Santanu Chaudhury   Sushmita Mitra
C.A. Murthy   P.S. Sastry
Sankar K. Pal (Eds.)

LNCS 5909

Pattern Recognition and Machine Intelligence

Third International Conference, PReMI 2009
New Delhi, India, December 2009
Proceedings

Springer

Santanu Chaudhury Sushmita Mitra
C.A. Murthy P.S. Sastry
Sankar K. Pal (Eds.)

Pattern Recognition and Machine Intelligence

Third International Conference, PReMI 2009
New Delhi, India, December 16-20, 2009
Proceedings

Springer

Volume Editors

Santanu Chaudhury
Indian Institute of Technology Delhi
Electrical Engineering Department
New Delhi 110016, India
E-mail: santanuc@ee.iitd.ernet.in

Sushmita Mitra
C.A. Murthy
Sankar K. Pal
Indian Statistical Institute
Center for Soft Computing Research
Machine Intelligence Unit
Kolkata 700108, India
E-mail: {sushmita, murthy, sankar}@isical.ac.in

P.S. Sastry
Indian Institute of Science
Department of Electrical Engineering
Bangalore 560012, India
E-mail: sastry@ee.iisc.ernet.in

Library of Congress Control Number: 2009940419

CR Subject Classification (1998): I.4, F.1, I.2, I.5, J.3, C.2.1, C.1.3

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

ISSN 0302-9743
ISBN-10 3-642-11163-7 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-11163-1 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12823147 06/3180 5 4 3 2 1 0

Preface

This volume contains the proceedings of the third international conference on Pattern Recognition and Machine Intelligence (PReMI 2009) which was held at the Indian Institute of Technology, New Delhi, India, during December 16–20, 2009. This was the third conference in the series. The first two conferences were held in December at the Indian Statistical Institute, Kolkata in 2005 and 2007.

PReMI has become a premier conference in India presenting state-of-art research findings in the areas of machine intelligence and pattern recognition. The conference is also successful in encouraging academic and industrial interaction, and in promoting collaborative research and developmental activities in pattern recognition, machine intelligence and other allied fields, involving scientists, engineers, professionals, researchers and students from India and abroad. The conference is scheduled to be held every alternate year making it an ideal platform for sharing views and experiences in these fields in a regular manner.

The focus of PReMI 2009 was soft-computing, machine learning, pattern recognition and their applications to diverse fields. As part of PReMI 2009 we had two special workshops. One workshop focused on text mining. The other workshop show-cased industrial and developmental projects in the relevant areas.

Premi 2009 attracted 221 submissions from different countries across the world. Each paper was subjected to at least two reviews; the majority had three reviews. The review process was handled by the Program Committee members with the help of additional reviewers. These reviews were analyzed by the PC Co-chairs. Finally, on the basis of reviews, it was decided to accept 98 papers. We are really grateful to the PC members and reviewers for providing excellent reviews. This volume contains the final version of these 98 papers after incorporating reviewers' suggestions. These papers have been organized into ten thematic sections.

For PreMI 2009 we had a distinguished panel of plenary speakers. We are grateful to Andrew G. Barto, Andrzej Skowron and Ramesh Jain for agreeing to deliver plenary talks. We also had invited talks delivered by Ajith Abraham, Pawan Lingras, Alfredo Petrosino, Partha Pratim Chakrabarti and Venu Govindaraju. Our Tutorial Co-chairs arranged an excellent set of pre-conference tutorials. We are grateful to all our invited and tutorial speakers.

We would like to thank the host institute, IIT Delhi, for providing all facilities for organizing this conference. We are grateful to Springer and National Centre for Soft Computing Research, ISI Kolkata for the necessary logistics and financial support. The success of the conference is also due to funding received from different government agencies and industrial partners. We are grateful to all of them for their active support. We are grateful to the Organizing Committee for their endeavor in making this conference a success.

The PC Co-chairs would like to especially thank our Publication Chair, Sumantra Dutta Roy, for his excellent contributions toward the publication process. We are also grateful to Dominic Ślęzak for his co-operation and help. Our patron, Surendra Prasad, and members of our Advisory Committee provided the required guidance.

PReMI 2005 and PReMI 2007 were successful conferences. We believe that you will find the proceedings of PReMI 2009 to be a valuable source of reference for your ongoing and future research activities.

Santanu Chaudhury
C.A. Murthy
S. Mitra
P.S. Sastry
S.K. Pal

Organization

Patron

Surendra Prasad	IIT Delhi

General Chair

Sankar Pal	ISI Kolkata

Advisory Committee

Amitava Bagchi	IISER Kolkata
B.L. Deekshatulu	University of Hyderabad
Dwijesh Dutta Majumdar	ISI Kolkata
S.C. DuttaRoy	IIT Delhi
Madan Gopal	IIT Delhi
R.K. Shyamasundar	TIFR
R.M.K. Sinha	IIT Kanpur

Program Co-chairs

Santanu Chaudhury	IIT Delhi
Sushmita Mitra	ISI Kolkata
C.A. Murthy	ISI Kolkata
P.S. Sastry	IISc Bangalore

Program Committee

Sanghamitra Bandyopadhyay	ISI Kolkata
Anupam Basu	IIT Kharagpur
Pushpak Bhattacharyya	IIT Bombay
K.K. Biswas	IIT Delhi
Isabelle Bloch	ENST France
Lorenzo Bruzzone	University of Trento
Bhabatosh Chanda	ISI Kolkata
Niladri Chatterjee	IIT Delhi
Subhasis Chaudhuri	IIT Bombay
Sukhendu Das	IIT Madras
Rajat K De	ISI Kolkata
Andreas Dengel	DFKI
Lipika Dey	TCS
Sumantra Dutta Roy	IIT Delhi

Asish Ghosh	ISI Kolkata
H. Ghosh	TCS
Mark A. Girolami	University of Glasgow
Larry Hall	USF
Gaurav Harit	IIT Kharagpur
C.V. Jawahar	IIIT Hyderabad
Jayadeva	IIT Delhi
Mohan Kankanhalli	NUS
I.N. Kar	IIT Delhi
Ravi Kothari	IBM-IRL
Anjaneyulu Kuchibhotla	HP Labs India
Arun Kumar	IIT Delhi
Krishna Kummamuru	IBM-IRL
Malay Kumar Kundu	ISI Kolkata
Brejesh Lall	IIT Delhi
Wang Lei	NTU
Sriganesh Madhvanath	HP Labs India
A.K. Majumdar	IIT Kharagpur
Dipti Prasad Mukherjee	ISI Kolkata
Jayanta Mukherjee	IIT Kharagpur
Sudipto Mukherjee	IIT Delhi
Hema A. Murthy	IIT Madras
Mike Nachtegael	Ghent University
Nasser Nasrabadi	ARL
Olfa Nasraoui	University of Louisville
Ram Nevatia	USC
N.R. Pal	ISI Kolkata
Swapan K. Parui	ISI Kolkata
Alfredo Petrosino	University of Naples Parthenope
Arun K. Pujari	University of Hyderabad
A.G. Ramakrishnan	IISc Bangalore
Gerald Schaefer	Loughborough University, UK
Srinivasan H. Sengamedu	Yahoo
Rahul Sukthankar	CMU
Jason T.L. Wang	NJIT
Jakub Wroblewski	Infobright

Publication Chair

Sumantra Dutta Roy	IIT Delhi

Tutorial Co-chairs

Jayadeva	ISSA Delhi
Ganesh Ramakrishnan	IBM-IRL Delhi

Organizing Co-chairs

H. Ghosh	TCS
I.N. Kar	IIT Delhi

Organizing Committee

Ramesh Agarwal	JNU
C. Anantaram	TCS
Ashok Chakravarty	MIT
Niladri Chatterjee	IIT Delhi
Sumantra Dutta Roy	IIT Delhi
Poonam Gupta	CDAC NOIDA
S. Indu	DCE
R.S. Jadon	MITS Gwalior
Ashish Khare	TCS
B.M. Mehtre	CMC Ltd., Hyderabad
Sona Minz	JNU
Sukumar Mishra	IIT Delhi
Sumant Mukherjee	ISSA, Delhi
K.R. Murali Mohan	DST
B.K. Panigrahi	IIT Delhi
Ram Ramaswamy	JNU
Sohan Ranjan	GE Research
Geetika Sharma	TCS
R. Shivakumar	DST/NSDI
Ajay Shukla	PMI, NTPC

International Liason and Co-ordination Co-chairs

Simon C.K. Shiu	Hong Kong Polytechnic University, Hong Kong
Dominik Ślęzak	Infobright, Poland

External Reviewers

Tinku Acharya
M. Abulaish
Avinash Achar
Amir Ahmad
Ramakrishnan Angarai G
Sanghamitra Bandyopadhyay
Subhashis Banerjee
Rana Barua
Jayanta Basak
Anupam Basu
Laxmidhar Behera
Narayan Bhamidipati
Rajen Bhatt
R.K.P. Bhatt
Pushpak Bhattacharyya
Arijit Bishnu

S.N. Biswas
K.K. Biswas
Prabir Kumar Biswas
Isabelle Bloch
Lorenzo Bruzzone
Partha Pratim Chakrabarti
Bhabatosh Chanda
Sharat Chandran
Niladri Chatterjee
Subhasis Chaudhuri
Tapan Kumar Chaudhuri
Sukhendu Das
Zhang David
Rajat Kumar De
Lipika Dey
Chitra Dutta
Sumantra Dutta Roy
Utpal Garain
Anil K. Ghosh
Ashish Ghosh
Hiranmay Ghosh
Mark Girolami
Sujit Gujar
Amarnath Gupta
Phalguni Gupta
Larry Hall
Sk. Mirajul Haque
Gaurav Harit
Rakesh Jadon
C.V. Jawahar
Jayadeva
Joby Joseph
Shiv Dutt Joshi
Mohan Kankanhalli
Indra Narayan Kar
Prem Karla
Sunil Kumar Kopparapu
Ravi Kothari
Arun Kumar
Krishna Kummamuru
Malay Kumar Kundu
Brejesh Lall
Srivatsan Laxman
Sriganesh Madhvanath
Anuj Mahajan
Subhamoy Maitra
Santi Prasad Maity
Pradipta Maji
A.K. Majumdar
Sharmila Mande
Naresh Manwani
Mona Mathur
Ujjwal Maulik
Shashi Mehta
Suman K. Mitra
Pabitra Mitra
Atanendu Mondal
Jayanta Mukherjee
Dipti Prasad Mukherjee
M.N. Murthy
Hema Murthy
Mike Nachtegael
Sarif Kumar Naik
Anoop M. Namboodiri
Sukumar Nandi
P.J. Narayanan
Nasser Nasrabadi
J. Saketha Nath
Ram Nevatia
Pinakpani Pal
Sarbani Palit
P.C. Pandey
Bijaya Ketan Panigrahi
Swapan Kumar Parui
Debprakash Patnaik
Amit Patra
Alfredo Petrosino
Arun Pujari
B. Ravindran
Shubhra Sankar Ray
Dipanwita Roy Chowdhury
Sanjay Saha
Sudeshna Sarkar
Palash Sarkar

Gerald Schaefer
Srinivasan H. Sengamedu
Debasis Sengupta
Debapriya Sengupta
Shesha Shah
B. Uma Shankar
Dominik Slezak
Rahul Sukthankar
Susmita Sur-Kolay
Jason T.L. Wang
Jakub Wroblewski

Table of Contents

Pattern Recognition and Machine Learning

Soft Computing and Applications

Bio and Chemo Informatics

Text and Data Mining

Image Analysis

Document Image Processing

Watermarking and Steganography

Biometrics

Image and Video Retrieval

Speech and Audio Processing

Applications

Evolutionary Computing

New Approaches to Design and Control of Time Limited Search Algorithms

Partha Pratim Chakrabarti[1] and Sandip Aine[2]

[1] Dept of CSE, Indian Institute of Technology Kharagpur, 721302, India
ppchak@cse.iitkgp.ernet.in
[2] Mentor Graphics (India) Pvt. Ltd., Noida, UP 201301, India
sandip_aine@mentor.com

Abstract. We talk about two key aspects of the quality-time trade-offs in time limited search based reasoning namely, design of efficient anytime algorithms and formulations for meta-reasoning (or control) to optimize the computational trade-off under various constrained environments. We present the ideas behind novel anytime heuristic search algorithms, both contract and interruptible. We also describe new meta-control strategies that address parameter control along with time deliberation.

1 Introduction

Optimizing the quality-time trade-off while solving optimization problems has been a major area of interest for Artificial Intelligence (AI) researchers. Since many of the optimization problems encountered in practice such as design, planning and scheduling are *NP-hard*, optimal rationality is not to be expected within limited resources (like computational time). In AI literature, this trade-off is addressed in two phases, through *design* of efficient algorithms which provide handles to control their execution, and through intelligent *meta-control* mechanisms which automatically decide such execution strategies under various constrained scenarios. The use of *Anytime algorithms* [1] was proposed as a basic computational framework for handling the quality-time trade-offs. These algorithms work under *any time* limitations producing different quality solutions, and thus, provide opportunities to reason about their time deliberation. Anytime algorithms are classified into two categories namely, interruptible algorithms and contract algorithms. Interruptible anytime algorithms are expected to regularly produce solutions of improved quality and when suddenly terminated, return the best solution produced so far as the result. For a contract algorithm, time is allocated a priori and the algorithm is required to provide a high-quality solution at the end of the time contract. Using these anytime algorithms as the base level computational model, several *meta-reasoning* frameworks have been proposed to control the quality-time trade-offs for hard optimization problems.

This paper addresses two dimensions of the quality-time trade-off paradigm, namely design and meta-control [2]. In the design domain, we concentrate on Anytime Heuristic Search techniques and present the key ideas behind two novel

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 1–6, 2009.

Anytime Heuristic Search algorithms, suitable for contract and interruptible purposes. In the control domain, we introduce the concept of parameter control of anytime algorithms and propose integrated meta-level frameworks capable of taking a unified decision on the parameter configuration chosen along with the deliberation control. We focus on the formulations and strategies required for an integrated control of iterative stochastic optimization algorithms such as Simulated Annealing [3] or Evolutionary Algorithms [4] under a variety of scenarios.

2 Design of Anytime Heuristic Search Techniques

In the past few years, a number of interruptible heuristic search algorithms have been developed. In general, these algorithms follow two basic principles: they use depth guided non-admissible pruning to obtain fast solutions and work in multiple iterations with gradual relaxation of the constraints to produce a stream of gradually improving solutions. These algorithms can be broadly classified into two categories, weighted A* [5,6] approaches and beam based approaches [7,8]. One of the major disadvantages of using the anytime techniques is the lack of models, which can provide the user an apriori estimate about the quality-time trade-off. Therefore, to use these techniques, lengthy simulations are required to find appropriate parameter settings. In the next sub-sections we discuss two novel techniques of heuristic search. The first one is designed to work under node limitation contract (Contract Search) and the second one is an interruptible algorithm (Anytime Window A*).

2.1 Contract Search

We explore the possibility of developing a heuristic search algorithm under a contract. Since time spent can be determined in terms of number of nodes expanded (as node expansion is the heaviest task for a search algorithm), we work under a node expansion contract. The A* algorithm performs a global competition performed among nodes in the open list, select nodes in ascending order of their $f(n)$ values. While this global competition guarantees optimality of solutions produced by the algorithm (when heuristics are admissible), it also ensures that all nodes which come into the open list and have $f(n)$ less than the optimal cost solution are necessarily explored. Our work is based on the hypothesis that for each level of a search space, if we only expand nodes up to the 'optimal-path node', we can attain optimality. To formalize this we introduce the concept of ranking among nodes in the search space. The rank denotes the position of a node (in terms of f-value) among a chosen set of nodes. While A* uses a global ranking scheme (without any restrictions) we introduce a local ranking among the nodes in the same level of a search. Our study of state spaces of some important practical problems reveals that the nodes on the optimal path are usually very competitive among nodes at their same level. We reinforce this observation by obtaining analytical results on search tree models (uniform and non-uniform costs).

Our analysis on best-first search shows that stricter pruning can be obtained using *rank based restrictions*. For Contract Search, we propose to obtain a probabilistic estimation of rank values, and use that information to guide the search. For this, we put forward the concept of Probabilistic Rank Profile (PRP). Probabilistic Rank Profile $P(S|l, k(l))$ is a model which represents the chance of expanding the 'optimal-path' node at a level if a maximum number of $k(l)$ nodes are expanded at that level. These values can be obtained either through profiling or by using search space characteristics to generate the PRP values. For Contract Search, we use the PRP to compute the expansion bounds for each level of the search space depending on the contract specification. We formulate the bound selection (k-selection) problem under a given contract C as follows: For a problem having PRP $P(S|l, k(l))$ we generate a k-cutoff for each level l $(0 \leq l \leq h)$ such that,

$$\textstyle\sum_l k(l) \leq C \ and \ \ P_S(C) \text{ is maximized} \tag{1}$$

With this formulation, we define the k-selection strategy for a particular level as a Markov Decision Process (MDP) which is solved using dynamic programming. The strategy can be computed off-line and then used to guide the actual search under specified node limitations.

Contract Search uses these $k(l)$ values to limit the expansions for a particular level. A naive way of incorporating the restrictions would be to expand the best $k(l)$ nodes at level l. In Contract Search, nodes are expanded in the best first mode across all levels. However, if the number of nodes expanded at a given level l equals the $k(l)$ value, the level is suspended. The search terminates when all levels are suspended or there are no more nodes having $f(n)$ less than the obtained solution. It may be noted that when a given level l is suspended, the suspension is propagated to upper levels to reduce useless expansions. Contract Search has been applied on a number of search problems namely, Traveling Salesperson Problem (TSP), 0/1 Knapsack Problem, 15-Puzzle Problem and Instruction Scheduling Problem and have yielded better quality results as compared to schemes such as ARA* [6] and beam search [8].

2.2 Anytime Window A*

As discussed earlier, the algorithm A* considers each node to be equivalent in terms of information content and performs a global competition among all the partially explored paths. In practice, the heuristic errors are usually distance dependent [9]. Therefore, the nodes lying in the same locality are expected to have comparable errors, where as the error may vary substantially for nodes which are distant from each other. Thus, if the global competition performed by A* is localized within some boundaries, tighter pruning can be obtained. Based on this observation, we present an alternative anytime heuristic search algorithm Anytime Window A* (AWA*), which localizes the global competition performed by A* within a fixed-size window comprising of levels of the search tree/graph.

The AWA* algorithm works in two phases, in the inner loop it uses the Window A* routine with a fixed window size and computes the solution under the current restriction. In the outer loop the window restrictions are gradually relaxed to obtain iteratively improving solutions. The Window A* routine works with a pre-specified window size. For a given window size ω, the expansions are localized in the following way. When the first node of any level (say l) is expanded, all nodes in the open list which are from level $(l - \omega)$ or less will be suspended for this iteration. The window slides in a depth-first manner when deeper nodes are explored. Window A* terminates when the first goal node is expanded or if there is no chance to obtain a better solution (than previously generate). The AWA* algorithm calls the Window A* routine multiple times with gradual relaxation of the window bounds. At the start, window size is set to 0 (depth-first mode). Once the Window A* routine terminates, AWA* checks whether there are any nodes suspended in the previous iteration. If the suspended list is empty, the algorithm is terminated returning the optimal solution. Otherwise, the window size is increased in a pre-decided manner, and Window A* is called again.

We analyze the characteristics of the presented algorithm using a uniform search tree model. This reveals some very interesting properties of the algorithm in terms of heuristic accuracy versus solution quality and related search complexity. The experimental results obtained on TSP and 0/1 Knapsack corroborate the analytical observations, thus, validating the applicability of analysis in real life domains. We get considerable improvement in convergence probabilities as well as intermediate solution qualities over the existing approaches.

3 Integrated Frameworks for Meta-level Control

In the domain of meta-level frameworks, most of the suggested methodologies attempt to solve the problem of time (resource) deliberation only. However, for many algorithms targeted to solve *hard* problems, time allocation is not the only criterion that can influence the solution quality. The progress of these algorithms is strongly affected by some other parameter choices. Thus, an integrated framework for deliberation and control is necessary to effectively handle constrained scenarios. Also, many of the available utility based formulations are for single problem instances. On the other hand, in many real world situations, the requirement is usually to solve single/multiple problem instances under time/quality constraints. In this section, we introduce the concept of parameter control of anytime algorithms under various constrained environments. We mainly concentrate on two specific stochastic techniques namely, simulated annealing (SA) [3] and evolutionary algorithms(EA) [4].

The basic problem for the meta-level controller that we address in this part is to decide on a control parameter setting (along with the time allocation for utility based framework) for anytime algorithms which optimizes the objective function chosen under the specified constraints. For this, the meta-level control framework requires a model to measure the improvement of the solution quality with time for different control settings. Dean & Boddy [10] introduced the term *performance profile* of an algorithm, which for an interruptible algorithm, represents the expected solution quality as a function of the allocated time. This model suggested in [10] was suitable for static reasoning only. Hansen and Zilberstein [11] extended the static profiling concept for dynamic decision making.

