a highly sensitive, adaptive and regulated complex adaptive system that involves numerous interactions among different types of cells, virus and bacteria. The immune system serves as the protector of body by killing virus or foreign bacteria, because it can distinguish non-self bacteria and remember the types of invaders. The immune system is composed of lymphoid and white blood cells, B cells and T cells are major categories of white blood cells.

T cells play a central role in human immune system. They can find an invader by recognizing its envelope molecule segments, and inspire an immune response. In the immune response, the immune system creates lots of T cells and antibodies that serve as the killers of infected cells and the recognized virus respectively.

1.2 HIV Infection

HIV is a type of RNA virus. A virion consists of RNA molecules, reverse transcriptase and integrase that are surrounded by several layers of protein coats.

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 988–997, 2005.

Out of the envelope membrane there are many molecules named GP120 that have selective tropism for CD4 T cells, and also segments that could be recognized by CD4 T cells. When a virion infects a T cell, it will bind and enter the cell first, then replicate its RNA into many copies, and assemble new virus with the same RNA. The infected cell will die for cytopathic effect. Thus the population of CD4 T cells decreases sharply at the initial stage of infection, and HIV population increases rapidly simultaneously. Along with the infection, the immune system will recognize virus and inspire an immune response. Lots of T cells and antibodies will be produced to kill infected cells and virus respectively. The initial stage often lasts several weeks.

Because HIV replication fidelity is very low, and HIV RNA determines HIV surface molecules segments, Some mutated virus with different surface molecules segments will avoid being killed. The remaining virus may continue to infect healthy T cells. Then the infection goes into the clinical latency stage. The stage may last several years comparing the short initial stage. Finally, the immune system collapses, and the third stage, onset of AIDS comes. Details on HIV structure, life cycle and infection process can be found in [2][3].

1.3 Related Work

The unique HIV-immune dynamics has attracted many researchers to model HIV infection. Modeling HIV infection may help people to realize the underneath mechanisms in HIV dynamics that are hard observed with biological methods. Next, we will review the models on HIV dynamics in two aspects: (1) whether a model reflects the reality of HIV infection; (2) whether the typical three stages HIV dynamics is well reproduced.

The models of HIV infection can be classified into two categories: (1) cellular automata models, including the Monte Carlo model by Pandey et al. [4][5][6][7][8], the physical space model by Santos et al.[9][10], the "shape" space model by Hershberg et al. [11], the sequence space model by Kamp et al. [12][13], Here "shape" space and sequence space are all space constructed by the genome of HIV; (2) System equation models, including Kirsher's ordinary differential equation (ODE) model on HIV infection and therapy [14][15][16], Hraba's model on immune cells population [17]. CA models emphasize the spatial structure of the immune system. However, because of the heavy computational cost caused by numerous interactions, many simplifications are employed in these models, which decrease the reliability of the result. System equation models adopt mathematical equations to simulate interactions among entities, while they cannot capture the unique spatial emergence from local interactions.

In the above developed models, only the physical space model by Santos et al. and "shape" space model by Hershberg et al. reproduced the typical three stages HIV dynamics. However, the two models all are deficient for unreasonable simplifications. The physical space model ignores the inherit inhomogeneities in HIV strains by setting the probability of different HIV strain being rooted, while the "shape" space model omits the spacial interactions of HIV infection, and supposes that immune cells may diffuse on the "shape" space, which means the immune cells can mutate too. In fact, the immune cells are produced by the immune system and cannot mutate.

We have reported an MMAS model for discovering HIV-immune infection dynamics in [1]. In the MMAS model, there are three types of agents: HIV, T cells and O cells. System mathematical equations are adopted to simulate agents interactions within the site of a two dimensional lattice. Because the average high density within a site, the setting does not affect the performance of the system. However, there are still flaws in the MMAS model. For example, the model does not account for immune memory. Human immune system can remember invaders that have stimulated immune response, and kill them soon when they enter human body again. Simulation result discovers the typical three stages HIV dynamics. However, the onset of AIDS is only inspired after the factor that the immune system are damaged to reproduce healthy immune cells, which is still in debate now.

In order to obtain a more convincing model that reflects the characteristics of human immune system and HIV infection, we present an enhanced MMAS model to discover HIV dynamics, in which the sequence representation of HIV genome, immune memory and agent remote diffusion among sites are incorporated. The meaning of our work is twofold: (1) to better understand HIV dynamics and the mechanism that HIV destroy human immune system eventually, (2) to obtain and design new HIV drug therapies. Through the simulation of HIV infection, we can design new drug therapies which prevent different phases of HIV infection and validate the efficiencies of the therapies.

2 The Computational Model

In this section, we will introduce the MMAS model briefly, then present the enhanced MMAS model.

2.1 The MMAS Model

In the MMAS model, HIV strains are denoted as an integer randomly selected from a gene space. All agents interact on a two dimensional lattice that simulates the organism of human body. The interactions among the agents are listed below:

– **T cell recognize and stimulate an immune response**: T cells can recognize HIV at a site if their "gene" are same. If a T cell recognizes a virion, it will kill all virus belonging to the recognized strain at the same site and stimulate the neighboring sites to reproduce the same T cells.
– **HIV infection and propagation**: Within a site, HIV can infect T cell or O cell with probability P_{inf}. After being infected, a cell will die and produce N_{ch} new virus with the same gene as that of the infecting virus. Simultaneously when virus are reproduced, they may mutate into a new strain with probability P_m.