We shift our focus from the time adjustment frameworks to a more generic multi-dimensional control space, the quality-time models no longer remain sufficient. We require to model an algorithm's progress not only with time but also with the different parameter settings. We introduce the concept of Dynamic Parameterized Performance Profile (DPPP) The dynamic parameterized performance Profile of an algorithm, $P(S_j|S_i, \Delta t, C)$, is defined as the probability of reaching a solution state S_j if an additional time Δt is given and the control parameter/vector is C, when the current solution state is S_i. Using this we generate the profiles for SA (with temperature as the parameter) and EA (with mutation and crossover rates as parameters).

Using these profiles we generate the optimal strategies for parameter control of an anytime algorithms for the runtime constrained scenario. The optimal selection strategy can be formulated as:

$$EQ_D(S_i, T_{left}) = \max_{\Delta t, C} \begin{cases} \Sigma_j P(S_j|S_i, T_{left}, C) * Q(S_j) \text{ If } \Delta t = T_{left} \\ \Sigma_j P(S_j|S_i, \Delta t, C) * EQ_D(S_j, T_{left} - \Delta t) - M_p \\ \text{Otherwise} \end{cases} \tag{2}$$

The probabilities are obtained from the DPPP, S_i is the current solution state, T_{left} is the time left before reaching the deadline and $Q(S_j)$ is the quality component of the solution state S_j. M_p is the quality penalty for each monitoring step.

The strategies obtained can also be extended for a *utility* based system. Utility is a function of the solution quality and the computation effort (time) which represents the intended trade-odd between quality and time. For a utility based system the framework takes a combined decision on the time allocation and control parameters following the same dynamic programming based approach. In many real-world cases, it may be required to solve a number of problem instances under given run-time limitations. The individual problem instances may depend on each other or may be independent. In our formulation with multiple independent problems, the number of problems to be solved and the total time allocated (i.e., the global deadline) are known a priori. With this formulation, the objective of the meta-controller is to maximize the overall quality within the global deadline. Considering dynamic strategies for multiple problems, we have two options namely, preemptive and non-preemptive strategies. In the non-preemptive case, the problem instances are solved in a particular order and once a problem is evicted from execution queue it is not taken back. In the case of preemptive strategy, any problem instance can be chosen from the problem set at run-time, irrespective of whether it has been stopped earlier or not and the parameter configuration and the time allocation are decided accordingly. Similar control strategies can be generated for the preemptive case. For multiple dependent problems, we consider a pipeline like structure to model an optimization flow, where output obtained from a given stage is fed to the next stage as input.

In this formulation, the controller has an n-stage pipeline where each stage is an optimization algorithm, the total allocated time (or the deadline) D is known a priori. The controller decides on the time allocation and parameter configuration for each stage, such that the expected output quality is maximized. To generate the meta-level strategies for the pipeline structure we modify our profiling technique. Up to this stage, we have generated independent profiles for the individual problems. However, when we consider the optimization process as a multi-stage pipeline, the performance of a given stage will be dependent on the solution obtained from its previous stage. Thus, we condition the profiles by including of information about the input.

Experiments performed (using both SA and EA) on classical optimization problems like TSP and 0/1 Knapsack and other real-life problems in the domain of CAD for VLSI optimizations, have demonstrated the efficacy of the proposed frameworks.

References

1. Boddy, M., Dean, T.: Deliberation scheduling for problem solving in time-constrained environments. Artificial Intelligence 67, 245–285 (1994)
2. Aine, S.: New Approaches to Design and Control of Anytime Algorithmd. PhD thesis, Indian Institute of Technology Kharagpur, Department of Computer Science and Engineering, IIT Kharagpur 721302 India (2008)
3. van Laarhoven, P., Aarts, E.: Simulated Annealing: Theory and Applications. Kluwer, Dordrecht (1992)
4. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston (1989)
5. Pohl, I.: Heuristic search viewed as path finding in a graph. Artif. Intell. 1(3), 193–204 (1970)
6. Likhachev, M., Gordon, G.J., Thrun, S.: Ara*: Anytime A* with provable bounds on sub-optimality. In: Advances in Neural Information Processing Systems, vol. 16, MIT Press, Cambridge (2004)
7. Zhang, W.: Complete anytime beam search. In: Proceedings of 14th National Conference of Artificial Intelligence AAAI 1998, pp. 425–430. AAAI Press, Menlo Park (1998)
8. Zhou, R., Hansen, E.A.: Beam-stack search: Integrating backtracking with beam search. In: Proceedings of the 15th International Conference on Automated Planning and Scheduling (ICAPS 2005), Monterey, CA, pp. 90–98 (2005)
9. Pearl, J.: Heuristics: intelligent search strategies for computer problem solving. Addison-Wesley Longman Publishing Co., Inc., Boston (1984)
10. Dean, T., Boddy, M.: An analysis of time-dependent planning. In: Proceedings of 6th National Conference on Artificial Intelligence (AAAI 1988), St. Paul, MN, pp. 49–54. AAAI Press, Menlo Park (1988)
11. Hansen, E.A., Zilberstein, S.: Monitoring and control of anytime algorithms: A dynamic programming approach. Artificial Intelligence 126(1-2), 139–157 (2001)

Feature Selection Using Non Linear Feature Relation Index

Namita Jain and C.A. Murthy

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
namita.saket@gmail.com,
murthy@isical.ac.in
http://www.isical.ac.in/~murthy/

Abstract. In this paper we propose a dependence measure for a pair of features. This measure aims at identifying redundant features where the relationship between the features is characterized by higher degree polynomials. An algorithm is also proposed to make effective use of this dependence measure for the feature selection. Neither the calculation of dependence measure, nor the algorithm need the class values of the observations. So they can be used for clustering as well as classification.

1 Introduction

The problem of reducing the number of features is also known as *dimensionality reduction*. Mathematically the problem of dimensionality reduction can be stated as follows. Given a p-dimensional random variable $x = (x_1, x_2, \cdots, x_p)$, find a lower dimensional representation of data $s = (s_1, \cdots, s_k)$ with $k \leq p$, that captures the content of the original feature set according to some criterion [1]. The two approaches used for dimensionality reduction are feature selection and feature extraction. Feature extraction aims at building a transformed feature set, such that the new feature set has some advantages over the original feature set. Principal Component Analysis is an example of a feature extraction technique. Feature selection aims at finding a representative feature set from the original set. Here the new representative set is the subset of original feature set. Irrelevant features are the ones which can be removed without making a negative impact on the learning task. Redundant features are features which become irrelevant due to presence of other features.

1.1 Feature Selection

Two important steps of feature selection are feature search and evaluating the feature subset according to some criterion [2][3][4].

Feature Search. Several search algorithms are available for feature selection in literature. None of them guarantees to provide the optimal feature subset for any choice of criterion function. Some of the search methods are branch

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 7–12, 2009.

and bound (performs poorly on non monotonic criterion functions), Sequential Forward search, Sequential Backward search, Sequential forward floating search, Sequential backward floating search [5]. Some algorithms use randomized search [1].

Evaluation criteria. Once a subset is generated, an evaluation criterion is used to assign a goodness level to a given subset of features. The criterion function can take different aspects into consideration. These include (i) Distance measure, (ii) Information measure, (iii) Dependence measure, (iv) Consistency measure [4].

Dependence Measure. Even if two features are individually relevant, the presence of one feature may make the other feature dispensable. Such features can be identified using a dependence measure [4]. Identifying such features is the main purpose of the criterion function defined in this paper.

Filter approach to feature selection. A feature selection method may follow either a filter approach or a wrapper approach[3]. The wrapper approach uses a learning algorithm that will ultimately be applied, as its guide, for selecting the features. The filter approach however uses the data alone to select the features. For an algorithm depending on filter approach, the evaluation of the feature set is not done in reference with any particular classifier. If the criterion function, used to determine the features to be added or removed to the subset, considers the effect of the change on entire subset rather then evaluating features separately, the whole set is evaluated at each step.

2 Proposed Feature Relation Index

In this section we propose a feature selection index which aims at identifying statistical dependence between features. A few existing dependence measures are also discussed here.

Correlation coefficient($\mid \rho \mid$). A simple measure of relationship between two variables x and y is absolute value of correlation coefficient [6]. The correlation coefficient is given by given by $\rho(x, y) = \sum \frac{(x_i - \overline{x})(y_i - \overline{y})}{n\sqrt{var(x)var(y)}}$ where $\overline{x}$ and $\overline{y}$ are mean values of the two variables. The magnitude of $\rho(x, y)$ determines the strength of the relationship.$\mid \rho(x, y) \mid$ is invariant to translation and scaling. $\mid \rho(x, y) \mid$ is not rotation invariant [6].

Least Square Regression Error (e). Another measure which has been used as a measure of dependence between variables is Least Square error [6]. A straight line $y = a + bx$ is fit to a set of points $(x_i, y_i), i = 1, \cdots, n$ in the x, y plane, such that it minimizes the mean square error given by $e(x, y) = \frac{1}{n} \sum (e(x, y)_i)^2$, where $e(x, y)_i = y_i - a - bx_i$. The coefficients are given by $a = \overline{y} - b\overline{x}$ and $b = \frac{cov(x,y)}{var(x)}$. The mean square error is given by $e(x, y) = var(y)(1 - \rho(x, y)^2)$. Mean square error is the residual variance unexplained by the linear model.

The value e denoting least square error (linear) will be 0 only if a perfect linear relationship exists between x and y. The measure is invariant to translation and sensitive to scaling. It is also non-symmetric. Having a non symmetric measure allows us to choose the variable which is more likely to have higher information content.

Maximal Information Compression Index (MICI)($\boldsymbol{\lambda_2}$)[7]. The linear dependence between two random variables x, y can be judged from smallest eigenvalue λ_2 of their covariance matrix $\sum$. The value of MICI is given by $2\lambda_2(x,y) = var(x) + var(y)\sqrt{(var(x) + var(y))^2 - 4var(x)var(y)(1 - \rho(x,y)^2)}$ The measure λ_2 can be seen as the eigenvalue for the direction normal to the Principal Component Axis [17]. Therefore λ_2 indicates the minimum amount of error incurred on projecting the data on a single dimension.

The value of λ_2 will be 0 if and only if a perfect linear relationship exists between x and y. The measure is invariant to translation. It is sensitive to scaling. The measure is symmetric. The measure is invariant to rotation. This method will not be able to find non-linear realtionship between the variables.

2.1 Non-linear Feature Relation Index

Here, we try to approximate one variable using a polynomial of another variable. This is done in such a way, that error function is minimized. We try to predict the value of $\hat{y}$ as a polynomial function of x

$$\hat{y} = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

The error function which is minimized is given by

$$e_{yx} = \sum (\hat{y} - y)^2$$

Thus, the value of e_{yx} is minimized to find the coefficients of the polynomial used to model the relationship between the variables. we obtain (k+1) equations for (k+1) unknowns.

For first degree polynomial this error is same as the least square error. Here the use of higher degree polynomilas allow us to discover relationships described by polynomial curves. Just as the error in that case was proportional to variance the error here is proportional to higher degree moments of the variable y.

The value of proposed measure e_{yx} will be 0 only if y can be represented perfectly as a polynomial function of x. The measure is invariant to translation. It is sensitive to scaling. The error function obtained is proportional to higher degree central moments of the variable y. Since, central moment is a good indicator of information content, this allows us to choose a variable with more information content. The measure is non symmetric. The measure is not invariant to rotation. The measure is able to find non-linear relationship between the variables.

2.2 Selecting a Value for k

In the given method we try to get a better estimation of y as a higher degree polynomial of x. This will lead to a smaller error value. It may be noted that increasing the power of the polynomial will not adversely effect the performance of the algorithm even if y can be represented as a lower degree polynomial of x. Let us consider the case where y can be represented perfectly as a polynomial (of dgree k) of x, leading to zero error. Let the representation be $y = \beta_{k,0} + \beta_{k,1}x + \cdots + \beta_{k,k}x^k$. Now we try to represent y as a k+1 degree polynomial of x as $y = \beta_{k+1,0} + \beta_{k+1,1}x + \cdots + \beta_{k+1,k+1}x^{k+1}$. We find that the value of the coefficients we obtain are given by

$$\beta_{k+1,i} = \beta_{k,i} for 1 <= i <= k, \beta_{k+1,k+1} = 0$$

3 Proposed Feature Selection Algorithm

First step of this feature selection method is to calculate the value of Non linear feature relation index e_m for each pair of features. The objective is to iteratively select features for which the elimination of dependent variables will lead to minimum loss. We select a parameter r which indicates the maximum number of features which can be removed at each step. Select number of features to be removed in the next step as $n = min(r, c)$ where c represents the number of features left to be considered. For each feature the n pairs which have minimum value of e_m are selected and summation of e_m is calculated over these pairs. The feature for which the calculated sum is minimum is selected and the corresponding dependent features are eliminated at each step. These steps are repeated till all the features are included either in set of selected features or in the set of rejected features.

3.1 Algorithm

Let F be the original set of features, S be the set of selected features and E be the set of eliminated features. Choose a value k giving the degree of expectation function to be used to Regression. Choose r giving the maximum number of features to be eliminated at each step. Let ϵ be the maximum error caused by eliminating a feature.

```
1. For each feature $i \in F$
   (a) For each feature $j \in F$ $M(i, j) \leftarrow e_m(i, j)$
   (b) $\epsilon = \infty$
   (c) $r = min(r, card(F - (S \cup E)))$
   (d) Find the set of features $D_i = [j_1, \cdots, j_r]$ such that

          $\forall(a, b)\{(a \in D_i, b \in F - D_i) \Rightarrow M(i, a) \leq M(i, b)\}$

   (e) if $M(i, r) > \epsilon$ then
         i. $r = r - 1$
        ii. goto 1.(c)
   (f) $S(i) = \sum M(i, j_k)$,where $k = 1, \cdots, r$
```

2. Find the feature i such that

$$\forall(i,j)\{i \in F-(S \cup E) \wedge j \in F-(S \cup E \cup \{i\}) \Rightarrow S(i) < S(j)\}$$

3. if $card(S) = 0$ then $\epsilon = M(i,r)$, where i is the feature found in step 2.
4. $S = S \cup \{i\}$, where i is the feature found in step 2.
5. $E = E \cup D_i$, where D_i is the set found in step 1.(d) corresponding to the feature i selected in step 3.
6. For each $d \in D_i$, where D_i is same as in step 4 $M(d,t) = \infty$ and $M(t,d) = \infty$ for all $t \in F$
7. If $card(F-(S \cup E)) > 0$ goto step 1.(c)

4 Results

The non Linear Feature relation index has been used by the algorithm given in previous section. We have tested this algorithm on several real life datasets [8] such as Iris, Cover Type (10 numeric attributes are used.), Ionosphere, Waveform, Spambase, Isolet, and Multiple features. A polynomial of degree 3 has been used for curve fitting while finding the non Linear Feature relation index.

As per three dimensional framework for classification given by Huan Liu and Lei Yu, the proposed method falls in the category of dependence measure, using a sequential search strategy, useful for clustering task[3] along with MICI. The performance of the proposed algorithm has been compared to the results given by MICI [7][9]. The proposed method gives better performance for all the datasets except Isolet and Spambase. It can be seen that the proposed method results in a slightly increased error rate for Spambase and Isolet datasets. In Isolet dataset the data is given in a normalized form. In Spambase most of the features are scaled to be represented as percentage. Scaling of data might be a reason for reduced performance of the proposed algorithm in the two cases as the Non Linear Feature Relation Index is sensitive to scaling. The difference in performance has been found to be statistically significant in all the datasets using the Welch test for Behrens-Fisher problem.

Table 1. Classification performance of proposed feature selction algorithm as compared to MICI, where D denotes the number of features in original set and d_{MICI} and d denote the number of features in the reduced set obtained by MICI and the proposed method

Dataset	D	d_{MICI}	MICI	d	Proposed
Iris	4	2	97.33	2	97.47
Cover Type	10	5	63.55	5	66.56
Ionosphere	32	16	65.92	11	80.57
Waveform	40	20	63.01	20	87.94
Spambase	57	29	88.19	29	87.47
Isolet	610	310	95.01	307	94.65
Mfeat	649	325	78.34	324	94.11

5 Conclusion

In this paper we proposed a dependence measure based on polynomial regression between two features. Also, an algorithm is proposed which effectively uses this measure to identify the redundant features from a given set of features. Since the calculation of dependence measure and the algorithm are independent of the class value of the observations, the proposed method is suitable for clustering tasks along with classification. The algorithm is used to obtain reduced feature sets for real life datasets. The reduced feature sets are then used for classification task. A good classification performance here indicates that redundant features have been discarded from the original feature sets.

References

1. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman Hall/CRC
2. Dash, M., Liu, H.: Feature Selection for Clustering. In: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, April 18-20, 2000, pp. 110–121 (2000)
3. Liu, H., Yu, L.: Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17(4), 491–502 (2005)
4. Dash, M., Liu, H.: Feature Selection for Classification. Intelligent Data Analysis 1(3), 131–156 (1997)
5. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs (1982)
6. Hoel, P.G., Port, S.C., Stone, C.J.: Introduction to Statistical Theory. Houghton Mifflin, New York (1971)
7. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
8. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html
9. Pal, S.K., Mitra, P.: Patter Recognition Algorithms for Data Mining. ChapMan and Hall/CRC

Classification of Multi-variate Varying Length Time Series Using Descriptive Statistical Features

S. Chandrakala and C. Chandra Sekhar

Department of Computer Science and Engineering,
Indian Institute of Technology Madras, India
{sckala,chandra}@cse.iitm.ac.in

Abstract. Classification of multi-variate time series data of varying length finds applications in various domains of science and technology. There are two paradigms for modeling multi-variate varying length time series, namely, modeling the sequences of feature vectors and modeling the sets of feature vectors in the time series. In tasks such as text independent speaker recognition, audio clip classification and speech emotion recognition, modeling temporal dynamics is not critical and there may not be any underlying constraint in the time series. Gaussian mixture models (GMM) are commonly used for these tasks. In this paper, we propose a method based on descriptive statistical features for multi-variate varying length time series classification. The proposed method reduces the dimensionality of representation significantly and is less sensitive to missing samples. The proposed method is applied on speech emotion recognition and audio clip classification. The performance is compared with that of the GMMs based approaches that use maximum likelihood method and variational Bayes method for parameter estimation, and two approaches that combine GMMs and SVMs, namely, score vector based approach and segment modeling based approach. The proposed method is shown to give a better performance compared to all other methods.

Keywords: Time series classification, Descriptive statistical features, Speech emotion recognition, Audio clip classification.

1 Introduction

Classification of multivariate, varying length time series data is necessary in widely varying domains that include data such as speech, music, video, bioinformatics, bio-medicine and tasks such as speech recognition, handwritten character recognition, signature verification, speaker recognition, audio classification and speech emotion recognition.Time series data may be of discrete or real valued, uniformly or non-uniformly sampled, univariate or multi-variate and of equal or unequal length. The main issues in time series classification methods are related to (a) time series representation, (b) similarity measure and (c) choice of classifier. Approaches to time series classification focus on finding the relevant features for time series representation, or on the similarity metric between a pair of time series, and/or on modeling the time series data.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 13–18, 2009.

There are two paradigms for modeling a varying length time series, namely, modeling it as a sequence of feature vectors and modeling it as a set of feature vectors. Tasks such as speech recognition need modeling both temporal dynamics and correlations among the features in the time series. In these kind of tasks, each example belonging to a class has a fixed number of acoustic events. Hidden Markov models (HMMs) are the commonly used models for speech recognition [1]. In tasks such as speaker recognition, audio or music classification and speech emotion recognition [2] , the duration of sequences is large, the local temporal dynamics is not critical and there may not be any underlying constraint in the time series. Each example of a class has a different number of acoustic events. Gaussian mixture models (GMMs) are commonly used for these tasks.

Generative [1,2] and discriminative approaches [3] are two main approaches to designing classifiers. Generative approaches focus on estimating the density of the data. These models are not suitable for classifying the data of confusable classes since a model is built for each class using the data belonging to that class only. Discriminative classifiers such as support vector machines (SVMs) [3] focus on modeling the decision boundaries between classes and is shown to be effective for static data classification of confusable classes. However, these models require the data to be represented as a fixed dimensional feature vector.

The motivation for the proposed work is to make use of the advantage of discriminative classifiers such as SVM for varying length time series classification tasks that involve modeling a time series as a set of vectors. In this work, we propose a method based on descriptive statistical features for multi-variate, varying length time series classification. First, local domain-specific features are extracted from each window or short time frame of a time series signal. The sequence of feature vectors is then considered as a combination of several univariate time series. A set of descriptive statistical features such as mean, variance, skewness and kurtosis are extracted from each univariate time series that forms a fixed dimensional representation for the varying length time series. The proposed representation converts the difficult problem of classification of multivariate, varying length time series into a problem of classification of static points. These static points are then classified using the SVMs. Some time series classification algorithms may fail for time series with missing samples. The proposed method reduces the dimensionality of the time series significantly and is less sensitive to the missing samples.

The rest of the paper is organized as follows: Section 2 presents a review of methods for varying length time series classification. The proposed method is presented in Section 3. Section 4 presents the results of the studies carried out on audio clip classification and speech emotion recognition.