- **Agents diffusion**. Agents may diffuse from high-density sites to low-density sites according to the diffusion equation:

$$D^{t+1}(i,j) = D^t(i,j) + (\sum_{k=neighbors} D^t(k)/m - D^t(i,j)) * C_d$$

In the above equation, $D^{t+1}(i,j)$ means the density of agents at site $[i,j]$. C_d is the diffusion constant. m denotes the number of the site's neighbors.

- **Agents natural creation and death**. Cells and virus will die for natural reason with probability P_{nd} at each step. Cells also can be created by organism, which mimics the immune system's recover ability from the immune suppression generated by infection.

The above rules has captured main characteristics of HIV dynamics in human immune system. The simulation results display the typical three stages dynamics approximately. A "shape" space emergence in the model indicates that HIV population of various strain changes from a power-law distribution to an exponential distribution. However, the representation of HIV strains and the absent of immune memory limit the application of the MMAS model.

2.2 Enhanced MMAS Model

In the enhanced MMAS model, we will refine the model according to three aspects: sequence gene space, immune memory and agent diffusion style.

Sequence Representation of HIV genome A RNA of HIV consists of a sequence of genes. Virus will mutate on one or several locations of their RNA sequence, thereby the mutative virus will keep certain similarity with original HIV. Here we adopt a sequence of strings with length l to denote virus genome. Two alternative integer, 0 or 1 is assigned to each location of the sequence. At each step, every location of the sequence mutates with probability P_m. Thus a genome space with size $2.^l$ is constructed. Correspondingly, each T cell is also assigned with a sequence of strings that denote its complementary shape to virus, complementary shape denotes T cells surface molecules segment which could recognize HIV strains with typical surface molecules segments. Figure 1 displays the process of HIV propagation, HIV mutation and T cells creation in an immune response.

The molecules segments on virus surface are determined by virus genome, and T cells recognize virus by virus segments. Thus the process of T cells recognizing HIV is determined by the similarity between the genome of HIV and the complementary shape of T cells. The similarity is higher, the successful recognizing rate is higher too. The genome dissimilarity(distance) between the virion and the T cell is calculated according to Equation 1.

$$D_{vt} = \Sigma_{k=1}^{l} \mid (g_v(k) - g_t(k)) \mid \tag{1}$$

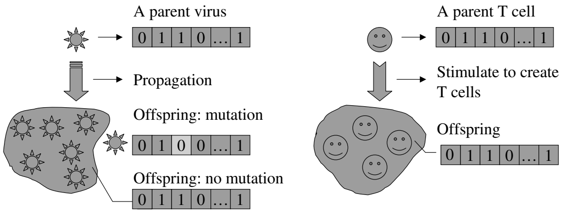


Fig. 1. The genome of a virion and the complementary shape of a T cell in HIV infection

In the above equation, G_v and G_t denote a genome of a virus and a T cell respectively, $g_v(i)$ and $g_t(i)$ denote the ith elements in the arrays respectively. The probability that the virion is recognized, P_r, is calculated according to Equation 2.

$$P_r = (l - D_{vt})/l \tag{2}$$

If there exist n virus and one cell, the probability for the cell to be infected is $1 - (1 - P_r)^n$.

Immune Memory Immune memory plays an important role in immunological activities. Human immune system can remember virus strains that have been recognized before. When the virus belonging to the strains that have been remembered enter body, the immune system will soon stimulate an immune response and kill the virus before they propagate. This is the theoretical basis of vaccines. In the enhanced MMAS model, a global memory repertory is added, which stores all HIV genomes that have been recognized and stimulate immune responses. At the beginning of a simulation, the memory repertory is empty, which means T cells have not recognized any virus in the body. However, along with the infection process, more and more HIV strains are recognized and added into the memory repertory.

The role of immune memory lies in the natural creation of T cells by organism. When a T cell is created, its complementary shape is determined by the repertory with probability P_c, otherwise it is randomly selected in “genome space”. The immune system will find the memorized strains of invaders easily and kill the invaders soon, thus the effect of immune memory is incorporated into the model.

Remote Diffusion In the MMAS model, a two dimensional lattice is set as an interacting environment. Virus and cells interact within a site of the lattice, and diffuse from high-density sites to low-density neighboring sites. However, the diffusion of virus may not be so regular for various topologies of organism and the transport effect of humoral liquid or capillary vessel. That is, virus or cells may jump from one site to a distant site with the help of humoral liquid or capillary vessel.

In our enhanced MMAS model, agents not only diffuse to its neighbor sites, but also diffuse to remote sites with probability P_j. The theoretical analysis and simulation result have proved that such remote diffusion constructs a dynamical small world network [18].

3 Experimental Result

In this section, we will present the experimental result and some discussions. First we will observe HIV dynamics in Experiment A. In the experiment, the size of the lattice is 50 × 50, the system updates in a synchronized parallel way. At the beginning of HIV infection, 10 virus with the same random selected genome are added into the lattice. Then the populations of HIV and cells are recorded in HIV infection, so are those of HIV strains and immune memory size. The simulation lasts 1000 steps, each step corresponds to one day. The parameters in Experiment A are listed in Table 1

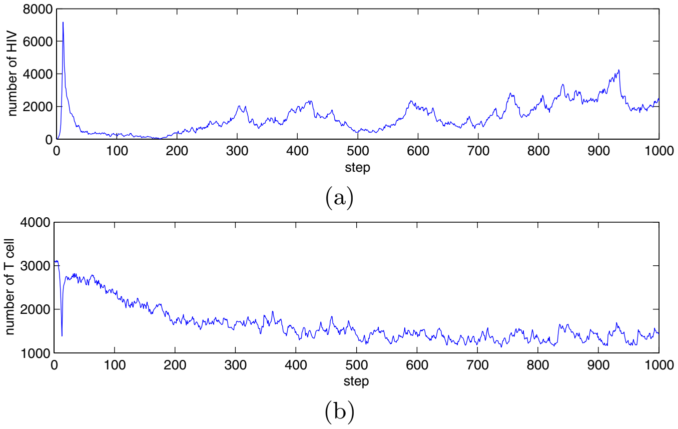
Table 1. The parameters in Experiment A

P_i	0.2	l	12
N_{ch}	6	P_c	0.25
P_m	0.005	C_d	0.1
P_j	0.1		

3.1 HIV Population Dynamics

Figure 2 (a),(b) display HIV and T cell population dynamics in Experiment A respectively. From the figure, we can observe the typical three stages dynamics easily. After the virus enter the body, an immune response that lasts several weeks is triggered, and virus are almost rooted up. It is the initial stage of HIV infection (step 0∼50). However, some variants of virus escape from the initial response and their population grows gradually in struggling with human immune system. It enters the clinical latency period (step 50∼800) which lasts several years. It could be observed that the immune system suppresses virus to low level for many times, while virus population becomes gradually larger in the clinical latency stage. Finally HIV destroy the immune system and it comes the onset of AIDS (after step 800). In Figure 2 (b), T cell population dynamics is complementary to that of HIV, which reflects the battle between HIV and the immune system.

The simulation reproduces two extreme time scales from the weeks of initial response to the years of the clinical latency period. In the simulation, the mechanism that decreases immune system's recover ability is not added, which is different from the MMAS model. Comparing Figure 2 (a),(b) to Figure 2 in [1] and Figure 1 in [19], we can conclude that HIV dynamics in Experiment A is closer to the empirical data in [19] than that of the MMAS model. The result indicates that the modifications in the enhanced model capture the key characteristics of HIV infection.



(a)

(b)

Fig. 2. (a) HIV population dynamics and (b) T cell population dynamics in Experiment A

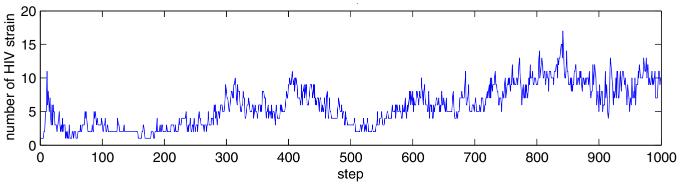


Fig. 3. HIV strain dynamics in Experiment A

Figure 3 displays HIV strain dynamics in Experiment A. HIV strain dynamics is similar to HIV population dynamics, which means the number of HIV strains is an important factor that affects HIV population dynamics. If the number of HIV strains increases, HIV population will increase correspondingly.

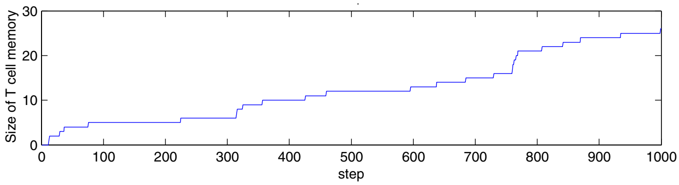


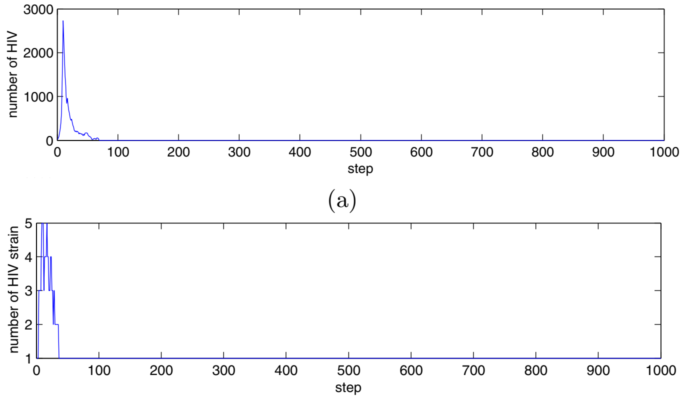
Fig. 4. Immune system's memory size dynamics in Experiment A

Figure 4 displays the size of the immune memory repertory in HIV infection. From the figure we can observe two phase transitions. The first one happens at about the 50th step and the second one happens at about 800th step, which indicate the start of the clinical latency period and AIDS period respectively. In the latency latency period, the immune system controls HIV propagation more and more difficultly.

In sum up, HIV infection after the initial response is a dynamical process in which the number of HIV strain increases, the size of immune memory increases, and T cells population decreases gradually.