2 Approaches to Classification of Multivariate, Varying Length Time Series Data

The two paradigms for modeling the varying length time series are modeling the sequences of vectors and modeling the sets of vectors. Approaches to the sequence modeling of multivariate, varying length time series data can be grouped into two categories depending on the method of handling the varying length patterns. Hidden Markov model based classifiers that can handle varying length time series without changing the length or structure of the time series, form the first category of approaches [1,2]. In the second category of approaches, each time series is represented as a fixed dimensional feature vector by converting a varying length sequence of feature vectors to a fixed length pattern (a static point) so that discriminative classifiers such as SVMs can be used for classification [4].

We proposed a hybrid framework [4] for modeling sets of vectors that first uses a generative model based method to represent a varying length sequence of feature vectors as a fixed length pattern and then uses a discriminative model for classification. We proposed two approaches namely, score vector based approach and segment modeling based approach under the hybrid framework using GMM and SVM. In the score vector based approach, each time series in the training data set is modeled by a GMM. The log-likelihood of a time series for a given GMM model is used as a similarity score that forms a feature for that time series. The similarity based paradigm recently introduced for classification tasks is shown to be effective. A score vector is formed by applying the time series to all the models. Likewise, a test time series is also represented as a score vector. An SVM based classifier is then used for time series classification considering each of the score vectors as a fixed length pattern.

In score vector based representation, temporal dynamics in the time series is not modeled and the dimension of score vector depends on cardinality of training data set. In tasks such as speech emotion recognition, though local temporal dynamics is not critical, some kind of sequence information is present at gross level in the time series. To address these issues, we proposed a segment modeling based approach. In this approach, temporal ordering of segments in a time series is maintained to capture the sequence information at gross level. The parameters of a statistical model of a segment are used as features for that segment. A time series is split into a fixed number of segments. Each segment is modeled by a multivariate Gaussian model or a GMM. The model parameters of segments are concatenated in the order of the segments to form a fixed length pattern.

3 Descriptive Statistical Features Based Approach to Time Series Classification

Let a multivariate time series be denoted by $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, ..., \mathbf{x}_N\}$, where $\mathbf{x}_j$ is a d-dimensional feature vector and N is the length of the time series. The assumption here is that the time series classification task involves modeling sets of vectors. In this proposed method, we consider a multivariate time series, $\mathbf{X}$ as a combination of d univariate time series. The i^{th} univariate time series of $\mathbf{X}$ is denoted by $\mathbf{x}_i = \{x_{1i}, x_{2i}, ..., x_{ni}, ..., x_{Ni}\}$ where x_{ni} is the i_{th} element of feature vector $\mathbf{x}_n$. We use a set of descriptive statistical features such as mean, variance, skewness and kurtosis for each of the univariate time series to describe its nature. Variance captures the degree of the deviation from the mean. Variance of a real valued univariate time series data is the second central moment. The moments about its mean are called central moments. Normalised third central moment is called skewness. It is a measure of asymmetry of a distribution. Skewness coefficient for a univariate time series Y is

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \frac{(y_i - \mu)^3}{\sigma^3} \tag{1}$$

where μ is the mean, σ is the standard deviation and N is the number of samples. The left skewed distribution denotes the negative skewness and the right skewed distribution denotes positive skewness. Normalised fourth central moment is called kurtosis. It is a measure of peakness of a distribution. Higher kurtosis means more of the variance is due to infrequent extreme deviations. The kurtosis for a univariate time series Y is

$$\mathbf{K} = \frac{1}{N}\sum_{i=1}^{N} \frac{(y_i - \mu)^4}{\sigma^4} \tag{2}$$

A set of descriptive features extracted from each univariate time series are then concatenated to form a fixed dimensional representation for the varying length time series. The block diagram of the proposed approach is given in Figure 1. The proposed method reduces the dimensionality of the time series significantly and is less sensitive to missing samples.

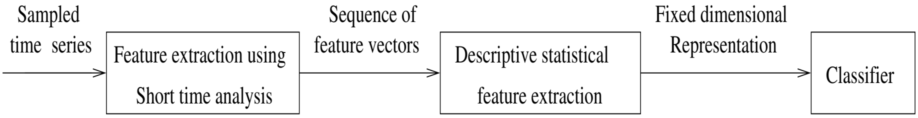


Fig. 1. Block diagram of descriptive statistical features based approach

4 Studies on Speech Emotion Recognition and Audio Clip Classification

Speech emotion recognition and audio clip classification tasks involve modeling the sets of feature vectors. Berlin emotional speech database [5] is used in our studies. A total of 494 utterances were divided among seven emotional classes: Neutral, Anger, Fear, Joy, Sadness, Disgust and Boredom. The duration of the utterances varies from one to two seconds. 80% of the utterances were used for training and the remaining for testing. A frame size of 20ms and a shift of 10 ms are used for feature extraction. The Mel frequency cepstral coefficient (MFCC) vector representing a given frame is a 39-dimensional vector, where the first 12 components are Mel frequency components and the 13th component is log energy. Remaining 26 components are delta and acceleration coefficients that capture the dynamics of the audio signal. The effectiveness of short time MFCC features in speech emotion recognition is shown in [6]. We also study classification of audio clips of TV programs in English and Hindi belonging to the following five categories: Advertisement(Advt), Cartoon, Cricket, Football and News. Audio data is sampled at 22050 Hz. Duration of the audio clips varies from 15 to 20 seconds. A total of 987 clips were divided among five categories. 39-dimensional MFCC feature vector represents each frame of a clip. Four descriptive statistical features, namely, mean, variance, skewness and kurtosis are extracted from each of the 39 univariate time series to form a 156 dimensional vector for a time series. SVM based classifier is then used for classification. The proposed method is compared with four methods evaluated on two data sets. Table 1 shows the performance of proposed method and methods used for comparison.

Table 1. Comparison of classification accuracy (in %) of descriptive statistical features based approach with GMM, VBGMM, score vector based approach and segment modeling based approach on Berlin Emotional database and audio clip database

Classifier	Input to the classifier	Accuracy	
		Emotion Data	Audio Data
GMM	Set of MFCC vectors	64.73	90.83
VBGMM	Set of MFCC vectors	65.71	90.83
SVM	Score vector	70.48	94.80
SVM	Segment Model parameters	72.38	86.85
SVM	Descriptive statistical features	79.05	96.33

First method is the GMM classifier with maximum likelihood method for parameter estimation. In the second method, variational Bayesian approach is employed for parameter estimation in GMM (VBGMM). Since prior distributions are assumed for mixture weights, means and precision matrices instead of using point estimates, VBGMM classifier performs slightly better than GMM classifier in case of emotion data. In case of GMM score vector based representation, effective discriminative ability of similarity space helps in achieving a much better performance than GMM and VBGMM methods. The dimension of the score vector based representation depends on the cardinality of training data set. In the segment modeling based approach that uses a single Gaussian with full covariance parameters, temporal dynamics of the segments in a time series is maintained to some extent and correlations among the features within a segment are also modeled. Hence, this method performs better than the score vector based approach in case of speech emotion recognition task which involves modeling sequence of subsets of vectors. Best performance is obtained with 5 segments for emotion data and 12 segments for audio data. Segmentation of varying length time series data is a critical step in this approach. In the descriptive statistical features based method, descriptive statistical features extracted from each univariate time series effectively describe the distribution. The proposed method outperforms all other methods used for comparison for both data sets. In addition, this method is simple with less computation than all other methods. Confusion matrices for the proposed method for both data sets are given in Table 2 and Table 3.

Table 2. Confusion Matrix for the proposed method on Berlin Emotion data

Classified	Correct Emotion Class						
Class	F	D	H	B	N	S	A
(F)ear	**100**	42.86	18.75	0	4.76	0	0
(D)isgust	0	**14.29**	0	0	0	0	0
(H)appy	0	14.29	**68.75**	0	9.52	0	3.85
(B)oredom	0	14.29	6.25	**100**	4.76	30.77	0
(N)eutral	0	14.29	0	0	**80.95**	0	7.69
(S)adness	0	0	0	0	0	**69.23**	0
(A)nger	0	0	6.25	0	9.5	0	**88.46**

Table 3. Confusion Matrix for the proposed method on audio data

Classified	Correct Audio Class				
Class	Advt	Cart	Cric	Foot	News
Advt	**89.83**	2.78	0	0	1.47
(Cart)oon	5.08	**95.83**	0	0	0
(Cric)ket	0	1.39	**98.59**	0	1.47
(Foot)ball	0	0	0	**100**	0
News	5.08	0	1.41	0	**97.06**

5 Conclusion

Classification of time series data of varying length is important in various domains. Depending upon the type and nature of time series, different methods have been used to represent the time series, compute similarity between two time series and design machine learning algorithms. In this paper, a method based on descriptive statistical features for multi-variate, varying length time series classification has been proposed. The proposed method converts a difficult problem of classification of multivariate, varying length time series into a problem of classification of static patterns. It reduces the dimensionality of the data significantly, less sensitive to missing samples and involves much less computation. The proposed approach is applied on speech emotion recognition and audio clip classification. It outperforms GMM, VBGMM, score vector and segment modeling based approaches. The proposed method for representation of time series data can be applied to any time series classification, clustering, matching and retrieval tasks that involve modeling sets of vectors.

References

1. Rabiner, L., Huang, B.-H.: Fundamentals of speech recognition. Prentice Hall, NewYork (1993)
2. Mishra, H.K., Sekhar, C.C.: Variational Gaussian mixture models for speech emotion recognition. In: International Conference on Advances in Pattern Recognition, Kolkata, India (February 2009)
3. Vapnik, V.: Statistical learning Theory. Wiley-Interscience, New York (1998)
4. Chandrakala, S., Sekhar, C.C.: Combination of generative models and SVM based classifier for speech emotion recognition. In: Proc. Int. Joint Conf. Neural Networks, Atlanta, Georgia (June 2009)
5. Burkhardt, F., Paeschke, A., Rolfes, M., Weiss, W.S.B.: A database of German emotional speech. In: Interspeech, Lisbon, Portugal, pp. 1517–1520 (2005)
6. Sato, N., Obuchi, Y.: Emotion recognition using Mel-frequency cepstral coefficients. Journal of Natural Language Processing 14(4), 83–96 (2007)

2D-LPI: Two-Dimensional Locality Preserving Indexing

S. Manjunath[1], D.S. Guru[1], M.G. Suraj[1], and R. Dinesh[2]

[1] Department of Studies in Computer Science, University of Mysore, Mysore, India
[2] Honeywell Technology Solutions, Bengaluru, India

Abstract. In this paper, we present a new model called two-dimensional locality preserving indexing (2D-LPI) for image recognition. The proposed model gives a new dimension to the conventional locality preserving indexing (LPI). Unlike the conventional method the proposed method can be applied directly on images in 2D plane. In order to corroborate the efficacy of the proposed method extensive experimentation has been carried out on various domains such as video summarization, face recognition and fingerspelling recognition. In video summarization we comapre the proposed method only with 2D-LPP which was recently used for video summarization. In face recognition and fingerspelling recognition we compare the proposed method with the conventional LPI and also with the existing two-dimensional subspace methods viz., 2D-PCA, 2D-FLD and 2D-LPP.

1 Introduction

Latent semantic indexing (LSI) was proposed in [1] to cluster text documents. LSI finds the best subspace which approximates the original document space onto a smaller subspace by minimizing the global reconstruction error. Initially LSI was used for text clustering / classification and later it was explored to index audio documents [2], image retrieval [3] and video content modeling [4]. In 1999, Kurimo [2] developed a method to enhance the indexing of spoken documents with the help of LSI and self organizing map. Zho and Grosky [3] extracted global color histogram from images and latent semantic analysis was performed to learn those extracted features to develop an image retrieval system. In [4], Souvannavong et al., used latent semantic indexing for modeling video content with Gaussian mixture model. Although LSI finds an appropriate subspace to obtain lower dimensional representation of original space, it is found that LSI seeks to uncover the most representative features rather than the most discriminative features for representation [5].

Hence, Locality Preserving Indexing (LPI) was developed for document representation [5] on the basis of Locality Preserving Projection [6] which preserves local structure of the given high dimensional space in lower dimension using spectral clustering. Later, it was modified by introducing an additional step of singular value decomposition in [7] to original LPI method [5] in order to ensure that the transformed matrix is of full order. The modified LPI [7] tries to ascertain both

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 19–24, 2009.

geometric and discriminating structure of the document space by discovering the local geometrical structures [7]. In order to project images onto the LPI subspace, images need to be transformed into one dimensional vector thereby requiring more memory and high computational time. However, a number of works are reported in the literature where vector based projections techniques are successfully extended to image based projection techniques [8][9][10]. The image based projection techniques project images directly onto lower dimensional space rather than transforming them into one dimensional vector [10], thereby working directly on matrices (images) and hence requiring less memory and less computational time. Motivated by the works of [8][9][10], in this paper we extend the conventional LPI [7] to two dimensional LPI, which can be applied directly on images to project them onto lower dimensions by preserving both representative as well as discriminative features of images. From the experimentation it is found that 2D-LPI is more effective as well as efficient when compared to the conventional LPI and also further study resulted that proposed 2D-LPI is better than the existing 2D-LPP [10], 2D-PCA [8] and 2D-FLD [9].

2 Two-Dimensional Locality Preserving Indexing (2D-LPI): Proposed Model

Consider a set of N images $I_1, I_2, I_3, \ldots, I_N$ each of $(m \times n)$ size. We propose a method that maps the original $(m \times n)$ dimensional space onto $(m \times d)$ dimension with $d \ll n$ by projecting each image I_i onto a subspace using W, a $(n \times d)$ dimensional transformation matrix as $Y = IW$ where Y is $(m \times d)$ dimensional projected features of image I.

2.1 The Algorithm

The proposed algorithm is as follows

1. **Construct the graph adjacency matrix:** Let G denote a graph with images as nodes. Image I_i is connected to image I_j if and only if I_j is one among the k-nearest neighbors of image I_i.
2. **Choose the weight:** Let S be the similarity weight matrix given by

$$S_{ij} = \begin{cases} \frac{\exp\{-\|I_i - I_j\|\}}{t} & If\, image\, I_i\, isconnected\, to\, I_j\, of\, G \\ 0 & Otherwise \end{cases} \qquad (1)$$

where t is a suitable constant.

The weight matrix S of the graph G gives the local structure of the image space. A diagonal matrix D whose entries are column (or row) sum of weight matrix S is computed as $D_{ii} = \sum_i S_{ji}$ and the Lapalacian matrix L is computed as $L = D - S$.
3. **Singular Value Decomposition:** Compute the weighted mean matrix of all the images, $\bar{I} = \frac{1}{N}(\sum_{i=1}^{N} I_i D_{ii})$. Decompose the obtained weighted mean matrix using singular value decomposition (SVD) to obtain matrices U and

V which contain left singular and right singular vectors as their columns respectively along with the diagonal matrix σ.

$$\hat{I} = U\sigma V^T \text{ where } \hat{I} = \left[\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_N\right]$$

The right singular matrix V is considered as transformation matrix of SVD and it is denoted by W_{SVD} i.e, $W_{SVD} = V$.

4. **Two-dimensional LPI (2D-LPI) Projection:** Compute the eigen vectors and eigen values by solving the generalized eigen problem as shown in (2), where W_{2DLPI} and λ are respectively the eigen vectors and the eigen value.

$$A^T\left(L \otimes I_m\right) AW_{2DLPI} = \lambda A^T\left(D \otimes I_m\right) AW_{2DLPI} \tag{2}$$

Here, A is an $(mN \times n)$ matrix generated by arranging all the images I in a column, $A = \left[A_1^T, A_2^T, A_3^T, \ldots, A_N^T\right]^T$. Operator $\otimes$ denotes Kronecker product and $I_m$represents an identity matrix of order m.

Let $W_{2DLPI} = [w_1, w_2, w_3, \ldots, w_d]$ and w_i $i = 1, 2, 3, \ldots, d$ be the first d unitary orthogonal solution vectors of (2) corresponding to the d smallest eigen values in the order of $0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_d$. These eigen values are nonnegative because both $A^T\left(L \otimes I_m\right) A$ and $A^T\left(D \otimes I_m\right) AW$ are symmetric $(n \times n)$ matrices and positive semi-definite. Thus W_{2DLPI} constitutes the $(n \times d)$ projection matrix, and the projection is done as follows.

$$Y_i = I_i W, \;\; i = 1, 2, 3, \ldots, n \tag{3}$$

Where $W = W_{SVD}W_{2DLPI}$ and $Y_i = \left[Y_i^{(1)}, Y_i^{(2)}, Y_i^{(3)}, \ldots, Y_i^{(d)}\right]$ is an $(m \times d)$ feature matrix of I_i.

2.2 Image Recognition

In the previous section 2.1, a set of images is projected on to 2D-LPI subspace. The distance between two projected matrices say $Y_i = \left[Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(d)}\right]$, $Y_j = \left[Y_j^{(1)}, Y_j^{(2)}, \ldots, Y_j^{(d)}\right]$ is calculated using Euclidean distance between the corresponding vectors i.e.,

$$dist(Y_i, Y_j) = \sum_{k=1}^{d} \left\| Y_i^{(k)} - Y_j^{(k)} \right\| \tag{4}$$

Suppose there are t_n number of training samples $T_1, T_2, \ldots, T_{t_n}$ which are projected onto a 2D-LPI space and each of these samples are assigned to a class C. The class C of a test image $T^{|}$ is identified by the use of nearest neighbor rule based Euclidean distances.

3 Experimentation 1

To study the efficacy of the proposed 2D-LPI we have carried out experimentation on video summarization, face recognition and fingerspelling recognition.

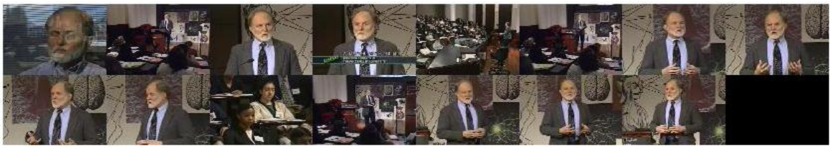


Fig. 1. Manually generated story board

3.1 Video Summarization

Video summarization is a process of representing a complete video with a small set of visually important frames. Video summarization is necessary for easy and fast content accessing and browsing of a large video database. Also it facilitates content based-video indexing and retrieval [11]. To study the efficacy of the proposed 2D-LPI, we conducted an experimentation to summarize a lecture video. While conducting the experimentation we decoded the video into frames. Then each frame is converted from RGB color to gray-level image and resized into 40 × 60 pixel image size. A lecture video contains 9 shots with 2743 frames at 30 frames per second. A story board of the lecture video is manually created by selecting one or more key frames from each shot as shown in Fig. 1.

Inorder to create automatic story board all frames of the video is used and projected onto the proposed 2D-LPI model. The number of eigen vectors (d = 4) were selected experimentally which gave best results and K = 5 is used to construct the adjacency matrix G. The obtained image features are used to cluster the frames using C-means clustering. C-means algorithm was applied 10 times with different starting points, and the best result in terms of the objective function of C-means was recorded. The centroid of each cluster is used as cluster representative which are used in the automatic generated story board and the generated story board using the proposed method is as shown in Fig. 2 (a). Inorder to know effectiveness of the proposed method on video summarization we compare the summarization results obtained by 2D-LPP [11] method, which is used for video summarization recently. Figure 2(b) shows the story board generated using 2D-LPP method.

Table 1 shows the details of the shots used for experimentation, row labeled expert depicts to number of keyframes identified by an expert and row corresponding to the proposed and 2D-LPP depicts the number of keyframes gen-

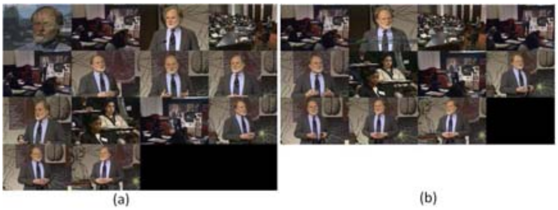


Fig. 2. (a) Automatic generated story board using proposed 2D-LPI method, (b) Automatic generated story board using proposed 2D-LPP method

Table 1. Shot details and the number of keyframes selected manually and proposed 2D-LPI method and existing 2D-LPP method

Shot	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Shot Details	[1-16]	[17-130]	[131-752]	[753-912]	[913-1007]	[1008-1657]	[1658-1847]	[1848-1999]	[2000-2743]
Expert	**1**	1	**2**	1	1	**4**	1	1	**3**
Proposed	**1**	1	**1**	1	1	**4**	1	1	**3**
2D-LPP	**0**	1	**1**	1	1	**1**	1	1	**4**

erated by the proposed method and 2D-LPP method [11]. From the Table 1, it can be seen that the existing 2D-LPP method omits a few keyframes where as the proposed method extracts keyframes approximately equal to the expert. Another important observation to be made is, shot 1 contains only 15 frames which has frames containing gradual transition from shot 1 to shot 2 and hence we have zero number of keyframes of shot 1 by 2D-LPP method. The proposed method extract key frames from shot 1, where as the 2D-LPP merges the shot 1 with shot 2. This type of video was intensionally taken to study the performance of both the methods.

3.2 Experimentation 2

In order to corroborate the efficacy of the proposed method we also conducted experimentation on ORL face, Yale face and fingerspelling datasets [12]useful for sign language recognition. The face images were resized into 40 × 50 pixels and the first six images of the ORL and Yale databases are used for training and the remaining four images are used for testing with K = 5 to construct the adjacency matrix G. The ORL and Yale achieved a maximum recognition rate of 96.88 and 88 percentage using conventional LPI with 25 and 50 dimensions and the proposed 2D- LPI method achieved a maximum recognition rate of 98.75 and 96 percentage for 40 × 7 dimensions respectively. On fingerspelling dataset conventional LPI achieved 86.88 percentage for 55 dimensions and proposed 2D-LPI achieved a maximum recognition rate of 89.88 percentage for 50 × 7 dimensions. Also we compare the recognition rate of the proposed method with 2D-PCA, 2D-FLD and 2D-LPP and the results are summarized in Table 2. From Table 2 it can be seen that the proposed method outperforms the other methods on both face and fingerspelling datasets.

Table 2. Maximum recognition rates on face datasets (ORL and Yale) and fingerspelling dataset

	2DPCA	2DFLD	2DLPP	2DLPI
ORL	96.88	97.50	96.25	98.75
Yale	90.67	89.00	88.00	96.00
Fingerspelling	88.75	89.27	72.50	89.38

4 Conclusion

In this paper we have proposed a two dimensional locality preserving indexing (2D-LPI) method suitable for images. The proposed method is more intuitive in handling images and performs better than the conventional LPI and is even better than the 2D-LPP. To study the efficacy of the proposed method, we have carried out experimentation on video summarization, face recognition and fingerspelling recognition. From the experimentation it is evident that 2D-LPI performs better than conventional LPI and also perform better than 2D-LPP, 2D-PCA and 2D-FLD method.

References

1. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society of Information Science 41, 391–407 (1990)
2. Kurimo, M.: Indexing audio documents by using latent semantic analysis and som. In: Kohonen maps, pp. 363–374. Elsevier, Amsterdam (1999)
3. Zhao, R., Grosky, W.I.: From features to semantics: Some preliminary results. In: Proceedings of International Conference on Multimedia and Expo (2000)
4. Souvannavong, F., Merialdo, B., Huet, B.: Video content modeling with latent semantic analysis. In: Proceedings of Third International Workshop on Content-based Multimedia Indexing (2003)
5. He, X., Cai, D., Liu, H., Ma, W.Y.: Locality preserving indexing for document representation. In: Proceedings of International Conference on Research and Development in Information Retrieval, pp. 96–103 (2004)
6. He, X., Niyogi, P.: Locality preserving projections. In: Advances in Neural Information Processing Systems (2003)
7. Cai, D., He, X., Han, J.: Document clustering using locality preserving indexing. IEEE Transactions on Knowledge and Data Engineering 17, 1624–1637 (2005)
8. Yang, J., Zhang, D., Frangi, A.F., yu Yang, J.: Two-dimensional pca a new approach to appearance-based face representation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(1), 131–137 (2004)
9. Li, M., Yuan, B.: 2d-lda a statistical linear discriminant analysis for image matrix. Pattern Recognition 26(5), 527–532 (2005)
10. Chen, S., Zhao, H., Kong, M., Luo, B.: 2d-lpp a two dimensional extension of locality preserving projections. Neurocomputing 70, 912–921 (2007)
11. Xu, L.Q., Luo, B.: Appearance-based video clustering in 2d locality preserving projection subspace. In: Proceedings of ACM International Conference on Image and Vide Retrieval (2007)
12. Guru, D.S., Suraj, M.G.: Fusion of pca and fld at feature extraction level for finger spelling recognition. In: Proceedings of the Third Indian International Conference on Atrificial Intelligence (IICAI), pp. 2113–2123 (2007)

A Geometric Algorithm for Learning Oblique Decision Trees

Naresh Manwani and P.S. Sastry

Indian Institute of Science, Bangalore-12, India
{naresh,sastry}@ee.iisc.ernet.in

Abstract. In this paper we present a novel algorithm for learning oblique decision trees. Most of the current decision tree algorithms rely on impurity measures to assess goodness of hyperplanes at each node. These impurity measures do not properly capture the geometric structures in the data. Motivated by this, our algorithm uses a strategy, based on some recent variants of SVM, to assess the hyperplanes in such a way that the geometric structure in the data is taken into account. We show through empirical studies that our method is effective.

1 Introduction

Decision tree is a well known and widely used method for classification and regression. The popularity of decision tree is because of its simplicity. In this paper we deal with oblique decision trees. These are binary decision trees where each non-leaf node is associated with a hyperplane (also called a split rule) and each leaf with a class label. To classify a pattern we follow a path in the tree by going down the left or right subtree at each node based on whether the pattern is on the positive or negative side of the hyperplane associated with the node; when we reach the leaf the pattern gets assigned the label of the leaf. Most approaches for learning decision trees are recursive algorithms for building the tree in a top down fashion. At each stage, we need to find a hyperplane to associate with the current node which is done by searching for a hyperplane which minimizes some impurity measure. Gini index and entropy are frequently used impurity measures. CART-LC [2], OC1 [8] are few examples of decision tree approaches which learn oblique decision trees by optimizing such impurity measures. All decision tree algorithms employ greedy search and hence a hyperplane assigned to a node once, is never changed. If a bad hyperplane is learnt at some node, it can not be rectified anymore and the effect of it may be a large tree which, in turn, may lead to poor generalization error.

A problem with all impurity measures is that they depend only on the number of (training) patterns of either class on each side of the hyperplane. If we change the class regions without changing the effective areas of class regions on either side of the hyperplane, the impurity measure will not change. Thus the impurity measures do not really capture the geometric structure of class distributions. In this paper we present a new decision tree learning algorithm which is based

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 25–31, 2009.

on the idea of capturing the geometric structure. For this we borrow ideas from some recent variants of the SVM method which are quite good at capturing (linear) geometric structure of the data. The Proximal SVM [5] finds two parallel clustering hyperplanes, one for each class, and takes the hyperplane situated in the middle of the two as the classifying hyperplane. However, often the points of the two classes are not clustered around parallel hyperplanes. Multisurface Proximal SVM [6] finds two clustering hyperplanes, one for each class, which are not necessary parallel, and the data is classified based on the nearness to the hyperplanes. Many times the class distribution is such that just one pair of hyperplanes is not good enough for classification. For such cases, they [6] extend this idea using the kernel trick of (effectively) learning the pair of hyperplanes in a high dimensional space to which the patterns are transformed.

Motivated by Proximal SVM, we derive our decision tree approach as follows. At each node, we use the idea of Multisurface Proximal SVM to find the linear tendencies in the data by learning hyperplanes around which the data of different classes are clustered. After finding these hyperplanes, the natural guess for a split rule at this node is the angle bisector of the two hyperplanes. This allows us to capture geometry of the classification boundary to some extent. Then, in the usual style of decision tree algorithms, we split the data based on this angle bisector and recursively learn the left and right subtrees of this node. Since, in general, there will be two angle bisectors, we select the one which is better based on an impurity measure. In this paper we discuss the algorithm for only two class classification problem. Since our decision tree approach is motivated by the geometry of the classification problem, we call it **Geometric Decision Tree** (Geometric DT).

2 The Algorithm

For a general decision tree algorithm, the main computational task is: given a set of data, find the best hyperplane to split the data. Let S^t be the set of points falling at node t. Let n_+^t and n_-^t denote the number of patterns of the two classes (say, '+1' and '-1') at that node. Let d be the data dimension. Let $A \in \Re^{n_+^t \times d}$ be the matrix representing points of class +1 as row vectors at node t. A_i is the ith row of A which is a row vector in $\Re^d$. Similarly $B \in \Re^{n_-^t \times d}$ be the matrix whose rows contain points of class -1 at node t. Let $h_1 : \mathbf{w}_1^t \mathbf{x} + b_1 = 0$ and $h_2 : \mathbf{w}_2^t \mathbf{x} + b_2 = 0$ be two clustering hyperplanes, one for each class. As mentioned before we use Multisurface Proximal SVM [6], also called Generalized Proximal SVM (GEPSVM), to find the two clustering hyperplanes. The hyperplane h_1 is closest to all points of class +1 and farthest from class -1. Similarly the hyperplane h_2 is closest to all points of class -1 and farthest from class +1. Then the optimization problems for finding $(\mathbf{w}_1, b_1)$ and $(\mathbf{w}_2, b_2)$ can be written as [6]

$$(\mathbf{w}_1, b_1) = \text{argmin}_{(\mathbf{w},b)\neq 0} \frac{||A\mathbf{w} + b\mathbf{e}_{n_+^t}||^2}{||B\mathbf{w} + b\mathbf{e}_{n_-^t}||^2} \tag{1}$$

$$(\mathbf{w}_2, b_2) = \text{argmin}_{(\mathbf{w},b)\neq 0} \frac{||B\mathbf{w} + b\mathbf{e}_{n^t_-}||^2}{||A\mathbf{w} + b\mathbf{e}_{n^t_+}||^2} \tag{2}$$

Where $||.||$ denotes the standard euclidean norm, $\mathbf{e}_{n^t_+}$ is $(n^t_+ \times 1)$ vector of ones and $\mathbf{e}_{n^t_-}$ is $(n^t_- \times 1)$ vector of ones. These optimization problems can be further reduced to [6]

$$\tilde{\mathbf{w}}_1 = \text{argmin}_{\tilde{\mathbf{w}}\neq 0} \frac{\tilde{\mathbf{w}}^T G \tilde{\mathbf{w}}}{\tilde{\mathbf{w}}^T H \tilde{\mathbf{w}}} \tag{3}$$

$$\tilde{\mathbf{w}}_2 = \text{argmin}_{\tilde{\mathbf{w}}\neq 0} \frac{\tilde{\mathbf{w}}^T H \tilde{\mathbf{w}}}{\tilde{\mathbf{w}}^T G \tilde{\mathbf{w}}} \tag{4}$$

where $\tilde{\mathbf{w}} = (\mathbf{w}, b)$, $G = [A\ \mathbf{e}_{n^t_+}]^T[A\ \mathbf{e}_{n^t_+}]$ and $H = [B\ \mathbf{e}_{n^t_-}]^T[B\ \mathbf{e}_{n^t_-}]$. This problem of finding the two clustering hyperplanes $(\mathbf{w}_1, b_1)$ and $(\mathbf{w}_2, b_2)$ gets reduced [6] to finding the eigenvectors corresponding to the maximum and minimum eigenvalue of the following generalized eigenvalue problem

$$H\tilde{\mathbf{w}} = \gamma G\tilde{\mathbf{w}} \tag{5}$$

Once we find these hyperplanes, the hyperplane we associate with the current node will be one of the angle bisectors of these two hyperplanes because it divide the space in such a way that on one side points of one class are clouded and on the other side points of second class are clouded. Let the hyperplanes $\mathbf{w}_3^t\mathbf{x} + b_3 = 0$ and $\mathbf{w}_4^t\mathbf{x} + b_4 = 0$ be the angle bisectors of $\mathbf{w}_1^t\mathbf{x} + b_1 = 0$ and $\mathbf{w}_2^t\mathbf{x} + b_2 = 0$, then

$$\tilde{\mathbf{w}}_3 \triangleq (\mathbf{w}_3, b_3) = (\frac{\mathbf{w}_1}{||\mathbf{w}_1||} - \frac{\mathbf{w}_2}{||\mathbf{w}_2||}, \frac{b_1}{||\mathbf{w}_1||} - \frac{b_2}{||\mathbf{w}_2||}) \tag{6}$$

$$\tilde{\mathbf{w}}_4 \triangleq (\mathbf{w}_4, b_4) = (\frac{\mathbf{w}_1}{||\mathbf{w}_1||} + \frac{\mathbf{w}_2}{||\mathbf{w}_2||}, \frac{b_1}{||\mathbf{w}_1||} + \frac{b_2}{||\mathbf{w}_2||}) \tag{7}$$

We choose the angle bisector which has lower impurity. In our work we use *gini* index to measure impurity. Let $\tilde{\mathbf{w}}_t$ be a hyperplane which is used for dividing the set S^t in two parts S^{t_l} and S^{t_r}. Let $n^{t_l}_+$ and $n^{t_l}_-$ denote the number of patterns of the two classes in the set S^{t_l} and $n^{t_r}_+$ and $n^{t_r}_-$ denote the number of patterns of the two classes in the set S^{t_r}. Then *gini* index of hyperplane $\tilde{\mathbf{w}}_t$ is given by,

$$gini(\tilde{\mathbf{w}}_t) = \frac{n^{t_l}}{n^t}\left[1 - (\frac{n^{t_l}_+}{n^{t_l}})^2 - (\frac{n^{t_l}_-}{n^{t_l}})^2\right] + \frac{n^{t_r}}{n^t}\left[1 - (\frac{n^{t_r}_+}{n^{t_r}})^2 - (\frac{n^{t_r}_-}{n^{t_r}})^2\right] \tag{8}$$

where $n^t = n^t_+ + n^t_-$ is the number of points in S_t. Also $n^{t_l} = n^{t_l}_+ + n^{t_l}_-$ is the number of points falling in the set S^{t_l} and $n^{t_r} = n^{t_r}_+ + n^{t_r}_-$ is the number of points falling in the set S^{t_r}. We choose $\tilde{\mathbf{w}}_3$ or $\tilde{\mathbf{w}}_4$ to be the split rule for S^t based on which of the two gives lesser value of *gini* index given by Eq.(8). Algorithm. 2 describes the complete decision tree learning algorithm.

3 Experiments

To test the effectiveness of Geometric DT we test its performance on several synthetic and real world data sets. We consider three synthetic problems

Algorithm 1. Geometric DT

```
Input: Sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1\ldots n}$, Max-Depth, $\epsilon_1$
Output: Pointer to the root of a decision tree
begin
    Root=$growTree(S)$;
    return $Root$;
end

Procedure $growTree(S^t)$
  Input: Set of patterns at node t, $S^t$
  Output: Pointer to a subtree
begin
    1.  – Find matrices $A$ and $B$ using the set $S^t$ of patterns at node $t$.
        – Find $\tilde{\mathbf{w}}_1$ and $\tilde{\mathbf{w}}_2$ according to Eq.(3),(4).
        – Compute angle bisectors $\tilde{\mathbf{w}}_3$ and $\tilde{\mathbf{w}}_4$ using Eq.(6),(7).
        – Calculate the $gini$ index (Eq.(8)) for both the angle bisectors $\tilde{\mathbf{w}}_3$ and $\tilde{\mathbf{w}}_4$.
        – Choose the one which gives lesser value of $gini$ index. Call it $\tilde{\mathbf{w}}^*$.
        – Let $\tilde{\mathbf{w}}_t$ denote the split rule at node $t$. Assign $\tilde{\mathbf{w}}_t \leftarrow \tilde{\mathbf{w}}^*$.
    2.  Let $S^{t_l} = \{\mathbf{x}_i \in S^t | \tilde{\mathbf{w}}_t^T \tilde{\mathbf{x}} < 0\}$ and $S^{t_r} = \{\mathbf{x}_i \in S^t | \tilde{\mathbf{w}}_t^T \tilde{\mathbf{x}} \geq 0\}$
        Define $\eta(S^t) = \frac{\#\text{points of minority class at node } t}{\#\text{points at node } t}$.
    3.  if $(\eta(S^{t_l}) < \epsilon_1)$ or (Tree-Depth=Max-Depth) then
        get a node $t_l$ and make $t_l$ a leaf node and assign appropriate class label to $t_l$;
        else $t_l$=growTree($S^{t_l}$);
        make $t_l$ left child of $t$;
    4.  if $(\eta(S^{t_r}) < \epsilon_1)$ or (Tree-Depth=Max-Depth) then
        get a node $t_r$ and make $t_r$ a leaf node and assign appropriate class label to $t_r$;
        else $t_r$=growTree($S^{t_r}$);
        make $t_r$ right child of $t$;
    5.  return $t$;

end
```

and two real world data sets from UCI ML repository [1]. First two synthetic problems are rotated 2×2 and 4×4 checkerboard data sets in two dimensions. The third problem is in $\Re^{10}$. We sample points uniformly from $[-1\ 1]^{10}$. Let $\tilde{\mathbf{w}}_1 = [1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0]'$ $\tilde{\mathbf{w}}_2 = [1, -1, 0, 0, 1, 0, 0, 1, 0, 0, 0]'$ and $\tilde{\mathbf{w}}_3 = [0, 1, -1, 0, -1, 0, 1, 1, -1, 1, 0]'$ be three hyperplane parameters in $\Re^{10}$. Now the points are labeled +1 or -1 based on the following rule

$$y = \begin{cases} 1, & \text{if} \quad (\tilde{\mathbf{w}}_1^t\tilde{\mathbf{x}} \geq 0 \ \&\ \tilde{\mathbf{w}}_2^t\tilde{\mathbf{x}} \geq 0 \ \&\ \tilde{\mathbf{w}}_3^t\tilde{\mathbf{x}} \geq 0) \ || \\ & \qquad (\tilde{\mathbf{w}}_1^t\tilde{\mathbf{x}} \leq 0 \ \&\ \tilde{\mathbf{w}}_2^t\tilde{\mathbf{x}} \leq 0 \ \&\ \tilde{\mathbf{w}}_3^t\tilde{\mathbf{x}} \geq 0) \ || \\ & \qquad (\tilde{\mathbf{w}}_1^t\tilde{\mathbf{x}} \leq 0 \ \&\ \tilde{\mathbf{w}}_2^t\tilde{\mathbf{x}} \geq 0 \ \&\ \tilde{\mathbf{w}}_3^t\tilde{\mathbf{x}} \leq 0) \ || \\ & \qquad (\tilde{\mathbf{w}}_1^t\tilde{\mathbf{x}} \geq 0 \ \&\ \tilde{\mathbf{w}}_2^t\tilde{\mathbf{x}} \leq 0 \ \&\ \tilde{\mathbf{w}}_3^t\tilde{\mathbf{x}} \leq 0) \ || \\ -1, & \text{else} \end{cases}$$

For each synthetic problem 2000 points were sampled for training set and 1000 for testing set. To evaluate the performance of the algorithm on real dataset we use Breast Cancer Data set and Bupa Liver Disorder Data set [1]. Breast

Cancer Data set is a 10 dimensional binary classification problem. It has total 683 instances, out of which we use 500 for training and 183 for testing. Bupa Liver Disorder Data set is a 6-dimensional binary classification problem. It has a total 345 instances, out of which we use 200 for training and 145 for testing. We compare performance of our approach with several algorithms, namely, CART-LC [2], OC1-AP, OC1(Oblique) [8], Support Vector Machine (SVM) [3] and GEPSVM [6]. For OC1 and CART-LC we use the downloadable [7] package. For SVM we use libsvm [4] code. Geometric DT is implemented in MATLAB. For GEPSVM also we wrote MATLAB code. We compare the performances on several measures like training time, training accuracy, test accuracy, number of leaves and depth of the tree. In the SVM method we used Gaussian kernel to capture non-linear classification boundaries. In case of GEPSVM also we used Gaussian kernel for all problems except the first problem where a pair of hyperplanes is good enough for classification. Best parameters for kernel were found by using 10-fold cross validation. The comparison results are shown in Table.1 (For OC1 and CART-LC which are randomized algorithms we show mean and standard deviation of results on 10 runs.) From Table. 1 we see that, over all the problems, Geometric DT performs better than all the other decision tree approaches in test accuracy, tree size and number of leaves. GEPSVM with linear kernel performs same as Geometric DT for the 2×2 checkerboard problem

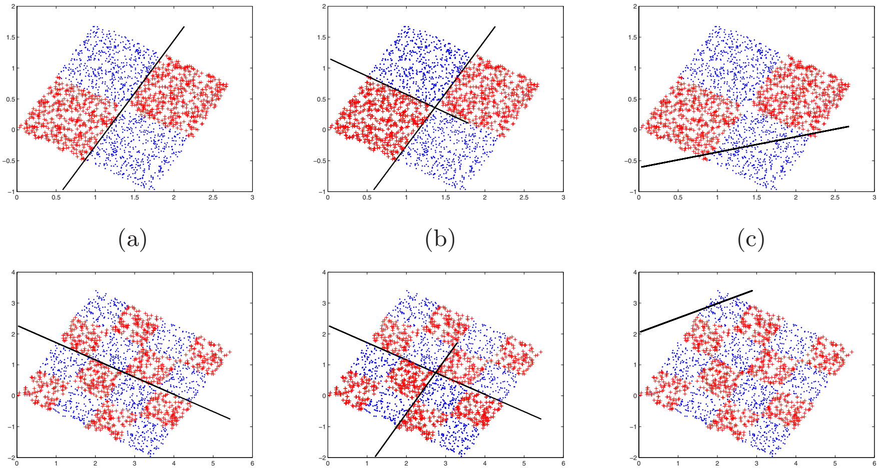
Table 1. Comparison results on synthetic and real world data sets

		Geometric DT	OC1(AP)	OC1 (Oblique)	CART-LC	SVM	GEPSVM
2×2 Checker-board	TrTime(sec)	0.064	0.031	3.112±0.83	1.049±0.8	0.132	**.004**
	TrAcc	**99.70**	99.34±0.3	99.40±0.92	98.57±1	99.25	**99.70**
	TesAcc	**99.90**	97.45±0.55	98.52±1.17	96.98±0.94	99.5	**99.90**
	# leaves	**4**	39±5.83	17.8±6.84	22±8.1	-	-
	Depth	**2**	11.8±1.93	8.4±3.24	9.6±2.8	-	-
4×4 Checker-board	TrTime(sec)	0.062	**0.047**	4.247±0.53	2.029±0.98	0.154	451.86
	TrAcc	97.75	98.28±0.4	**98.78**±0.76	98.08±0.71	98.10	90.70
	TesAcc	**98.50**	92.48±0.86	94.65±0.91	92.53±0.9	97.50	89.80
	# leaves	**16**	18.4±15.08	82.7±21.96	104.4±19.18	-	-
	Depth	**4**	16.3±2	14.6±2.32	15.4±2.46	-	-
10 Dim-ensional Dataset	TrTime(sec)	**0.1192**	0.249±0.02	18.37±1.73	31.54±9.67	1.048	384.10
	TrAcc	85.20	72.10±2.88	74.76±3.64	61.83±1.54	**98.65**	92.20
	TesAcc	**84.60**	60.40±1.86	63.50±0.74	61.29±1.46	74.60	67.30
	# leaves	26	55.8±15.27	31.7±12.81	**16.3**±7.6	-	-
	Depth	**7**	11.3±1.57	9.7±2.11	7.9±1.79	-	-
Breast Cancer	TrTime(sec)	0.0145	0.011±0.001	1.41±0.42	**0.10**±0.02	0.061	9.63
	TrAcc	96.4	97.30±1.03	97.02±0.76	95.14±1.65	**100**	**100**
	TesAcc	**100**	96.77±0.91	97.27±1.15	93.34±3.18	77.05	77.05
	# leaves	3	7.9±4.68	4.9±2.33	**2.8**±1.32	-	-
	Depth	2	4±1.63	2.9±1.29	**1.7**±1.06	-	-
Bupa Liver Disorder	TrTime(sec)	0.0137	**0.008**	0.74±0.15	0.06±0.01	0.018	0.83
	TrAcc	77	81.3±6.71	84.2±5.15	80.9±6.38	**100**	79.00
	TesAcc	72.41	59.3±2.28	60.76±2.67	59.66±2.33	**73.1**	71.72
	# leaves	**11**	11.5±6.45	12.9±5.51	12.5±7.09	-	-
	Depth	7	7±3.13	7.6±3.17	**6.7**±3.13	-	-

because for this problem the two approaches work in a similar way. But when there are more than two hyperplanes required, GEPSVM with Gaussian kernel performs worse than our decision tree approach. For example, for 4×4 checkerboard example, GEPSVM can achieve only about 89.8% test accuracy while our decision tree gives about 98.5% test accuracy. Moreover, with Gaussian kernel, GEPSVM solves the generalized eigenvalue problem of the size of number of points, whereas our decision tree solves the generalized eigenvalue problem of the dimension of the data (which is the case with GEPSVM only when it uses linear kernel). This gives us an extra advantage in computational cost over GEPSVM.