3.2 The Role of HIV Mutation Rate

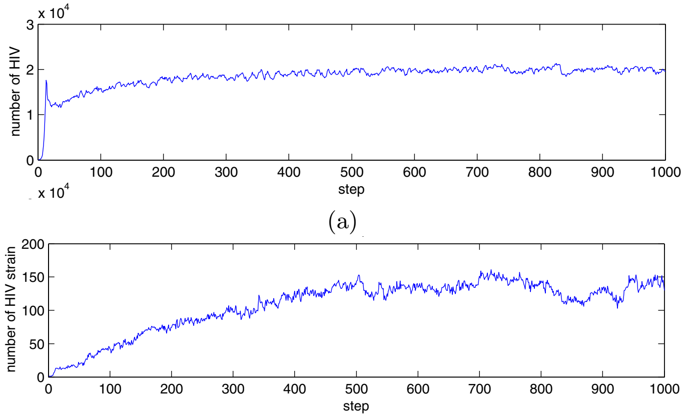
In order to examine the role of HIV mutation rate, we design Experiment B and C, in which HIV mutation rate is 0.001 and 0.01 respectively. Figure 5 (a) and (b) display HIV population and HIV strain dynamics respectively in Experiment B. From the figure, we can obtain that virus are eradicated in the initial immune response because they have not developed more strains to escape from the immune system. Figure 6 (a) and (b) display HIV population and HIV strain dynamics respectively in Experiment C. From the figures we can observe that HIV population recovers and increases soon after the initial immune response. The reason is that the number of HIV strain has become very large before the initial immune response. The Results from Experiment B and C reveal that HIV mutation rate plays a key role in HIV infection. It could control the length of clinical latency period and even the outcome of HIV infection.



(a)

(b)

Fig. 5. (a) HIV population dynamics and (b) HIV strain dynamics in Experiment B



(a)

(b)

Fig. 6. (a) HIV population dynamics and (b) HIV strain dynamics in Experiment B

4 Discussions and Conclusions

In the paper we present an enhanced MMAS model accounting for HIV infection. A comprehensive comparison between our model and other HIV dynamics models according to methodology and results is also presented. Despite the characteristics in the previous MMAS model, Such characteristics as sequence representation of HIV genome, immune memory and remote diffusion are incorporated into the model. HIV genome belonging to a certain strain is represented as a sequence of strings. HIV surface molecules segments are determined by their genome, so are the complementary shape of T cells. The similarity between the genome of virus and the complementary shape of T cells determines the probability of T cells recognizing HIV. The system can remember the recognized strains with a immune memory repertory. Agents adopt remote diffusion with a certain probability for the transport effect of humoral liquid or capillary vessel. The modifications make the enhanced MMAS model closer to the reality of HIV infection. The three stages HIV dynamics obtained from the enhanced model is more similar to the clinical data, which also makes the model more convincing. In the simulation, two phase-transitions are observed in the dynamics of the size of immune memory, the role of HIV mutation rate is also studied.

Because the enhanced MMAS model simulates the whole infection process accurately, it provides a good tools to understand the underneath mechanisms in HIV dynamics, and to design new HIV drug therapies in future. For example, we can examine the effects of new drug therapies that prevent different stages of HIV life cycle, such as preventing virus from binding cells (P_i), inhibiting the viral enzyme reverse transcriptase and viral protease (N_{ch}), disturbing HIV RNA replication fidelity(P_m). Furthermore, drug resistance and HIV vaccine can also be simulated and studied with the enhanced MMAS model.

References

1. Zhang, S., Liu, J.: A Massively Multi-agent System for Discovering HIV-immune Interaction Dynamics. LNAI, Vol. 3446 (2005) 161–173
2. Young, J.A.T.: The Replication Cycle of HIV-1. HIV InSite Knowledge Base Chapter, (1997) http://hivinsite.ucsf.edu/InSite.jsp?page=kb-02&doc=kb-02-01-01
3. Hope, T.J., MD, D.T.: Structure, Expression, and Regulation of the HIV Genome. HIV InSite Knowledge Base Chapter, (2000) http://hivinsite.ucsf.edu/InSite.jsp?page=kb-02&doc=kb-02-01-02
4. Pandey, R.B.: A Stochastic Cellular Automata Approach to Cellular Dynamics for HIV: effect of viral mutation. Theory in Bioscience, Vol. 117 (1998) 32
5. Mannion, R., Ruskin, H., Pandey, R.B.: Effect of Mutation on Helper T-cells and Viral Population: A Computer Simulation Model for HIV. Theory in Bioscience, Vol. 119 (2000) 10
6. Mannion, R., Ruskin, H., Pandey, R.B.: A Monte Carlo Approach to Population Dynamics of Cell in an HIV Immune Response Model. Theory in Bioscience, Vol. 119 (2000) 94
7. Mielke, A., Pandey, R.B.: A Computer Simulation Study of Cell Population in a Fuzzy Interaction model for Mutating HIV. Physica A, Vol. 251 (1998) 430
8. Pandey, R.B., Mannion, R., Ruskin, H.: Effect of Cellular Mobility on Immune Response. Physica A, Vol. 283 (2000) 447
9. Santos, R., Coutinho, S.: On the Dynamics of the Evolution of HIV Infection. (2000) http://arxiv.org/abs/cond-mat/0008081
10. R. Santos and S. Countinho: Dynamics of HIV infection: a cellular automata approach. Physical Review Letters, Vol. 16 (2001) 0168102
11. Hershberg, U., Louzoun, Y., Atlan, H., Solomon, S.: HIV Time Hierarchy: Winning the War while, Loosing all the Battles. Physica A, Vol. 289 (2000) 178–190
12. Kamp, C., Bornholdt, S.: From HIV infection to AIDS: A Dynamically Induced Percolation Transition?. Proc. R. Soc. London B, Vol. 269 (2002) 2035
13. Kamp, C., Bornholdt, S.: Co-evolution of Quasispecies: B-cell Mutation Rates Maximize Viral Error Catastrophes. Physical Review Letters, Vol. 88 (2002) 068104
14. Kirschner, D.E.: Using Mathematics to Understand HIV Immune Dynamics. Notices of the American Mathematical Society, (1996) 191–202
15. Kirschner, D.E., Mehr, R., Perelson, A.S.: Role of the Thymus in Pediatric HIV-1 Infection. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology, Vol. 18 (1998) 95–109
16. Kirschner, D.E., Webb, G.F.: A Mathematical Model of Combined Drug Therapy of HIV Infection. Journal of Theoretical Medicine, Vol. 1 (1997) 25–34
17. Hraba, T., Dolezal, J.: A Mathematical Model and CD4+ Lymphocyte Dynamics in HIV Infection. Emerging Infectious Diseases, Vol. 2(4) (1996) 301–305
18. Manrubia, S.C., Delgado, J., Luque, B.: Small-world Behavior in a System of Mobile Elements. Europhysics Letters, Vol. 53 (2001) 693–699
19. Sloot, P., Chen, F., Boucher, C.: Cellular Automata Model of Drug Therapy for HIV Infection. ACRI2002, LNCS Vol. 2493 (2002) 282–293

An Efficient Feature Extraction Method for the Middle-Age Character Recognition

Shahpour Alirezaee[1,2], Hasan Aghaeinia[1], Karim Faez[1],
and Alireza Shayesteh Fard[2]

[1] Electrical Engineering Department, Amirkabir University of Technology,
Hafez Ave., Tehran, Iran
{Alirezaee, Aghaeini, Kfaez}@Aut.ac.ir
[2] Electrical and Computer Engineering Department, Zanjan University,
Zanjan, Iran
{Alirezae, Shayestehfard}@mail.znu.ac.ir

Abstract. In this paper, we introduce an efficient feature extraction method for character recognition. The EA strategy is used to maximize the Fisher linear discriminant function (FLD) over a high order Pseudo-Zernike moment. The argument, which maximizes the FLD criteria, is selected as the proposed weight function. To evaluate the performance of the proposed feature, experimental studies are carried out on the historic Middle-Age Persian characters. The numerical results show 96.8% recognition rate on the selected database with the weighted Pseudo-Zernike feature (with order 10) and 65, 111,16 neurons for the input, hidden, and output layers while this amount for the original Pseudo-Zernike is 93%.

1 Introduction

Generally, an off-line handwritten character recognition system includes three parts: preprocessing, feature extraction, and classification [1,2,3]. Preprocessing is primarily used to reduce variations of handwritten characters. Feature extraction is essential for efficient data representation and extracting meaningful features for later processing. Up to now, many features have been used in this area among which moments can be considered as one of the widely used feature extractor technique [4]. The classifier is used to make a final decision according to extracted features and acquired knowledge. Generally, there are three categories of classifiers: Neural classifier, Statistical classifier and Structural classifier.

In order to design a high recognition rate system, the choice of feature extractor is very crucial and extraction of pertinent features from two-dimensional character images plays an important role in any OCR system. There are various techniques reported in the literature that deal with this problem. This study has been focused on the Pseudo-Zernike moment.

In this paper we are aiming to develop the Pseudo-Zernike moment feature for character recognition. The main idea is based on maximizing the Fisher Linear Discriminant function over a high order Psuedo-Zernike moment. The EA strategy

D.S. Huang, X.-P. Zhang, G.-B. Huang (Eds.): ICIC 2005, Part II, LNCS 3645, pp. 998 – 1006, 2005.

has been used to maximize the FLD function. For evaluating the proposed idea, it has been applied to the historic Middle-Age Persian characters. Unfortunately the automatic analysis of the Middle-Age Persian documents has not received any attention.

This paper is organized as follows. The second section introduces the Middle-Age Persian. The third section describes the Pseudo-Zernike moments. The linear fisher discriminant function and the weight function will be defined in the fourth section. The fifth section describes the EA strategy. The numerical results are presented in the sixth section. Finally, the conclusions will be discussed in the seventh section.

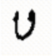
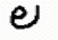
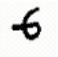
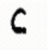
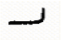
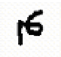
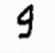
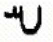
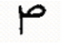
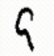
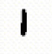
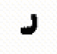
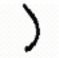
2 The Middle-Age Persian

Iranian languages are branch of "Iranian Indian" as well as "European Indian" languages. The Iranian Indian languages are divided into three distinct periods, "Ancient", "Middle-Age" and "Modern". While the Modern Persian language enjoys from the existence of large number of texts, books, and literatures, unfortunately the same is not true for the Middle-Age and Ancient Persian language. The aim of this research is to provide a recognition algorithm capable of recognizing Middle-Age Persian language texts.