In 10-dimensional problem, both in test accuracy and training time, Geometric DT performs better than all the other approaches. This shows that our algorithm scales with data dimension. For the real world problems also Geometric DT performs well. It gives highest test accuracy with Breast Cancer Data set. For Liver Disorder Data set also it performs better than all other decision tree approaches and the test accuracy of SVM is marginally better.

In Figure. 1 we show the effectiveness of our algorithm in terms of capturing the geometric structure of the classification problem. We show the first few hyperplanes generated by our approach and OC1(oblique) for 2×2 and 4×4 checkerboard data. We see that our approach learns the correct geometric structure of the classification boundary, whereas the OC1(oblique), which uses the *gini* index as impurity measure, does not capture that.



(a) (b) (c)

(d) (e) (f)

Fig. 1. Comparisons of Geometric DT with OC1(Oblique) on rotated 2×2 and 4×4 checkerboard data: (a)Hyperplane at root node learnt using Geometric DT on rotated 2×2 checkerboard data; (b) Hyperplane at left child of root node learnt using Geometric DT on rotated 2×2 checkerboard data; (c)Hyperplane at root node learnt using OC1(Oblique) DT on rotated 2×2 checkerboard data; (d), (e) and (f) show the hyperplanes for 4×4 checkerboard data in a similar way

4 Conclusions

In this paper we presented a new algorithm for learning oblique decision trees. The novelty is in learning hyperplane that captures the geometric structure of the data using the GEPSVM idea. Through empirical studies we showed that the method performs better than the other approaches and also captures the geometric structure well. Though we consider only the 2-class problems here, the approach can easily be generalized to multiclass problems by learning a split rule, at each node, to separate the majority class from the rest. We will be explaining this in our future work.

References

1. Asuncion, A., Newman, D.J.: UCI machine learning repository (2007)
2. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth and Brooks (1984)
3. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. In: Knowledge Discovery and Data Mining, vol. 2, pp. 121–167 (1998)
4. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. Fung, G., Mangasarian, O.L.: Proximal support vector machine classifiers. In: Knowledge Discovery and Data Mining (KDD), pp. 77–86 (2001)
6. Mangasarian, O.L., Wild, E.W.: Multisurface proximal support vector machine classification via generalized eigenvalues. IEEE Transaction on Pattern Analysis and Machine Intelligence 28(1), 69–74 (2006)
7. Murthy, S.K., Kasif, S., Salzberg, S.: The OC1 decision tree software system (1993), http://www.cs.jhu.edu/~salzberg/announce-oc1.html
8. Murthy, S.K., Kasif, S., Salzberg, S.: A system for induction of oblique decision trees. Journal of Artificial Intelligence Research 2, 1–32 (1994)

Kernel Optimization Using a Generalized Eigenvalue Approach

Jayadeva, Sameena Shah, and Suresh Chandra

Indian Institute of Technology Delhi, Hauz Khas,
New Delhi -110016, India

Abstract. There is no single generic kernel that suits all estimation tasks. Kernels that are learnt from the data are known to yield better classification. The coefficients of the optimal kernel that maximizes the class separability in the empirical feature space had been previously obtained by a gradient-based procedure. In this paper, we show how these coefficients can be learnt from the data by simply solving a generalized eigenvalue problem. Our approach yields a significant reduction in classification errors on selected UCI benchmarks.

Keywords: Data dependent kernel, Fisher's coefficient, generalized eigenvalue, kernel optimization, Rayleigh quotient.

1 Introduction

SVMs aim to minimize the upper bound on the generalization error by maximizing the margin between the separating hyperplane and the data. If a linear separator in the input space does not suffice, a non-linear mapping is used to map the input space into a higher dimensional Euclidean or Hilbert feature space. This embedding induces a Riemannian metric in the input space, which is either enlarged or reduced in the feature space. Thus, by choosing an "appropriate" mapping $\phi(\cdot)$, the data points become linearly separable or mostly linearly separable in the high dimensional feature space, enabling easy application of structural risk minimization. Each kernel induces a different mapping or structure of the data in the embedding space, which may or may not be suitable from the classification perspective. To obtain a good quality embedding, that is more appropriate for the estimation task, the choice of the kernel should be "learned" from the structure of the data rather than chosen through some trial and error heuristics. An improper choice of kernel may lead to worse classification. In this paper, we extend the approach of [3], which tries to optimize the data dependent kernel by maximizing the class separability criterion in the empirical feature space. The transformation from the input space to an r-dimensional Euclidean space is given by $x \longrightarrow D^{-1/2}P^T(k(x,x_1), k(x,x_2), \ldots, k(x,x_m))^T$, where, $K_{m\times m} = P_{m\times r}D_{r\times r}P_{r\times m}$. Their approach has led to obtaining kernels that yield a significantly better classification performance compared to primary Gaussian or polynomial kernels for k-Nearest Neighbor (KNN), Kernel Fisher Discriminant (KFD), Kernel Minimum Squared Error machine (KMSE), and SVM on the UCI benchmark data sets [4]. However, only a slight improvement

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 32–37, 2009.

in performance is observed on using the optimal kernels for SVMs. Our approach reduces to solving a single generalized eigenvalue problem. Our approach also has a number of other advantages, namely,

1. It avoids using an iterative algorithm to update alpha (coefficients of the optimal kernel) in each step, since now only a single generalized eigenvalue problem needs to be solved.
2. It avoids an ascent procedure that can not only potentially run into numerical problems when the Hessian matrices in question are not well conditioned, but also gets stuck in local maxima.
3. It also avoids tuning parameters such as learning rate and the number of iterations.

Section 2 formally defines our notion of optimality and develops the procedure to obtain the data dependent optimal kernel as a generalized eigenvalue solution. Section 3 gives the classification performance evaluation of our optimal kernel with a Gaussian kernel, and with the optimal kernel generated by the iterative update approach of [3] on UCI Benchmark data sets and section 4 concludes the paper.

2 Kernel Optimization

Maximizing the separation between the different classes is a natural step in the hope of obtaining better classification. To evaluate the measure of separability, we choose the Fisher Discriminant function [5]. It attempts to find a natural measure between examples induced by the generative model. In terms of the between class scatter matrix S_B and the within class scatter matrix S_W, the Fisher discriminant criterion $J(\cdot)$ can be written as (cf. [6]) $J(w) = \frac{w^T \ S_B \ w}{w^T \ S_W \ w}$. Substituting $\nabla J(w) = 0$ for maximizing $J(\cdot)$we get,

$$S_B w = J(w) S_W w. \tag{1}$$

Thus the extremum points w^* of the Rayleigh Quotient $J(w)$ are obtained as the eigenvectors of the corresponding generalized eigenvalue problem.

2.1 Data Dependent Optimal Kernel

Similar to Amari [2], we choose the data dependent kernel function, K, as a conformal transformation of the form $K(x, y) = q(x).q(y).k_0(x, y)$, where,

$$q(x) = \alpha_0 + \sum_{i=1}^{n_e} \alpha_i k_1(x, a_i) \tag{2}$$

In (2), k_1 denotes a primary Gaussian kernel of the form $k_1(x, a_i) = \exp(-\gamma \| x - a_i \|)^2$ and $\alpha = \{\alpha_0, \alpha_1, \ldots, \alpha_{n_e}\}$ is the vector of linear combination coefficients. The total number of empirical cores is denoted by n_e and a_i denotes the i-th empirical core. We choose a random set of one-third of the data points as the empirical cores. If the data is written as the first m_1 data points class C_1, followed by the rest m_2 data points of class C_2, then, the kernel matrix K_0 , corresponding to the primary kernel, can be written as

$$K_0 = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix} \tag{3}$$

where, K_{11}, K_{12}, K_{21} and K_{22} are sub-matrices of order $m_1 \times m_1,\ m_1 \times m_2,\ m_2 \times m_1$, and $m_2 \times m_2$ respectively. The "between-class" kernel scatter matrix B_0, and the "within - class" kernel scatter matrix W_0, can then be written as

$$B_0 = \begin{pmatrix} \frac{1}{m}K_{11} & 0 \\ 0 & \frac{1}{m}K_{22} \end{pmatrix} - \begin{pmatrix} \frac{1}{m}K_{11} & \frac{1}{m}K_{12} \\ \frac{1}{m}K_{21} & \frac{1}{m}K_{22} \end{pmatrix}, \tag{4}$$

and,

$$W_0 = \begin{pmatrix} k_{11} & 0 & \dots & 0 \\ 0 & k_{22} & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & k_{mm} \end{pmatrix} - \begin{pmatrix} \frac{1}{m}K_{11} & 0 \\ 0 & \frac{1}{m}K_{22} \end{pmatrix}. \tag{5}$$

Using (4) and (5), the Fisher's coefficient is equivalent to $J = \frac{q^T B_0 q}{q^T W_0 q}$, where,

$$q = \begin{pmatrix} 1 & k_1(x_1, a_1) & \dots & k_1(x_1, a_n) \\ 1 & k_1(x_2, a_1) & \dots & k_1(x_2, a_n) \\ \vdots & \dots & \dots & \vdots \\ 1 & k_1(x_m, a_1) & \dots & k_1(x_m, a_n) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = K_1\alpha \quad . \tag{6}$$

Here, K_1 is a matrix of order $m \times (n_e + 1)$ and α is as before. Hence J can now be used as a measure of separability of the data in the empirical feature space that is easy to compute. Maximizing J, to obtain the maximum separation of data requires finding the optimal q. Optimizing the projection q is in turn equivalent to optimizing α with respect to the data (cf. (2)). Maximization of this separability measure, to obtain the optimal kernel, thus amounts to finding the optimal set of coefficients of the vector α. These coefficients would optimally enlarge or reduce the spaces around the empirical cores to reduce the error maximally. To obtain this optimal α, an iterative procedure to maximize Rayleigh quotient J was proposed in [3]. The algorithm based on the general gradient descent method updates in all iterations. While a decreasing learning rate ensures convergence, the algorithm still requires an iterative update solution and is susceptible to getting trapped in local maxima. In the sequel we solve (11) as a generalized eigenvalue problem and obtain the optimal data dependent kernel.

2.2 The Generalized Eigenvalue Solution

On substituting (6), J can be written as

$$J = \frac{q^T B_0 q}{q^T W_0 q} = \frac{\alpha^T K_1^T B_0 K_1 \alpha}{\alpha^T K_1^T W_0 K_1 \alpha} \quad . \tag{7}$$

As a generalized eigenvalue problem, (6) can be written as $B_0 q = \lambda W_0 q$, where λ are the set of eigenvalues. This is solved by Lanczos method. For the optimal kernel, the eigenvector α, corresponding to the maximum eigenvalue is used. It also avoids the need for tuning parameters like number of iterations and learning rate. Further, it also avoids the possibility of getting stuck in local minima which is a generic problem associated with gradient descent based methods. The regularized measure of separability is then given by

$$J = \frac{\alpha^T K_1^T B_0 K_1 \alpha}{\alpha^T K_1^T (W_0 + DI) K_1 \alpha} \tag{8}$$

where D is a small constant and I is the identity matrix of the appropriate dimension.

3 Experimental Results

In this section we apply the algorithm to datasets of different level of complexity in terms of number of instances (m) and number of attributes (d). Ionosphere (m = 351 and d=34), Monks (m = 432 and d=6), Heart(m = 297 and d=13), Wisconsin (m = 569 and d=30). The data points with missing attributes, if any, have been removed. In the Ionosphere dataset, the column with zero variance has been removed. Each dataset has been normalized to lie between zero mean and unit variance. For all our experiments we have used (8), where D have been chosen through cross validation.

For evaluating our approach's performance, in each run, we randomly divide the data set into three disjoint sets. The data points in the first set are used for training while those in the second set are used for testing. The data points in the remaining third set are used as the empirical cores. Tables 1 through 4 compare the performance in terms of the classification error incurred (both training and testing separately) by using a primary Gaussian kernel, the kernel optimized using the approach proposed by Xiong et al. and the kernel optimized using the suggested approach illustrated in this paper. γ_0 corresponds to the parameter for the primary kernel while γ_1 is the parameter for the data dependent kernel. For the sake of comparison we choose the same set of values for γ_0 , γ_1 and C (for SVMs) as those in [3]. Each entry in the table, for a particular choice of kernel and a particular choice of value for γ_0 and γ_1 has been computed as the average of 20 random runs. The parameter for C has been chosen to be 1000 for all runs.

For most cases, the kernel optimized via the generalized eigenvalue approach gives significantly better classification errors than the corresponding other two methods. The comparison between the generalized eigenvalue approach based optimal kernel and using an arbitrary kernel clearly shows the huge difference in performance. This is because in [3], only an approximation to the optimal value of α is obtained for the equation $\frac{J_1}{J_2}\alpha = N_0^{-1} M_0 \alpha$ using an update of the form $\alpha^{n+1} = \alpha^n + \eta \left(\frac{1}{J_2} M_0 - \frac{J_1}{J_2^2} N_0 \right) \alpha^n$. On the other hand we mapped the problem to a generalized eigenvector problem. The generalized eigenvector problem is a quasiconvex optimization problem, since the constraint is convex and

Table 1. Training and testing error rates (%) using SVM classifier for the different kernels on Ionosphere data

(γ, γ_0)	Gaussian Kernel		Iterative update		Generalized eigenvalue	
	Training Error	Testing Error	Training Error	Testing Error	Training Error	Testing Error
(0.01, 0.0005)	1.4103	16.5812	0.3846	17.6923	0.2564	16.4530
(0.05, 0.0001)	3.8034	13.5470	5.3419	12.4786	0.5556	4.7009
(0.01, 0.0001)	4.5299	12.2650	3.9744	11.9658	2.7778	8.3761

Table 2. Training and testing error rates (%) using SVM classifier for the different kernels on the Monks-I dataset

(γ, γ_0)	Gaussian Kernel		Iterative update		Generalized eigenvalue	
	Training Error	Testing Error	Training Error	Testing Error	Training Error	Testing Error
(0.01, 0.0005)	32.7957	32.9655	33.5314	33.1353	9.2530	14.7142
(0.05, 0.0001)	31.5054	31.9624	33.4140	33.2258	7.7957	13.7366
(0.01, 0.0001)	31.2097	31.6129	33.4946	33.0645	6.2634	13.9785

Table 3. Training and testing error rates (%) using SVM classifier for the different kernels on the CLEVELAND HEART dataset

(γ, γ_0)	Gaussian Kernel		Iterative update		Generalized eigenvalue	
	Training Error	Testing Error	Training Error	Testing Error	Training Error	Testing Error
(0.01, 0.0005)	9.0377	17.9160	29.4524	30.8347	8.5061	21.1590
(0.05, 0.0001)	9.8990	16.5657	9.7475	16.4646	1.3131	28.9899
(0.01, 0.0001)	10.0000	16.3131	17.3737	22.0202	7.3232	21.6667
(0.05, 0.001)	7.3737	18.2323	12.2727	21.3131	1.1111	29.1919

Table 4. Training and testing error rates (%) using SVM classifier for the different kernels on the WISCONSIN BREAST CANCER dataset

(γ, γ_0)	Gaussian Kernel		Iterative update		Generalized eigenvalue	
	Training Error	Testing Error	Training Error	Testing Error	Training Error	Testing Error
(0.01, 0.0005)	0.6925	2.9640	9.4460	12.1330	0.8587	2.9640
(0.05, 0.0001)	1.1579	3.5789	1.8684	4.1053	0.6316	3.7632
(0.01, 0.0001)	1.4474	3.0526	10.2105	13.2632	1.1053	2.8684

the objective function is quasiconvex [7]. The Generalized eigenvalue problem is tractable and can be solved in polynomial time [7]. In practice they can be solved (a feasible point with an objective value that exceeds the global minimum by less than the prescribed accuracy can be obtained) very efficiently. Matlab uses semi-definite programming to obtain the global optimal of the GEVP problem [8]. Hence we could obtain the optimal solution, which is not possible in the other case. The approach in [3] was adopted to remove the problem of singularity. We corrected this problem using regularization. The generalized eigenvalue problem thus allows us to obtain the optimal solution without the need of any tuning of parameters or stopping conditions etc and without compromising the quality of the optimum. Overall one can conclude that if the empirical cores have the same distribution as the training and testing set, then classification accuracy can be improved significantly, sometimes by an order of magnitude, by the use of such data dependent kernels optimized by the generalized eigenvalue approach.

4 Conclusions and Discussion

The low classification error rates obtained by using the generalized eigenvalue approach to obtain the data dependent kernel implies that this approach yields a kernel that is better suited for the classification task because it can better adapt to the structure of the data and leads to a significant improvement in SVM's performance. This is because the generalized eigenvalue approach yields an exact solution that avoids getting stuck in a local maximum, and consequently leads us to obtain better kernels that yield a lower classification error. Moreover, we did not face any numerical problems in computing the generalized eigenvalue solution for the regularized Fisher's Discriminant function for various data sets. The performance of the proposed approach was compared with that of [3]. The experimental results convincingly demonstrated the effectiveness of our approach. Thus, it is evident that using the exact solution gives much better solutions that not only fit well, but also generalize well. In future, it would be interesting to explore the effect that the number of empirical cores would have on the classifier performance. We also plan to test our approach on other classifiers.