The oldest Middle-Age manuscript consists of several fragments of papyrus, which recently found in the ‘Fayûm’ region in Egypt, and now is in the Royal Museum at Berlin. It is supposed to written in the eighth century. Then four manuscripts written on Indian paper have been appeared in A.D. 1323-1324. They are two copies of the ‘Yasna’ and ‘Vendidad’, containing the ‘Avesta’ with its ‘Zand’ (Pahlavi translation and explanation). These old manuscripts are now preserved in London and Bombay. Next manuscripts, which have been probably written about fifty years later, are preserved in Copenhagen and Bombay. Another manuscript, nearly for the same age, is a mixed collection of Pahlavi texts, written in A. D. 1397. It is in Munich, where the oldest ‘Pâzand-Sanskrit’ (a copy of the ‘Ardâ-Vîrâf-nâmak’, written in A.D. 1410) is preserved. Another ‘Pâzand-Sanskrit’ manuscript, a copy of the ‘Khurdah Avesta’ exists in Bombay. Pahlavi and ‘Pâzand’ manuscripts of the sixteenth century are rather more abundant.

The Middle-Age Persian language has 16 characters (Table.1) and has a right to left direction. There are many similarities between Modern Persian and Middle-Age Persian characters. Some of Middle-Age Persian characters have more than one phonetic equivalent. The exact phonetic equivalent can be extracted from before and after characters. This multi-phonetic property has made this language as one of the most difficult ancient languages in the world. Fig.1 shows a page sample of the Middle-Age documents. We are aiming to provide through this research a feature extraction algorithm for recognizing the Middle-Age Persian characters.

Table 1. Middle-Age Persian alphabet

Character No	Character shape	Phonetic equivalence
1		/k/
2		/s/
3		/p/
4		/Q/, /m/
5		/c/
6		/A/
7		/b/
8		/E/
9		/z/
10		/k/
11		/s/
12		/t/
13		/c/
14		/w/,/r/,/O/,/n/
15		/d/, /g/, /y/
16		/l./