References

1. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control 25, 821–837 (1964)
2. Amari, S., Wu, S.: Improving support vector machine classifiers by modifying kernel functions. Neural Networks 12(6), 783–789 (1999)
3. Xiong, H., Swamy, M.N.S., Ahmad, M.O.: Optimizing the Kernel in the Empirical Feature Space. IEEE Trans. Neural Networks 16(2), 460–474 (2005)
4. Murphy, P.M., Aha, D.W.: UCI machine learning repository (1992), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
5. Jaakkola, T.S., Haussler, D.: Exploiting generating models in discriminative classifiers. In: Proc. of Tenth Conference on Advances in Neural Information Processing Systems, Denver (December 1998)
6. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn., pp. 117–124. John Wiley and Sons, Inc., Chichester (2001)
7. Boyd, S., El Ghaoui, L., Feron, E., Balakrishnan, V.: Linear Matrix Inequalities in System and Control Theory. Studies in Applied Mathematics, vol. 15. SIAM, Philadelphia (1994)
8. GEVP, Robust Control Toolbox. The Mathworks, `http://www.mathworks.com/access/helpdesk/help/toolbox/robust/gevp.html`

Zero Norm Least Squares Proximal SVR

Jayadeva[1], Sameena Shah[1], and Suresh Chandra[2]

[1] Dept. of Electrical Engineering, [2] Dept. of Mathematics, Indian Institute of Technology,
New Delhi 110016, India
jayadeva@ee.iitd.ac.in, sameena.shah@gmail.com,
chandras@maths.iitd.ac.in

Abstract. Least Squares Proximal Support Vector Regression (LSPSVR) requires only a single matrix inversion to obtain the Lagrange Multipliers as opposed to solving a Quadratic Programming Problem (QPP) for the conventional SVM optimization problem. However, like other least squares based methods, LSPSVR suffers from lack of sparseness. Most of the Lagrange multipliers are non-zero and thus the determination of the separating hyperplane requires a large number of data points. Large zero norm of Lagrange multipliers inevitably leads to a large kernel matrix that is inappropriate for fast regression on large datasets. This paper suggests how the LSPSVR formulation may be recast into one that also tries to minimize the zero norm of the vector of Lagrange multipliers, and in effect imposes sparseness. Experimental results on benchmark data show that a significant decrease in the number of support vectors can be achieved without a concomitant increase in the error.

Keywords: SVR, sparse representation, zero-norm, proximal, least squares.

1 Introduction

Support Vector Machines (SVMs) are computationally efficient for classification as well as regression tasks [1]. The elegance of SVMs lies in the fact that the nature of the optimization problem that needs to be solved for both linear and non-linear regression problems remains the same. The optimal hyperplane for the regression problem is determined by solving the problem stated in (1). Usually the dual of (1), a quadratic programming problem (QPP), is solved to obtain the support vectors.

$$\underset{q,q',w,b}{\text{Minimize}} \quad \frac{1}{2}w^T w + \mathrm{C}e^T\left(q+q'\right)$$

subject to

$$y-(\mathbf{P}w+\mathrm{b}) \le q,$$

$$(\mathbf{P}w+\mathrm{b})-y \le q',$$

$$q,q' \ge 0, \tag{1}$$

where, $w \in \Re^N$ is the weight vector, $\mathrm{b} \in \Re$ is the bias, q is the error for each data point, e is an M dimensional vector of ones, and $\mathrm{C} > 0$ is a parameter that trades off

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 38–43, 2009.

accuracy with the complexity of the regressor. **P** is the data matrix containing M data points, where each data point $x_i \in \Re^N$, $i=1,...,M$ has a corresponding function value y_i, $i=1,....,M$. LSSVMs [2] for regression solve the following optimization problem.

$$\underset{q,w}{\text{Minimize}} \ \frac{\text{C}}{2}\left(q^T q\right)+\frac{1}{2}\left(w^T w\right)$$

subject to

$$w^T x_i + b = y_i - q_i, \quad i=1,..., M \tag{2}$$

The introduction of the l_2 norm of the error variable in the objective function changes the nature of the optimization problem that needs to be solved to obtain the Lagrange multipliers. The solution of LSSVM can be obtained by solving the following system of equations

$$\begin{bmatrix} 0 & e^T \\ e & \mathbf{K}+\dfrac{I}{C} \end{bmatrix}\begin{bmatrix} \text{b} \\ u \end{bmatrix}=\begin{bmatrix} 0 \\ y \end{bmatrix} \tag{3}$$

where, **I** is an identity matrix (MxM), φ is the nonlinear mapping to higher dimension and **K** is the kernel matrix $K_{ij} = [\phi(P_i)]^T \phi\left(P_j\right), \ i, j = 1, 2, ..., M$.

LSSVMs require the solution of a set of linear equations rather than a QPP, for which the authors propose the use of iterative methods like SOR or conjugate gradient. Though LSSVM is computationally better than conventional SVMs, a disadvantage that least squares methods generally suffer from is the lack of sparseness. The support values are proportional to the errors at all the data points while in conventional SVMs many of them may be zero.

In this paper, we explore the possibility of using the zero norm in conjunction with least squares based methods to impose sparseness. The least squares based method we have chosen is Least Squares Proximal Support Vector Regression (LSPSVR) [3]. LSPSVR has been recently proposed and requires only the inversion of a single, positive definite matrix to obtain the Lagrange multipliers. However, LSPSVR also yields non-sparse solutions. It is important to have most components of β as zero, so as to obtain a kernel matrix of small size. In Section 2 we give the zero norm algorithm and show how it can be used to modify the LSPSVR formulation to obtain a sparse representation. Section 3 contains experimental results obtained for three data sets using this algorithm and finally section 4 is devoted to concluding remarks.

2 Zero Norm LSPSVR

The zero norm of a vector is defined as its number of non- zero components. Formally, the zero norm of a vector w is given by $\|w\|_0^0 = card\{w_i | w_i \neq 0\}$, where *card* denotes the set cardinality. The requirement of minimum number of non zero components of β can thus be formulated as the problem of minimizing the zero norm of β.

The l_p norm of the vector w is defined as $\|w\|_p = \left(\sum_{i=1}^{n} w_i^p\right)^{\frac{1}{p}}$. The minimization of the l_p norm of w in conjunction with the constraint $y_i(w.x_i + b) \geq 1$ solves the classical SVM problem. The generalization capabilities of such linear models have been studied for over a decade now. A general belief that has emerged out of this research is that for $p \geq 1$ the minimization of the l_p - norm of w is good for generalization. For $p = 0$, it has been shown in [4] that the problem of minimization of zero norm subject to the constraints $y_i(w.x_i + b) \geq 1$ is NP-hard. Therefore, one needs to use some approximation or local search method for solving the problem. Weston et al. [5] introduced an iterative algorithm, termed as 'Approximation of the zero-norm Minimization (AROM)' that performs a gradient step at each iteration and converges to a local minimum.

AROM Algorithm. AROM [6] solves the following optimization problem.

$$\underset{w}{\text{Minimize}} \quad \| w \|_0^0$$

subject to

$$y_i[w^T x_i + b] \geq 1, \qquad i = 1,...,l \tag{10}$$

i.e. it finds a separating hyperplane with the fewest nonzero elements as the coefficients of the vector ***w***. In order to minimize the zero norm of the vector $w = (w_1, w_2, ..., w_N)^T$, a vector of coefficients $z = (z_1, z_2, ..., z_N)^T$ is introduced.

```
1. Set $z = (1,\ldots,1)^T$
2. Solve
       $\text{Min} \sum_{j=1}^{n} | w_j |$
     subject to:
       $y_i\left(w(x_i * z) + b\right) \geq 1$
3. Let $\overline{w}$ be the solution of the previous problem. Set $z \leftarrow z * \overline{w}$.
4. Go back to 2 until convergence.
```

AROM requires the solution of a succession of linear programs combined with a multiplicative update to reduce the number of nonzero components of w, i.e. the dimension of the input data points. Following AROM we introduce a diagonal matrix **Z** in the LSPSVR formulation and propose the optimization problem.

Zero Norm LSPSVR

$$\underset{q,w,b}{\text{Minimize}} \quad \text{C} \cdot \frac{1}{2}\left(q^T q\right) + \frac{1}{2}\left(w^T w + \text{b}^2\right)$$

subject to

$$\left(\mathbf{K}(\mathbf{P}, \mathbf{P}^{\mathbf{T}})w + eb\right) - y + \mathbf{Z}q = 0. \tag{11}$$

The Lagrangian is given by

$$L(w,\mathrm{b},q,\beta)=\frac{C}{2}\|q\|^2+\frac{1}{2}\left\|\begin{bmatrix}w\\ \mathrm{b}\end{bmatrix}\right\|^2-\beta^T\left[\mathbf{K}w+e\mathrm{b}-y+\mathbf{Z}q\right]. \tag{12}$$

Substituting the KKT conditions, we get

$$\left(\mathbf{KK}^T+ee^T\right)\beta+\frac{\mathbf{ZZ}^T\beta}{C}=y\,. \tag{13}$$

On simplifying we obtain $\beta=\left[\mathbf{GG}^{\mathrm{T}}+\frac{\mathbf{ZZ}^{\mathbf{T}}}{\mathrm{C}}\right]^{-1}*y$ where $\mathbf{G}=[\mathbf{K}\quad e]$. The form of the equation that needs to be solved to obtain the Lagrange multipliers remains the same as in LSPSVR, and β can still be obtained by just a single matrix inversion, thus preserving the advantage offered by LSPVR. A modification we have introduced is that instead of minimizing the zero norm of weight vector i.e. the number of dimensions, we attempt to minimize the number of patterns. This implies the minimization of the number of components in the error vector $\boldsymbol{q}$. In other words, we attempt to have the same classification performance but such that comparatively large amount of error is contributed by only a few points rather than comparatively smaller amount of error contributed by a large number of points. This is done by multiplying **Z** by the error variable instead of the weight vector as done in feature selection. Since the support values for LSPSVR are related to the error vector, $q=\frac{\mathrm{Z}^T\beta}{\mathrm{C}}$, therefore the minimization of the zero norm of $\boldsymbol{q}$ implies the minimization of the zero norm of β.

The zero norm LSPSVR algorithm is summarized as below.

1. Set **Z** = diagonal matrix of ones.

2. Solve the problem in (11).

3. If $\bar{\beta}$ is the solution of the current problem. Update $\mathbf{Z}\longleftarrow\mathbf{Z}*\boldsymbol{q}=\mathbf{Z}*\mathbf{Z}^{\mathbf{T}}*\frac{\bar{\beta}}{\mathrm{C}}$

4. Eliminate all the data vectors i, st. $\mathbf{Z}_{ii}<0$, $\forall i$.

5. Go to Step 2 and iterate till the termination condition is met. There are many possible termination criteria, for example, allowed error bounds, bound on the number of data points, or the required sparseness.

3 Experimental Results

All the program codes were written in MATLAB 7.0 and executed on a Pentium III PC with 256 MB RAM. Co-ordinates of data samples were normalized to lie between zero and one. For each data set, the parameter C was determined by choosing a small tuning set of 20% samples. In all cases, regression was performed by using a polynomial kernel of degree 2, of the form $\mathrm{K}(x_i,x_j)=(x_i.x_j+1)^2$. Figures 1, 2 and 3 show how training and testing errors change as the number of data samples is reduced according to the zero norm LSPSVR algorithm on Comp activ, Census house and Boston housing benchmark data sets [7, 8]. The x-axis in each figure depicts the number

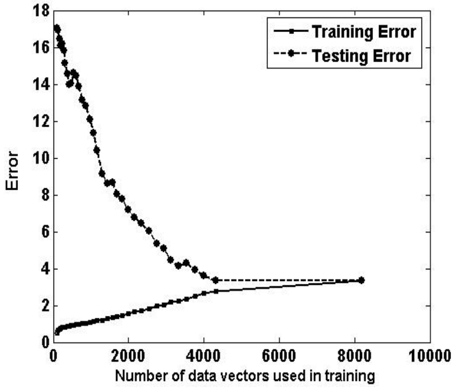


Fig. 1. Training and Testing errors on Comp Active data using 10 fold cross validation

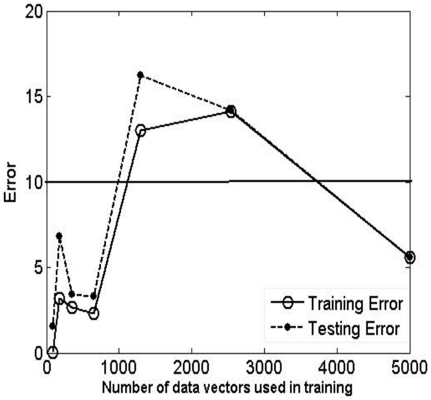


Fig. 2. Training and Testing errors on Census House data using 10 fold cross validation

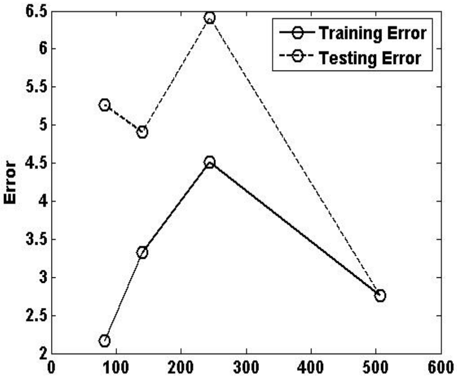


Fig. 3. Training and Testing errors on Housing data using 10 fold cross validation

of data points that were used in training, whereas the *y* axis indicates the average percentage error over 10 sample runs. The testing error was computed over all the data points, whereas the training error was determined as the total error over the data points used in training. The number of training patterns reduces after each iteration, while the number of testing patterns is a constant and equal to the size of the complete data set. Comp active has 8192 points lying in 22 dimensions. Figure 1 shows that the number of data points reduces to around 4000 in the first iteration itself without a significant increase in the error. The figure indicates that it may be best to use the result obtained after three or four iterations, because after that, the training error keeps decreasing while the testing error keeps increasing. Figure 2 shows the results for Census house data set of which we have used the first 5000 data points. We have also eliminated the dimensions whose value is a constant. The number of effective dimensions is thus 120. Although in the first couple of iterations the error goes up, but later on it comes down significantly and is much smaller than the initial value. This shows that it is possible to achieve better generalization while using fewer support vectors. An important point to note is that the number of data points has been reduced to about 60 from an initial value of 5000. This marks an enormous reduction in the size of the kernel from 5000 x 5000 to 60 x 60. Figure 3 shows the results for zero norm LSPSVR applied on Boston Housing data set that has 506 data points, each of dimension 14. The results show that the training error comes down and the testing error goes up with a marked reduction in the data set to about 16% of its initial size.

4 Conclusion

In this paper we present a novel algorithm that requires a single matrix inversion and a multiplicative update in each iteration to reduce the number of support vectors. Results on benchmark data sets show that it is possible to obtain a significant reduction in the number of support vectors within three or four iterations of the proposed algorithm. At each iteration, a significant number of data points are eliminated. Therefore, for later iterations the size of the data set reduces and problem (11) can be solved much faster. The zero norm LSPSVR approach was able to yield a sparse solution that did not significantly increase the error rate. However, in a couple of cases the generalization error actually reduced below the initial value. This could be because of elimination of outliers, but is an observation that merits further investigation. Since each iteration only requires a single matrix inversion, fast implementations of the proposed scheme are possible. The matrices to be inverted in successive steps are all positive definite, and related to each other, and it is therefore interesting to explore more efficient methods of implementing the proposed scheme. The zero norm approach can be applied to other least squares methods, most notably for imposing sparseness in LSSVM.

References

1. Cristianini, N., Taylor, J.S.: An Introduction to Support Vector Machines and other kernel based learning methods. Cambridge University Press, Cambridge (2000)
2. Suykens, J.: Least Squares Support Vector Machines. In: IJCNN (2003), `http://www.esat.kuleuven.ac.be/sista/lssvmlab/`
3. Jayadeva, R.K., Chandra, S.: Least Squares Proximal Support Vector Regression. Neurocomputing (communicated)
4. Amaldi, E., Kann, V.: On the approximability of minimizing non zero variables or unsatisfied relations in linear systems. Theoretical Computer Science 209, 237–260 (1998)
5. Weston, J., Elisseeff, A., Scholkopf, B.: Use of the l_0-norm with linear models and kernel methods. Technical report (2001)
6. Weston, J., Elisseeff, A., Scholkopf, B., Tipping, M.: Use of Zero Norm with Linear Models and Kernel Machines. Journal of Machine Learning Research 3, 1439–1461 (2003)
7. Murphy, P.M., Aha, P.M.: UCI Repository of Machine learning Databases (1992), `http://www.ics.uci.edu/mlearn/MLRepository.html`
8. Data for Evaluating Learning in Valid experiments, `http://www.cs.utoronto.ca/~delve`

Effect of Subsampling Rate on Subbagging and Related Ensembles of Stable Classifiers

Faisal Zaman* and Hideo Hirose

Kyushu Institute of Technology,
680-4 Kawazu, Iizuka-shi, Fukuoka, Japan
zaman@ume98.ces.kyutech.ac.jp,
hirose@ces.kyutech.ac.jp

Abstract. In ensemble methods to create multiple classifiers mostly bootstrap sampling method is preferred. The use of subsampling in ensemble creation, produce diverse members for the ensemble and induce instability for stable classifiers. In subsampling the only parameter is the subsample rate that is how much observations we will take from the training sample in each subsample. In this paper we have presented our work on the effect of different subsampling rate (SSR) in bagging type ensemble of stable classifiers, Subbagging and Double Subbagging. We have used three stable classifiers, Linear Support Vector Machine (LSVM), Stable Linear Discriminant Analysis (SLDA) and Logistic Linear Classifier (LOGLC). We also experimented on decision tree to check whether the performance of tree classifier is influenced by different SSR. From the experiment we see that for most of the datasets, the subbagging with stable classifiers in low SSR produces better performance than bagging and single stable classifiers, also in some cases better than double subbagging. We also found an opposite relation between the performance of double subbagging and subbagging.

Keywords: Subsample rate, Stable Classifiers, Subbagging, Double Subbagging.

1 Introduction and Notation

An important theoritical appraoch to analyzing the performance of combination of learning machines is through studying their stability. Various notions of stability of ensemble of machines (combination of machines) have been proposed in the last few years [2], [3], and have been used to study how combining machines can influence the generalization performance [5]. Numerous theoretical and empirical studies have been published to establish the advantages of combining learning machines, for example using Boosting or Bagging methods, [4] [9] very often leads to improved generalization performance, and a number of theoretical explanations have been proposed [5], [9].

[*] Corresponding author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 44–49, 2009.

Subbagging [6] is a particular ensemble architecture of bagging, consisting of a voting combination of a number of learning machines. Each machine is obtained by training the same underline learning algorithm, e.g. a decision tree, on a dataset drawn randomly with replacement from an initial training set with having some part of it (50%). In the double bagging framework proposed by Hothorn and Lausen [8], the out-of-bag sample is used to genrate an additional classifier model to integrate with the base learning model. In the setup of Hothorn and Lausen, the double-bagging uses the values of the linear discriminant functions trained on the out-of-bag sample as additional predictors for bagging classification trees only. The discriminant variables are computed for the bootstrap sample and a new classifier is constructed using the original variables as well as the discriminant variables. Using subbagging in double bagging rather than bagging, will result in larger out-of-bag sample for the additional classifier and if the additional classifier is a strong classifier it could result in an efficient ensemble method [10].

In this paper we have presented our experimental results, where we have constructed ensemble of three stable (linear) classifiers using different subsample rate (SSR) and check their accuracy in several benchmark datasets. We also compared their performance with aingle and bagged stable classifiers. In [7] authors analyzed the characteristics of subbagging. In particular, they developed probabilistic bounds on the distance between the empirical and the expected error that, unlike in the standard analysis of bagging are not asymptotic. These bounds formally suggest that subbagging can increase the stability of the learning machines when these are not stable and decrease otherwise. Keeping this in mind we develop our experiments. The results indicate that subbagging can significantly increase the stability of these machines whereas the test error is close (and sometimes better) to that of the same machines trained on the full training set. Another important practical consequence of our theoretical results is that, since subbagging improves the stability, it can be easier to control the generalization error. The paper is organized as follows: in Section 2 we have reviewed the theory of stability of subsampled ensemble. In Section 3, have presented the experiment and the results of the experiments with some discussion. It is followed by the Conclusion of the work.

2 Theory

Let $D_l = \{(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}\}_{i=1}^{l}$ be the training set. A learning algorithm is a function $\mathcal{A} : (\mathcal{X} \times \{-1, 1\})_{i=1}^{l} \to (\mathcal{X})^{\mathbb{R}}$ which give an training set D_l, returns a real valued classifier $f_{D_l} : \mathcal{X} \to \mathbb{R}$. Classification of a new input $\mathbf{x}$ is done by taking the sign of $f_{D_l}(\mathbf{x})$. Now we discuss the notion of β_l stability and relate it to the generalization performance of a subsampled ensemble based on [7]. The output of a larning machine on a new (test) point $\mathbf{x}$ should not change more than β_l when we train the machine with any training set of size l and when we train the machine with the same training set but one training point (any point) removed. Now let $F_{D_l}^{r} = E_{D_r}[f_{D_r}]$ be the expected combinations of machines trained on subsamples of size r. Next, we state a theorem modified from [7] to support our findings

Theorem 1. *For any given δ with probability $1 - \eta$ the generalization misclassification error of the expected combination $F^r_{D_l}$ of machines each using a subsample of size r of the training set and each having stability β_r is bounded by:*

$$R^{\delta}_{emp}(F^r_{D_l}) + \frac{4r}{l}\frac{\beta_r}{\delta} + \sqrt{\frac{1}{2l}\left(\frac{4r\beta_r}{\delta} + 1\right)^2 ln(\frac{1}{\eta})}$$

Thus we see from the theorem that the control of the test error by empirical error ($R^{\delta}_{emp}(F^r_{D_l})$) is manily due to the subsampling effect rather than the size of the ensemble. This indicates that increasing the number of base classifiers should not imporve this control as much as reducing the value of r. This we experimentally verified in Section 3. It should also be noted from Theorem 1 that if the stability of a subsampled machine $F^r_{D_l}$ is better than a single machine then the ensemble will give better bound otherwise the bound will be worse and subbagging should be avoided.