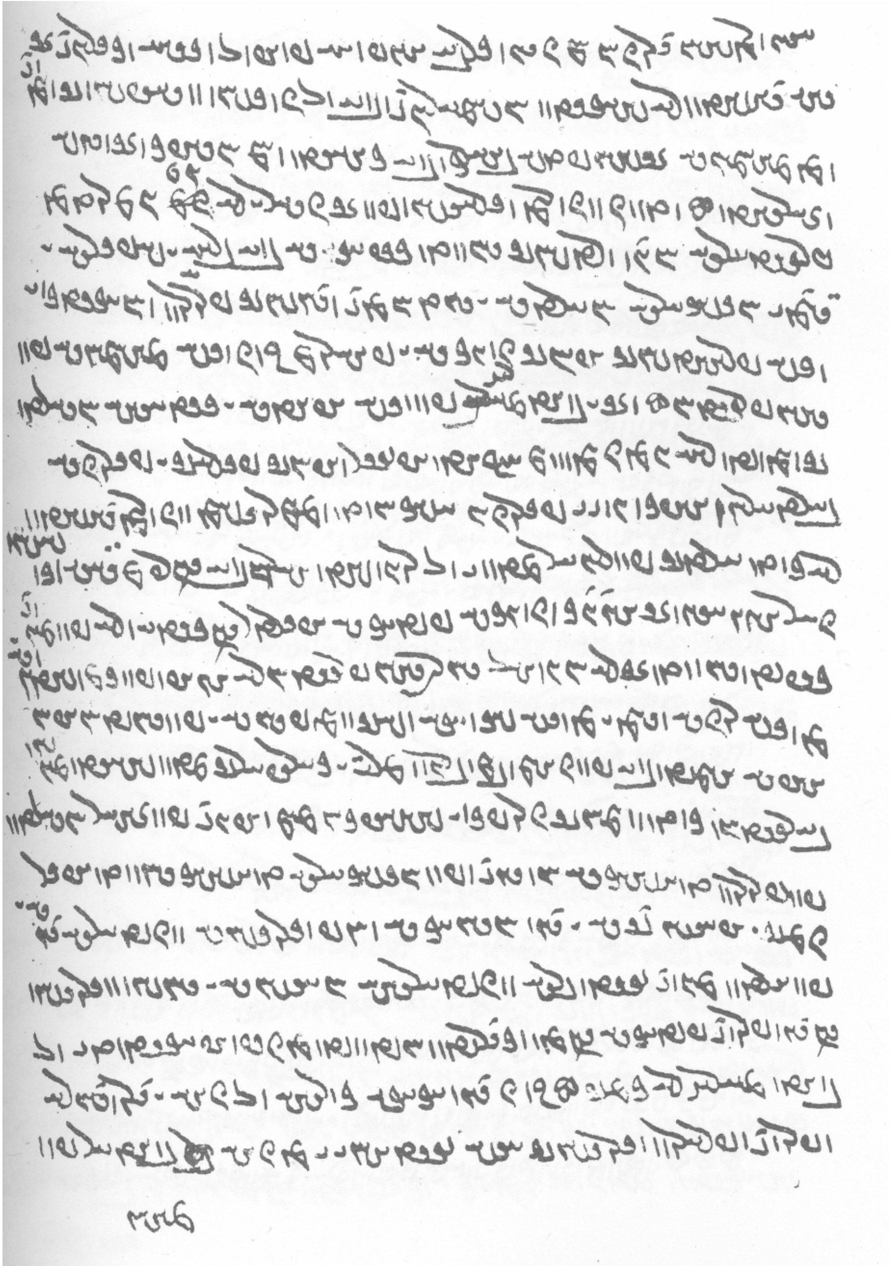


Fig. 1. A page sample of the Middle-Age Persian

3 Pseudo-Zernike Moment Invariants

Statistical-based approaches for feature extraction are very important in pattern recognition for their computational efficiency and their use of global information in an image for extracting features [4], [5]. The advantages of considering orthogonal moments are, that they are shift, rotation and scale invariant and very robust in the presence of noise. The invariant properties of moments are utilized as pattern sensitive features in classification and recognition applications. Pseudo-Zernike polynomials are well known and widely used in the analysis of optical systems. Pseudo-Zernike polynomials are orthogonal sets of complex-valued polynomials defined as [6]:

$$V_{n,m}(x,y) = R_{n,m}(x,y)\cdot \exp(j\cdot m\cdot \tan^{-1}({}^{y}\!/_{x})) \tag{1}$$

$$x^2 + y^2 \le 1$$

Where $|m| \le n$, $n \ge 0$ and Radial polynomials $\{R_{n,m}\}$ are defined as:

$$R_{n,m}(x,y) = \sum_{s=0}^{n-|m|} D_{n,|m|,s}\cdot (x^2+y^2)^{\frac{n-s}{2}} \tag{2}$$

Where,

$$D_{n,|m|,s} = (-1)^s . \frac{(2.n+1-s)!}{s!.(n-|m|-s)!.(n+|m|+1-s)!} \tag{3}$$

The Pseudo-Zernike of order "n" and repetition "m" can be computed using the scale invariant central moments GM_{pq} and the radial geometric moments RM_{pq} as follows:

$$\begin{aligned} A_{n,m} &= \frac{n+1}{\pi} \sum_{(n-s-m)even,s=0}^{{}^{n-|m|}\!/_{2}} D_{n,|m|,s} \\ &\times \sum_{a=0}^{k}\sum_{b=0}^{m} (-j)^b \cdot \binom{k}{a} \cdot \binom{m}{b} \cdot GM_{2k+m-2a-b,2a+b} \\ &+ \frac{n+1}{\pi} \sum_{(n-s-m)odd,s=0}^{{}^{n-|m|}\!/_{2}} D_{n,|m|,s} \\ &\times \sum_{a=0}^{d}\sum_{b=0}^{m} (-j)^b \cdot \binom{d}{a} \cdot \binom{m}{b} \cdot RM_{2d+m-2a-b,2a+b} \end{aligned} \tag{4}$$

where $k = (n-s-m)/2$, $d = (n-s-m+1)/2$, and RM_{pq} is as follows:

$$GM_{pq}=\frac{\sum_X\sum_y f(x,y)(x-x_0)^p(y-y_0)^q}{\alpha^{(p+q+2)/2}} \tag{5}$$

$$RM_{pq}=\frac{\sum_x\sum_y f(x,y)((x-x_0)^2+(y-y_0)^2)^{1/2}(x-x_0)^p(y-y_0)^q}{\alpha^{(p+q+2)/2}} \tag{6}$$

4 The Fisher Linear Discriminant

The Fisher Linear Discriminant (FLD) gives a projection matrix W that reshapes the scatter of a data set to maximize class separability, which can be defined as the ratio of the between-class scatter matrix to the within-class scatter matrix [7]. This projection defines features that are optimally discriminating.

Let $\{\vec{x_i}\}$ be a set of N column vectors of dimension D. The mean of the dataset is:

$$\overrightarrow{\mu_x}=\frac{1}{N}\sum_{i=1}^{N}\vec{x_i} \tag{7}$$

There are K classes $\{C_1,C_2,...,C_k\}$. The mean of class k containing N_k members will be:

$$\overrightarrow{\mu_{x,k}}=\frac{1}{N}\sum_{\vec{x_i}\in C_k}\vec{x_i} \tag{8}$$

The between class scatter matrix is defined as:

$$S_B=\sum_{k=1}^{K}N_k\left(\overrightarrow{\mu_{x,k}}-\overrightarrow{\mu_x}\right)\left(\overrightarrow{\mu_{x,k}}-\overrightarrow{\mu_x}\right)^T \tag{9}$$

The within class scatter matrix is defined as:

$$S_W=\sum_{k=1}^{K}\sum_{\vec{x_i}\in C_k}\left(\vec{x_i}-\overrightarrow{\mu_{x,k}}\right)\left(\vec{x_i}-\overrightarrow{\mu_{x,k}}\right)^T \tag{10}$$

The transformation matrix that repositions the data to be most separable is the matrix W that maximizes:

$$\frac{\det\left(W^T S_B W\right)}{\det\left(W^T S_W W\right)} \tag{11}$$

The analytical solution of the LDA criteria leads to the generalized eigenvectors of S_B and S_W which are the eigenvectors of:

$$S_B . S_W^{-1} \tag{12}$$

In this paper, we are aiming to maximize class separability of the Pseudo-Zernike moments. In the rest, the "W" matrix will be called the weight function. Therefore, our target is to select a weight function, which maximizes the class separability. We have used the EA strategy for maximizing the class separability (Eq.11). The next section presents the EA strategy in details.