3 Experiments and Discussion of Results

We conducted a number of experiments using six datasets from UCI [1]: Diabetes, German Credit, Glass Identification, Cleveland Heart, Ionosphere and Iris. We focused mainly on the stable classifiers Linear Support Vector Machine

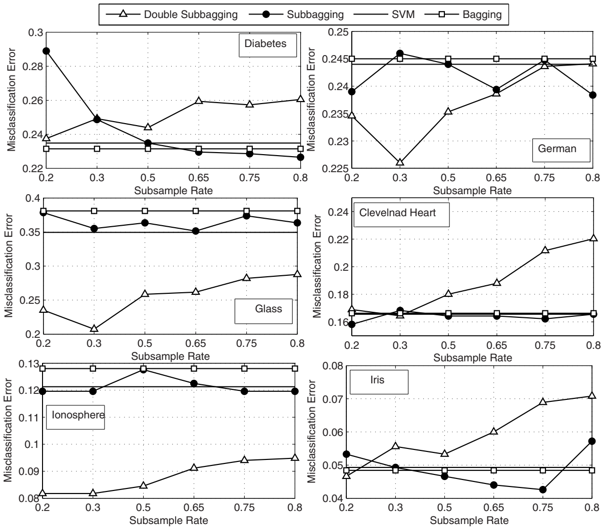


Fig. 1. Performance of *LSVM*, *Bagging LSVM*, *Subbagging LSVM* and *Double Subbagging LSVM* with different SSR. In X axis the SSR are given, in Y axis, average of 10 repititions of 10-fold cross-validation misclassification error is given.

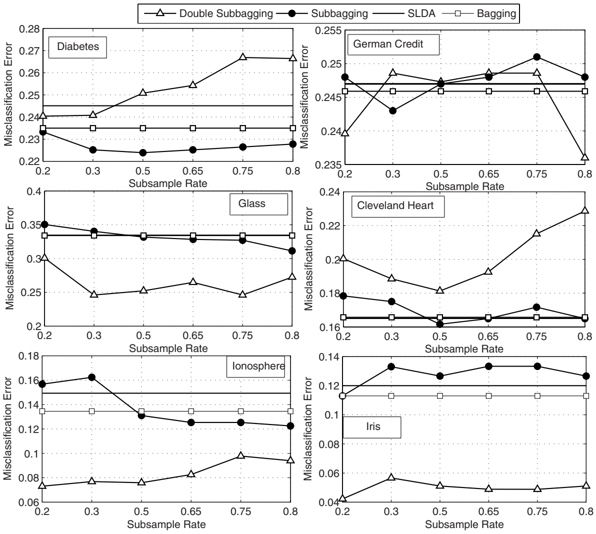


Fig. 2. Performance of *SLDA*, *Bagging SLDA*, *Subbagging SLDA* and *Double Subbagging SLDA* with different SSR. In X axis the SSR are given, in Y axis, average of 10 repititions of 10-fold cross-validation misclassification error is given.

(LSVM), Stable Linear Discriminant Analysis (SLDA) and Logistic Linear Classifier (LOGLC), we have also checked the performance of Classification and Regression Tree (CART) in different SSR. The first goal of the experiments was to check whether in subbagging with low SSR, the performance of the linear classifiers is better than single classifier and bagged version of the classifier. Secondly we wanted to check whether smaller SSR improve the accuracy of double subbagging with linear classifiers as the additional classifiers. Thirdly, we wanted to check the performance of CART with smaller SSR; whether with smaller SSR, it produce competent accuracy.

In all the experiment we have fixed the size of the ensemble size to 20, for all the classifiers. For LSVM, first we have optimized the cost parameter with 5-fold cross-validation. Then we use that parameter to create the ensembles. For each data set, ten stratified ten-fold cross-validation was performed.

3.1 Discussion of the Results

In case of LSVM we see that the performance of double subbagging is best among all the classifiers. In case of subbagging it showed superior performance with relatively low SSR in case most of the datasets. It showed better accuracy than bagging and single LSVM in all datasets except Ionpsphere dataset, (Fig. 1). This is becuase the test performance of the subbagged machines are bounded by the empirical error and its stability does not depend on the training size [7]. In case SLDA the scenario is nearly same, with double subbagging performing poorly

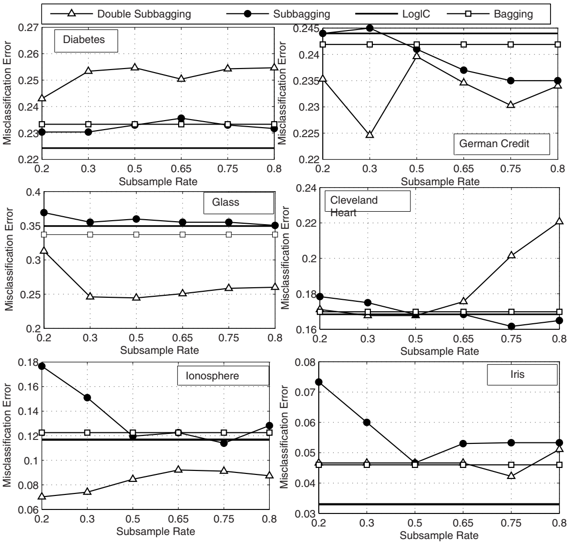


Fig. 3. Performance of *LOGLC*, *Bagging LOGLC*, *Subbagging LOGLC* and *Double Subbagging LOGLC* with different SSR. In X axis the SSR are given, in Y axis, average of 10 repititions of 10-fold cross-validation misclassification error is given.

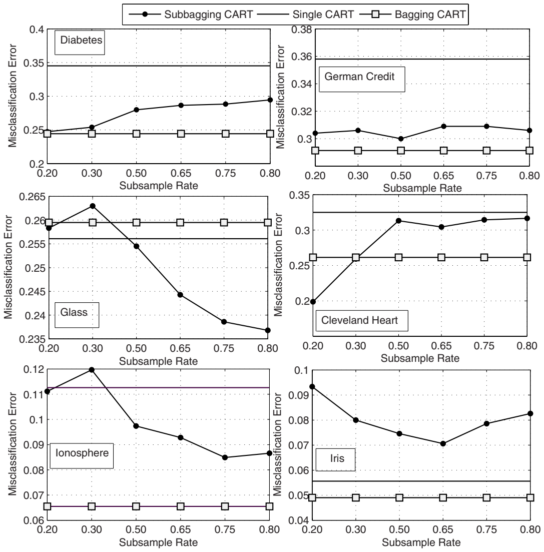


Fig. 4. Performance of *CART*, *Bagging CART*, *Subbagging CART* with different SSR. In X axis the SSR are given, in Y axis, average of 10 repititions of 10-fold cross-validation misclassification error is given.

in two datasets (Diabetes and Cleveland Heart), check Fig. 2. For LOGLC, the single LOGLC performed better than the ensemble of LOGLC in three datasets and it is aparent from the Fig. 3. In case of CART the performance of subbagging is satisfactory, as it showed competent performance with lower SSR in case of four datasets, than bagging CART in Fig. 4.

4 Conclusion

In this paper we have experimented with ensemble of several stable (linear) classifiers, where the ensembles are constructed with different subsampling rates. It is interesting to notice that ensemble with low SSR $(0.2 - 0.5)$ producing better accuracy with approximately all the linear classifiers, compared with bagging and single linear classifier. This implies that, one can get similar performance similar to a single classifier (machine), by training faster machines using smaller traning sets. This is an important practical advantage of ensembles of machines to reduce the computational time significantly, specially in the case of large datasets. In case of decision tree it also produced better accuracy with low SSR (0.2 – 0.5). The main direction which emerges from this work is that, time complexity of ensemble of machines can be reduced without deteriorating its accuracy.

References

1. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/mlearn/MLRepository.html
2. Bousquet, O., Elisseeff, A.: Stability and generalization. J. Mach. Lear. Res. 2, 499–526 (2002)
3. Elisseeff, A., Evegniou, T., Pontil, M.: Stability of Randomized Algorithm. J. Mach. Lear. Res. 6, 55–79 (2005)
4. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996a)
5. Breiman, L.: Heuristics of instability and stabilization in model selection. Annals of Statistics 24(6), 2350–2383 (1996c)
6. Bühlman, P.: Bagging, subbagging and bragging for improving some prediction algorithms. In: Arkitas, M.G., Politis, D.N. (eds.) Recent Advances and Trends in Nonparametric Statistics, pp. 9–34. Elsevier, Amsterdam (2003)
7. Evgeniou, T., Pontil, M., Elisseeff, A.: Leave one out error, stability, and generalization of voting combinations of classifiers (Preprint) (2001)
8. Hothorn, T., Lausen, B.: Double-bagging: combining classifiers by bootstrap aggregation. Pattern Recognition 36(6), 1303–1309 (2003)
9. Shapire, R., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics (1998)
10. Zaman, M.F., Hirose, H.: A New Double Bagging via the Support Vector Machine with Application to the Condition Diagnosis for the Electric Power Apparatus. In: International Conference on Data Mining and Applications (ICDMA 2009), pp. 654–660 (2009)

Metric in Feature Space

C.A. Murthy and Sourav Pradhan

Machine Intelligence Unit, Indian Statistical Institute, Kolkata- 700108, India
{murthy,sourav_r}@isical.ac.in

Abstract. In this paper our purpose is to define a metric in feature space. The metric finds the correlation distance between two features. It may be used in many applications. An application showing the utility of the proposed metric for feature selection is described here. The performance of the feature selection method is found to be comparable to other feature selection methods.

Keywords: metric, correlation, PCA, partition, entropy, regression.

1 Introduction

In pattern recognition literature, to evaluate the significance of features, many measures and their variations like distance measures [1], consistency measures [2], dependency measure [3], mutual information measures [4], classification error [5] have been introduced. But there are very few measures to find distance between two features. Mitra *et al.* [6] defined an index called maximal information compression index, which is like a distance measure between two features. This measure is invariant to translation but sensitive to scaling. In this present article we will define a distance measure which is invariant to both translation and scaling. This distance measure may be used effectively in several contexts. Here we have used it in feature selection.

2 Definitions and Basic Mathematics

This section provides the definitions and basic mathematics regarding features and metric. We shall assume that Ω is the space under consideration.

Definition 1 (Feature). *A feature X is a function from Ω to* $\mathbb{R}$.

$$X : \Omega \to \mathbb{R}.$$

Let S be the set of all features.

Definition 2 (Metric). *For a nonempty set M, ρ is said to be a metric on M if $\rho : M \times M \longrightarrow \mathbb{R}$ such that*

(i) $\rho(x, y) \geq 0 \qquad \forall x, y \in M$ (non-negativity)
(ii) $\rho(x, y) = 0 \qquad$ if and only if $x = y$

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 50–55, 2009.

(iii) $\rho(x, y) = \rho(y, x) \quad \forall x, y \in M$ (symmetry)
(iv) $\rho(x, z) \leq \rho(x, y) + \rho(y, z) \;\; \forall x, y, z \in M$ (triangular inequality)

If Ω is a finite set then we can have

$$l_m(X, Y) = \left(\sum_{w \in \Omega} |X(w) - Y(w)|^m \right)^{1/m} \qquad \forall m \geq 1; \quad X, Y \in S$$

If Ω is an uncountable set and integration is possible then we can have

$$L_m(X, Y) = \left(\int_{\Omega} |X(w) - Y(w)|^m dw \right)^{1/m} \qquad \forall m \geq 1$$

Note that l_m and L_m are metrics for every $m \geq 1$ and they are known as Minkowski metrics.

One can compute the distance between two features using these metrics but results may or may not be meaningful. For example, if the feature space is transformed by multiplying a feature vector by an arbitrary constant, then distance relationships in the transformed space can be very different from the original distance relationships, even though the transformation merely amounts to a different choice of units for the feature. Such scale changes can have a major impact on feature subset selection problem. We would like to make the distance function between two features to be translation and scale invariant. So we want that the following two properties be satisfied by the feature distance measure denoted by d.

P1. $d(X, aX + b) = 0 \qquad \forall a \in \mathbb{R}, \; a \neq 0$ *and* $\forall b \in \mathbb{R}$
P2. $d(X, Y) = d(aX + b, Y) \;\; \forall a \in \mathbb{R}, \; a \neq 0$ *and* $\forall b \in \mathbb{R}$

where X and Y are two features. First property says that features X and $aX + b$ are equivalent. Because feature X is scaled and translated to $aX + b$. As features X and $aX + b$ are equivalent then distances from Y to X and $aX + b$ must be same. For proper mathematical formulation of the above two properties let S be the set of all features. Let

$$S_X = \{aX + b : \quad a \in \mathbb{R}, \; a \neq 0 \text{ and, } \; b \in \mathbb{R}\}$$

Proposition. *For any two features X and Y either*

1. $S_X \cap S_Y = \emptyset$ *or*
2. $S_X = S_Y$

S_X's form a partition of S.
The proposition can easily be proved. The following property is also needed to be satisfied

P3. $d(X, Y) = d(X_1, Y_1)$ *if* $X_1 \in S_X$ *and* $Y_1 \in S_Y$

We would like to make the distance measure to be invariant to scaling and translation of the variables.

One property of any metric ρ is

$$\rho(x, y) = 0 \quad \text{if and only if} \quad x = y;$$

Note that d does not satisfy it since

$$d(X, aX + b) = 0 \quad \forall a \in \mathbb{R},\ a \neq 0 \text{ and, } \forall b \in \mathbb{R}$$

Thus to make proper mathematical formulation, let us assume that $\mathcal{A} = \{S_X \colon X \in S\}$. Note that $\mathcal{A}$ is the set of all partitions created by S_X. Let distance function D be defined on $\mathcal{A}$, i.e., $D : \mathcal{A} \times \mathcal{A} \longrightarrow [0, \infty)$. Let

$$D(S_X, S_Y) = d(X, Y)$$

where d satisfies properties ***P1***, ***P2***, and ***P3*** stated above. Then metric property (ii) is satisfied if d is defined properly. One of the quantities which is translation and scale invariant is $|r_{XY}|$ where

$$r_{XY} = \frac{\text{cov(X,Y)}}{\sqrt{\text{var(X)var(Y)}}}. \quad -1 \leq r \leq 1$$

var() denotes the variance of a variable and cov() the covariance between two variables. A possible way of defining metric is to make it a function of r. In fact $d(X, Y) = \sqrt{1 - r_{XY}}$ can be shown to satisfy triangular inequality. But $d(X, aX + b) \neq 0 \quad \forall a \in \mathbb{R}$ and, $a \neq 0$. Thus we have not considered the following distance measures.

(i) $d(X, Y) = \sqrt{1 - r_{XY}}$
(ii) $d(X, Y) = 1 - r_{XY}$

Some of the functions which don't depend on the sign of r are

(i) $d(X, Y) = \sqrt{1 - r_{XY}^2}$
(ii) $d(X, Y) = 1 - |r_{XY}|$

In this article we shall concentrate on these two distance measures. The above functions follow the properties ***P1***, ***P2***, and ***P3***. The functions also follow (i), (ii), (iii) properties of metric. But triangular inequality has not yet been proved. We have tested for triangular inequality property on 10 lakh data sets, each data set having 10 ordered triples of points, generated randomly from (0, 1). The triangular inequality is experimentally verified in each of these 10 lakh data sets for each of the above two d's. For both functions, namely, $d(X, Y) = \sqrt{1 - r_{XY}^2}$ and $d(X, Y) = 1 - |r_{XY}|$, the triangular inequality is experimentally seen to be true.

We have tested the utility of the proposed distance measures in feature selection problem. The algorithm for feature selection and the experimental results are given in next section.

3 Feature Selection Algorithm, Results and Comparison

Algorithm. Let the original number of features be N, and the original feature set be $O = \{F_i, i = 1, \cdots, N\}$. Represent the dissimilarity between the features F_i and F_j by the distance function $d(F_i, F_j)$. Higher the value of d is, the more dissimilar are the features. Let t_i^k represents the dissimilarity between feature F_i and its kth nearest neighbor feature in R. Then

Step 1: Choose an initial value of $k \leq (N-1)$. Set the number of iteration $iter \leq \lfloor \frac{N}{k} \rfloor$. Initialize the reduced feature subset R to the original feature set O, i.e., $R \longleftarrow O$.
Step 2: Compute $d(F_i, F_j) \quad \forall i, j = 1, \ldots, N$
Step 3: For each feature $F_i \in R$, compute t_i^k.
Step 4: Find the feature $F_{i'}$ for which $t_{i'}^k$ is minimum. Retain this feature in R and discard k nearest features of $F_{i'}$.
(Note: $F_{i'}$ denotes the feature for which removing k nearest neighbors will cause minimum error among all the features in R).

```
Step 5: If  $iter > 0$
              $iter = iter - 1$;
              Go to Step 3.
        else
             Return feature set $R$ as the reduced feature set.
             Stop.
```

Results and Comparison. For our experiment we have used 6 real-life public domain data sets [7] from three different categories: low-dimensional ($D \leq 10$) (Iris, Wisconsin Breast Cancer), medium-dimensional ($10 < D \leq 100$) (Ionosphere, Spambase), and high-dimensional ($D > 100$) (Multiple features, Isolet), containing both large and relatively smaller number of points.

Four indices, namely, entropy (E) [1], fuzzy feature evaluation index (FFEI) [6], class separability (S) [5], and K-Nearest Neighbor (K-NN) accuracy [6] are considered to demonstrate the efficiency of the proposed methodology and for comparing it with other methods. Representation entropy (RE) [5] is used to show the redundancy present in reduced feature subset. Five feature selection schemes considered for comparison are branch and bound (BB) [5], sequential forward search(SFS) [5], sequential floating forward search (SFFS) [1], and stepwise clustering (SWC)(using correlation coefficient) [6], relief-F [1]. We have got exactly same result for each data set for the two proposed distance functions. The performance of proposed method is also compared with other two feature similarities least square regression error [6] and, maximal information compression index [6] using our algorithm. In our experiment, we have used interclass distance [5] measure as the feature selection criterion for BB, SFS, SFFS algorithms. Following table provide such a comparative result corresponding to high, medium, and low-dimensional data sets when the size of the reduced feature subset is taken to be about half of the original size. Since the branch and bound (BB) and sequential floating forward search (SFFS) algorithms require infeasible

Table 1. Comparison of Feature Selection Algorithms for real-life data sets

Data Set	Method	RE	E	FFEI	S	KNN Acuuracy	
						Mean (%)	SD (%)
Isolet N=617 d=309 c=26 k=7	$\sqrt{1-r_{XY}^2}$	5.99	0.87	0.42	0.07	77.73	1.26
	Least Square Regression error	5.40	0.88	0.42	0.05	75.75	1.17
	Maximal Information Compression Index	5.12	0.88	0.42	0.05	70.55	1.29
	SWC	5.98	0.87	0.42	0.06	78.25	1.22
	Relief-F	4.68	0.90	0.43	0.03	80.32	0.93
	SFS	4.73	0.92	0.44	0.03	74.45	1.20
Multiple Features N=649 d=324 c=10 k=6	$\sqrt{1-r_{XY}^2}$	3.83	0.01	0.38	0.10	87.87	0.01
	Least Square Regression Error	3.75	0.02	0.38	0.10	87.12	0.01
	Maximal Information Compression Index	3.2	0.02	0.37	0.08	81.54	0.02
	SWC	0.24	0.02	0.33	0.02	72.62	0.01
	Relief-F	3.36	0.02	0.38	0.08	86.35	0.01
	SFS	0.61	0.02	0.33	0.02	79.35	0.02
Spambase N=57 d=27 c=2 k=6	$\sqrt{1-r_{XY}^2}$	0.01	0.01	0.03	9.34	75.68	0.79
	Least Square Regression error	0.39	0.01	0.04	7.10	70.64	0.92
	Maximal Information Compression Index	0.03	0.01	0.04	6.92	68.60	0.59
	SWC	0.01	0.01	0.03	9.34	76.40	1.05
	Relief-F	0.15	0.01	0.03	9.53	79.08	0.90
	SFS	0.39	0.01	0.04	7.10	70.73	0.77
	SFFS	0.39	0.01	0.04	7.10	70.73	0.77
	BB	0.38	0.01	0.03	7.11	70.93	0.70
Ionosphere N=32 d=14 c=2 k=6	$\sqrt{1-r_{XY}^2}$	3.36	0.78	0.41	5.65	73.75	5.46
	Least Square Regression Error	3.21	0.80	0.42	30.73	70.60	4.58
	Maximal Information compression Index	3.31	0.80	0.41	12.16	70.60	4.58
	SWC	3.38	0.78	0.41	11.59	71.39	5.27
	Relief-F	3.18	0.81	0.42	8.21	71.58	6.36
	SFS	2.68	0.82	0.42	11.35	70.73	7.23
	SFFS	2.68	0.82	0.42	11.35	70.73	7.23
	BB	3.18	0.81	0.42	11.20	72.14	7.50
Cancer N=9 d=5 c=2 k=4	$\sqrt{1-r_{XY}^2}$	1.46	0.37	0.35	0.31	94.32	1.90
	Least Square	1.51	0.37	0.36	0.34	94.25	0.89
	Regression error						
	Maximal Information Compression Index	1.53	0.38	0.35	0.33	94.19	1.50
	SWC	1.46	0.37	0.35	0.31	95.03	0.94
	Relief-F	1.44	0.39	0.35	0.29	95.57	0.54
	SFS	1.23	0.36	0.33	0.26	95.65	0.47
	SFFS	1.23	0.36	0.33	0.26	95.65	0.47
	BB	1.23	0.36	0.33	0.26	95.65	0.47
Iris N=4 d=2 c=3 k=2	$\sqrt{1-r_{XY}^2}$	0.27	0.35	0.56	0.03	93.48	2.03
	Least Square	0.23	0.57	0.35	0.04	92.29	2.57
	Regression Error						
	Maximal Information Compression Index	0.73	0.58	0.37	0.09	93.03	2.50
	SWC	0.27	0.56	0.35	0.03	93.48	2.03
	Relief-F	0.08	0.58	0.34	0.02	94.44	3.71
	SFS	0.23	0.57	0.35	0.05	92.29	2.57
	SFFS	0.23	0.57	0.35	0.05	92.29	2.57
	BB	0.23	0.57	0.35	0.05	92.29	2.57

high computation time for the large data sets, we could not provide the results for them in the table. For the classification accuracy using K-NN, both the mean and standard deviations (SD) computed for ten independent runs are presented. In relating to the search-based algorithms (BB, SFFS, and SFS) and two feature similarities (least square regression error and maximal information compression index), the performance of the proposed method is comparable or sometimes better. In a part of the experiment, we compared the performance with a supervised method Relief-F, which is widely used. We have used 50 percent of the samples as design set for the Relief-F algorithm. Sometimes, the Relief-F algorithm provides classification performance better than our method but its performance in terms of the unsupervised indices is poor. Relief-F has a higher time requirement, specially for data sets with large number of samples because the number of iterations is huge. The performance of our method is better, in terms of unsupervised indices, than least square regression error and maximal information compression index. Representation entropy performance for proposed method is comparable to other methods but better for large dimensional data sets.