5 EA Strategy

Evolutionary algorithms (EA) are stochastic search methods that have been applied successfully in many search, optimization, and machine learning problems [8]. An EA (see the following pseudo code) proceeds in an iterative manner by generating new populations $P(t)$ of individuals from the old ones ($t = 0, t = 1, t = 2, \ldots$). Every individual in the population is the encoded (binary, real) version of a tentative solution. An evaluation function associates a fitness value to every individual indicating its suitability to the problem. The canonical algorithm applies stochastic operators such as selection, crossover, and mutation on an initially random population in order to compute a whole generation of new individuals. In a general formulation, the variation operators are applied to create a temporary population, evaluate the resulting individuals, and get a new population by $P(t+1)$ either using $P^{'}(t)$ or, optionally, $P(t)$. The halting condition is usually set as reaching a preprogrammed number of iterations of the algorithm, or to find an individual with a given error if the optimum, or an approximation to it, is known beforehand.

```
Evolutionary algorithms
    $t = 0$;
    Initialize and evaluate [ $P(t)$ ]
    While not stop condition do
        $P^{'}(t)$ =Variation [ $P(t)$ ];
        Evaluate [ $P^{'}(t)$ ];
        $P(t+1)$ =Select ( $P^{'}(t)$, $P(t)$ );
        $t = t + 1$;
    End while;
```

We have applied the EA strategy to maximize the FLD in the Eq.11 for the 10th order Pseudo-Zernike moment.

6 Numerical Results

Training and testing data sets respectively consist of 150 and 50 samples for each character which were` collected from different ancient documents [9].

In this study, absolute value of the 10th order Pseudo-Zernike moment has been selected as a feature (normalized to A_{00}). We are aiming to select a weight function, which produces maximum separability on the selected database. The EA strategy can yield the weight function, which maximizes the character class separability over the Pseudo-Zernike moment.

After finding the weight function, we will apply it to classify the characters. For classification, a one hidden layer feedforward neural network has been trained [10]. The inputs to the neural network are feature vectors derived from the proposed feature extraction technique. We have set the number of input nodes in the input layer of the neural network equal to the number of feature vector elements. The number of nodes in the output layer is then set to the number of character classes. The number of the hidden layer neurons will be obtained experimentally. In this way, the hidden layer neurons have been increased to yield the best classification rate. The best network specifications are 65, 111, 16 neurons for the input, hidden, and output layers which achieves 96.8% classification rate over the Middle-Persian database.

7 Conclusions

In this paper a weighted Pseudo-Zernike feature was presented. The main idea is based on the optimization of the Fisher linear discriminant function using the EA strategy. The argument, which maximizes the FLD, was selected as the proposed weight function. To evaluate the performance of the proposed feature, experimental studies were carried out on the Middle-Persian character images. The numerical results show 96.8% recognition rate with the weighted Pseudo-Zernike feature (with order 10) and 65, 111,16 neurons for the input, hidden, and output layers while this amount for the original Pseudo-Zernike is 93%.

Acknowledgements

The authors would like to Acknowledge Iran Telecommunication Research Center (ITRC) for supporting this work.

References

1. Cai, J., Liu, Z.: Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(3), (1999)
2. Arica, N., Yarman-Vural, F.T.: An Overview of Character Recognition Focused on Off-Line Handwriting. IEEE Trans. on Sys., Man., and Cybernetics, 31(2), (2001).
3. Casey, R.G. , Lecolinet, E.: A Survey of Methods and Strategies in Character Segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 18(7), (1996)
4. Baily, R. R. , Srinath, M. : Orthogonal moment features for use with parametric and non-parametric classifiers. IEEE Trans. Pattern Analysis and Machine Intelligence, 18(4), (1996) 389-398

5. Belkasim, S.O., Shridhar, M. , Ahmadi, M. : Pattern recognition with moment invariants, a comparative study and new results. Pattern Recognition, 24(12), (1991) 1117-1138
6. Haddadnia, J. , Ahmadi, M. , Faez, K. : An efficient feature extraction method with pseudo-Zernike moment in RBF neural network-based human face recognition system. EURASIP journal on applied signal processing, 9, (2003) 890-901
7. Feng, Y., Shi, P. : Face Detection Based on Kernel Fisher Discriminant Analysis. Automatic Face and Gesture Recognition (2004) 381-384
8. Alba, E. ,Tomassini, M. : Parallelism and Evolutionary Algorithms", IEEE Trans. on Evolutionary Computation, 6(5), (2002)
9. Pahlavy Handwritten Documents. Asian Institute of Shiraz University (1972)
10. Jain, A.k., Mao, J., Mohiuddin, K.M.: Artificial neural networks; A tutorial. Computer, 29(3) (1996) 31 – 44

Author Index