4 Conclusion

A distance measure, invariant to translation and scaling, is proposed for feature subset selection. This measure is fast to compute as computational complexity for finding correlation between two variables is linear. The performance of this distance measure using the algorithm described in this paper is comparable to other methods, generally better for unsupervised case.

Acknowledgement

The authors wish to thank Mr. B. L. Narayan, Yahoo India, Bangalore for his helping in this work.

References

1. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman and Hall/Crc Data Mining and Knowledge Discovery Series (2007)
2. Dash, M., Liu, H.: Consistency-based search in feature selection. Artif. Intell. 151(1-2), 155–176 (2003)
3. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng. 17(4), 491–502 (2005)
4. Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M.: Normalized mutual information feature selection. IEEE Trans. on Neural Networks 20(2), 189–201 (2009)
5. Devijver, P., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Engle Cliffs (1982)
6. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised Feature Selection Using Feature Similarity. IEEE Trans. Pattern Analysis and Machine Intelligence 24(3), 301–312 (2002)
7. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Database, Univ. of California, Irvine, Dept. of Information and Computer Science (1998)

Speeding-Up the K-Means Clustering Method: A Prototype Based Approach

T. Hitendra Sarma and P. Viswanath

Pattern Recognition Research Laboratory
NRI Institute of Tehnology, Guntur-522009, A.P., India
{t.hitendrasarma,viswanath.pulabaigari}@gmail.com

Abstract. The paper is about speeding-up the k-means clustering method which processes the data in a faster pace, but produces the same clustering result as the k-means method. We present a prototype based method for this where prototypes are derived using the leaders clustering method. Along with prototypes called leaders some additional information is also preserved which enables in deriving the k means. Experimental study is done to compare the proposed method with recent similar methods which are mainly based on building an index over the data-set.

1 Introduction

k-means clustering method, for a given data-set, finds k patterns called the k centers (or k means) which represents the k clusters. The centers need to be found such that the sum of squared distances from each data point to its nearest center is minimized. An iterative procedure to find the k centers is given by Lloyd [4] and this is what conventionally called *the k-means clustering algorithm*. There are several approximate methods like *single pass k-means method* [2] which scans the data-set only once to produce the clustering result. Other approximate methods are to keep important points in the buffer (primary memory) while discarding unimportant points as done by Bradley *et. al.* [1]. There are other algorithmic approaches which speeds-up the k-means method without compromising with the the quality of the final result. These methods are primarily based on building an index like data structure over the data-set which speeds-up the nearest neighbor finding process. Using a data structure similar to *kd-tree* to speed-up the process was given by Kanungo et al. [3] who gave a *filtering algorithm* to reduce the number of centers to be searched for a set of points which are enclosed in a hyper-rectangle. It proposes a variation of the kd-tree called *balanced box-decomposition tree (BBD-tree)* and under certain favoring conditions, like clusters being well separated, proves that the total running time of the k-means algorithm using their filtering approach is $O(dn \log n + 2^d mk \log n)$, where n is the data-set size, d is the dimensionality of data, m is the number of iterations taken by the k-means algorithm, and k is the number of clusters derived. Clearly, the filtering approach is not a good choice for high dimensional data-sets.

This paper proposes to use only a few selected prototypes from the data-set along with some additional information. Prototypes are selected by using

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 56–61, 2009.

a fast clustering method called the *leaders clustering method* [5]. Along with leaders additional information like number of patterns and linear sum of patterns that are grouped with a leader are preserved. Recently this kind of approach is adopted to devise approximate hybrid density based clustering methods like *rough*-DBSCAN [7] and to improve prototype based classification methods like rough fuzzy weighted k-nearest leader classifier [6].

The proposed method called *leader-k-means (lk-means)* clustering method runs in two stages. In the first stage it applies the k-means method over the leaders set. The second stage checks for correctness of the results and if needed applies a correcting step to guarantee that the final clustering result is same as that would have obtained if the k-means method is applied over the entire data-set.

The paper is organized as follows. Section 2 briefly reviews the k-means clustering method while Section 3 reviews the leaders clustering method along with a modified leaders method called *modified-leaders* to derive leaders along with the number of patterns that are grouped with it, etc. The proposed hybrid method *lk*-means clustering method is described in Section 4. Experimental results are given in Section 5 and Section 6 gives some of the conclusions.

2 K-Means Clustering Algorithm

k-means clustering method is a partition based method and each cluster is represented by its centroid (mean). Let $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ be the data-set of dimensionality d. The means has to be found such that the criterion $J = \sum_{i=1}^{n} ||x_i - m(x_i)||^2$ is minimized, where $m(x_i)$ is the nearest mean to x_i. The iterative procedure given by Lloyd [4] to find the k means is given in the Algorithm 1.

Algorithm 1. K-means($\mathcal{D}$, k)

```
1. Randomly choose $k$ patterns from $\mathcal{D}$. Let this be $M^{(0)} = \{m_1^{(0)}, m_2^{(0)}, \ldots, m_k^{(0)}\}$.
2. Let $i = 0$; /* $i$ is the iteration number */
repeat
    3. $i = i + 1$;
    4. Form $k$ clusters by assigning each pattern in $\mathcal{D}$ to its nearest mean in $M^{(i-1)}$.
    5. Find new centroids (means) of the $k$ clusters, i.e., $M^{(i)} = \{m_1^{(i)}, m_2^{(i)}, \ldots, m_k^{(i)}\}$.
until ($M^{(i)} == M^{(i-1)}$)
6. Output $M^{(i)}$.
```

3 Leaders Clustering Method

In leaders clustering method each cluster is represented by a pattern called *leader* and all other patterns in the cluster are its *followers*. For a given threshold distance τ, leaders method maintains a set of leaders $\mathcal{L}$, which is initially empty and is incrementally built. For each pattern x in the data set $\mathcal{D}$, if there is a leader $l \in \mathcal{L}$ such that distance between x and l is less than or equal to τ, then x is assigned to the cluster represented by l. Otherwise x itself becomes new leader and is added to $\mathcal{L}$. The algorithm outputs the set of leaders $\mathcal{L}$.

The leaders method is modified in-order to be used with k-means method. The modifications are, (i) to replace a leader by its cluster centroid (so, a leader is no more a pattern in the data-set, but all its followers are patterns from the data-set), (ii) to store along with leaders, a count, which is the number of patterns that are present in its cluster, and (iii) linear sum of all followers. The modified leaders method is given in Algorithm 2. The output of the method is leaders set along with count, linear sum, and also the data-set rearranged and stored according to the clusters.

Algorithm 2. Modified-Leaders($\mathcal{D}$, τ)

```
$\mathcal{L} = \emptyset$;
**for** each $x \in \mathcal{D}$ **do**
  Find a $l \in \mathcal{L}$ such that $||l - x|| \leq \tau$
  **if** there is no such $l$ or when $\mathcal{L} = \emptyset$ **then**
    $\mathcal{L} = \mathcal{L} \cup \{x\}$;
    $count(x) = 1$;
    $linear\text{-}sum(x) = x$;
    $followers(x) = \{x\}$;
  **else**
    $count(l) = count(l) + 1$;
    $linear\text{-}sum(l) = linear\text{-}sum(l) + x$;
    $followers(l) = followers(l) \cup \{x\}$;
  **end if**
**end for**
**for** each $l \in \mathcal{L}$ **do**
  Replace $l$ by centroid of its cluster, *i.e.,*
      $$l = \frac{linear\text{-}sum(l)}{count(l)};$$
**end for**
Output:
$\mathcal{L}^* = \{< l, count(l), linear\text{-}sum(l), followers(l) > | l \text{ is a leader } \}$.
```

4 *lk*-Means: Leader Based K-Means Clustering Method

The proposed method *lk*-means clustering method runs in two stages. The first stage called *lk*-means-*first-stage* basically runs the Lloyd's k-means method, but using leaders set, their count and linear-sum values. The method is iterated till it converges. The second stage is called *lk*-means-*correcting-stage* which checks for correctness of the results and if needed applies the correcting step so that the final result is same as that obtained by applying(including itself) the Lloyd's algorithm using the entire data-set.

The first step called *lk*-means-*first-stage* is same as Lloyd's k-means as given in Algorithm 1, but is applied using the leaders set (instead of the entire data-set). The method starts with initially chosen random patterns from the data-set as its seed-points, and in each iteration, each leader is assigned to the nearest mean pattern to form the k clusters of leaders. The new means (*i.e.,* the new centroids) of each cluster (cluster of leaders) is found as explained. Let $l_1, l_2, \ldots, l_p$ be the leaders in a cluster. Then its new centroid is $\frac{\sum_{j=1}^{p} \textit{linear-sum}(l_j)}{\sum_{j=1}^{p} \textit{count}(l_j)}$. The method is iterated till it converges. The clustering result of this stage consists clusters of leaders. By replacing each leader by the set of its followers, we get a partition of the data-set $\mathcal{D}$. Let this partition be π^l, and that obtained by employing Algorithm 1 (*i.e.,* the original k-means method) over the entire data-set (keeping the seed points same) be π. Now, π^l need not be same as π, because there may be a leader l which is assigned to a mean m_i (according to *lk*-means-*first-stage*), but a follower of l may be actually closer to some other mean m_j such that $m_i \neq m_j$. Let the follower be x_f. So, this situation arises when $||l - m_i|| < ||l - m_j||$, but $||x_f - m_i|| > ||x_f - m_j||$. The pattern x_f according to the original k-means

Algorithm 3. *lk*-means-*correcting-stage*$(M^{(0)})$

```
/* $M^{(0)} = \{m_1^{(0)}, m_2^{(0)}, \ldots, m_k^{(0)}\}$ is the set of means that is obtained as output of
lk-means-first-stage */
1. Assign each leader to its nearest mean in $M^{(0)}$ to obtain a partition of the set of
leaders which is $\pi_{\mathcal{L}} = \{L_1, L_2, \ldots, L_k\}$.
2. For $j = 1$ to $k$, find $\textit{Total-Sum}(L_j) = \sum_{\forall l \in L_j} \textit{linear-sum}(l)$, and $\textit{Total-Count}(L_j)$
$= \sum_{\forall l \in B_j} \textit{count}(l)$;
3. $i = 0$.
repeat
  4. $i = i + 1$;
  5. Find the set of border leaders.
  for each border leader $l$ do
    6. Find nearest mean for $l$ in $M^{(i-1)}$. Let this be $m_l^{(i-1)}$ which corresponds to
    the cluster $L_l$.
    7. $\textit{Total-Sum}(L_l) = \textit{Total-Sum}(L_l) - \textit{linear-sum}(l)$;
    8. $\textit{Total-Count}(L_l) = \textit{Total-Count}(L_l) - \textit{count}(l)$;
    for each $x \in \textit{follower}(l)$ do
      9. Find the nearest mean of $x$ in $M^{(i-1)}$. Let this be $m_x^{(i-1)}$ and the corre-
      sponding cluster be $L_x$;
      10. $\textit{Total-Sum}(L_x) = \textit{Total-Sum}(L_x) + x$;
      11. $\textit{Total-Count}(L_x) = \textit{Total-Count}(L_x) + 1$;
    end for
  end for
  12. Find new means,
  $M^{(i)} = \{\textit{Total-Sum}(L_j)/\textit{Total-Count}(L_j) \mid j = 1 \text{ to } k\}$;
until $(M^{(i)} == M^{(i-1)})$
13. Output $M^{(i)}$.
```

method should be included with the cluster for which m_j is the mean, but *lk*-means-*first-stage* assigned this to the mean m_i. Such leaders are named as *boarder-leaders*. The set of such border-leaders can be found as part of last iteration of the method *lk*-means-*first-stage*. The second stage is called *lk*-means-*correcting-stage*. If the set of border leaders is non-empty, then π^l and π need not be same. A correcting step is applied over the result of *lk*-means-*first-stage* in order to get the same clustering result as that would have obtained by using the original Lloyd's algorithm (Algorithm 1). Only the followers of border leaders are reconsidered in the correcting stage as given in Algorithm 3. Each border leader l is removed from its cluster and each follower of this leader l is reconsidered as an individual pattern and assigned to its nearest mean. The process is repeated till convergence.

5 Experimental Results

Experiments are done with (i) the Pendigits data-set available at the UCI Machine Learning Repository, and (ii) a series of synthetic data-sets of varying dimensionality. 39 different synthetic data-sets of dimensionality ranging from 2 to 40 are generated as follows. Each data-set has 60000 patterns. Each data-set is generated from a tri-modal Gaussian distribution $p(x) = \frac{1}{3}N(\mu_1, \Sigma_1) + \frac{1}{3}N(\mu_2, \Sigma_2) + \frac{1}{3}N(\mu_3, \Sigma_3)$. For dimensionality 2, $\mu_1 = (0, 0), \mu_2 = (5, 5), \mu_3 = (-5, 5)$; for dimensionality 3, $\mu_1 = (0, 0, 0), \mu_2 = (5, 5, 5), \mu_3 = (-5, 5, -5)$; and so on upto dimensionality 40. In Each case the covariance matrix is taken as the Identity matrix of size $d \times d$, where d is the dimensionality of the data.

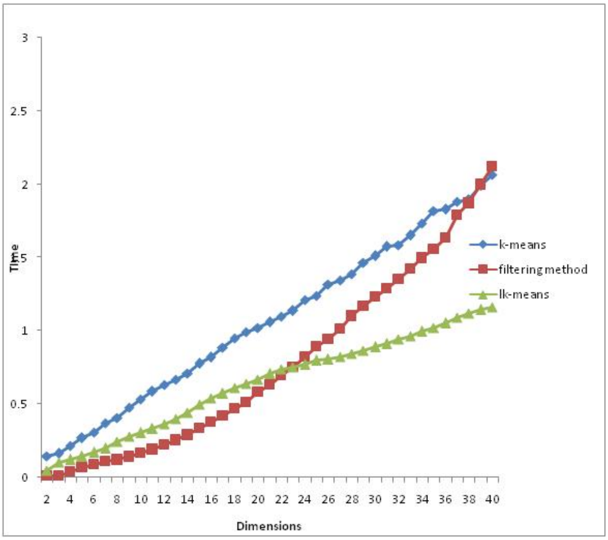


Fig. 1. Comparison of time taken by various methods

For Pendigits data-set k value taken is 10, and for synthetic data-sets it is 3. The parameter τ used to generate the set of leaders is taken as 5% of the average distance between a pair of distinct random patterns over 10 random trials.

For Pendigits data-set, the k-means method (Algorithm 1) has taken 0.463 seconds, whereas the filtering approach has taken time equal to 0.326 seconds. But the proposed method has taken time equal to 0.294 seconds only. This is because, for the Pendigits data, its dimensionality is 16, and filtering approach can work well only for low dimensional data-sets. This point is amplified by the synthetic data-sets of varying dimensionality, which clearly shows that the filtering approach can reduce the time requirement considerably for low dimensional data-sets, but fails to do so for high dimensional data-sets. See Figure 1.

6 Conclusions

The paper presented a hybrid method to speed-up the k-means clustering method. The proposed method scales well for large and high dimensional data-sets, in contrast to other approaches which are not suitable to work with high dimensional data-sets. The proposed method is a prototype based one, where the prototypes called leaders are derived from the data-set using the leaders clustering method.

References

1. Bradley, P.S., Fayyad, U., Raina, C.: Scaling clustering algorithms to large databases. In: Proceedings of Fourth International Conference on Knowledge Discovery and Data Mining, pp. 9–15. AAAI Press, Menlo Park (1998)
2. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)
3. Kanungo, T., Mount, D.M., Netanyahu, N.S.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 881–892 (2002)
4. Lloyd, S.P.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28, 129–137 (1982)
5. Spath, H.: Cluster Analysis Algorithms for Data Reduction and Classification. Ellis Horwood, Chichester (1980)
6. Babu, V.S., Viswanath, P.: Rough-fuzzy weighted k-nearest leader classifier for large data sets. Pattern Recognition 42, 1719–1731 (2009)
7. Viswanath, P., Suresh Babu, V.: Rough-DBSCAN: A fast hybrid density based clustering mehtod for large data sets. Pattern Recognition latters (2009) doi:10.1016/j.patrec.2009.08.008

Constructive Semi-Supervised Classification Algorithm and Its Implement in Data Mining

Arvind Singh Chandel[1], Aruna Tiwari[1], and Narendra S. Chaudhari[2]

[1] Department of Computer Engg.
Shri GS Inst of Tech.& Sci.
SGSITS, 23, Park Road,
Indore (M.P.) India 452001
atiwari@sgsits.ac.in, asc_chandel@rediffmail.com
[2] Department of Computer Science and Engineering (CSE)
IIT, Indore
M-Block, IET-DAVV Campus,
Khandwa Road, Indore-452017(M.P.)
NSchaudhari@iitbombay.org, NSC183@gmail.com

Abstract. In this paper, we propose a novel fast training algorithm called Constructive Semi-Supervised Classification Algorithm (CS-SCA) for neural network construction based on the concept of geometrical expansion. Parameters are updated according to the geometrical location of the training samples in the input space, and each sample in the training set is learned only once. It's a semi-supervised based approach, the training samples are semi-labeled i.e. for some samples, labels are known and for some samples, data labels are not known. The method starts with clustering, which is done by using the concept of geometrical expansion. In clustering process various clusters are formed. The clusters are visualizes in terms of hyperspheres. Once clustering process over labeling of hyperspheres is done, in which class is assigned to each hypersphere for classifying the multi-dimensional data. This constructive learning avoids blind selection of neural network structure. The method proposes here is exhaustively tested with different benchmark datasets and it is found that, on increasing value of training parameters number of hidden neurons and training time both are getting decrease. Through our experimental work we conclude that CS-SCA result in simple neural network structure by less training time.

Keywords: Semisupervised classification, Geometrical Expansion, Binary Neural Network, Hyperspheres.

1 Introduction

Constructive learning begins with a minimal or empty structure, and dramatically increases the network by adding hidden neurons until a satisfactory solution is found. Numbers of Constructive learning algorithms are available to overcome the problem of traditional algorithms for classification. It includes Fast Covering Learning Algorithm (FCLA) for Supervised Learning [2], Constructive Unsupervised Learning

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 62–67, 2009.

Algorithm (CULA) [1], Constructive set Covering learning algorithm (CSCLA) by Ma and Yang [4], Boolean-like training algorithm (BLTA) by Gray and Michel [5], expand and truncate learning algorithm (ETL) by Kim and Park [3]. BLTA is a dynamic technique and derives its original principle from Boolean algebra with extension. ETL finds a set of required separating hyperplanes and automatically determines a required number of neurons in the hidden layer based on geometrical analysis of the training set. CSCLA was proposed based on the idea of weighted Hamming distance hypersphere. In general, ETL, IETL, CSCLA, have no generalization capability. BLTA has generalization capability, but needs more hidden neurons.

Moreover to it, FCLA is supervised learning algorithm, thus labeled samples are used for learning but labeled data sample are expensive to obtain as they require the effort of experienced human annotators. On the other hand unlabeled data samples are easy to obtain but there are very few way to process them. Thus approach of semi-supervised was used that uses large amount of unlabeled data sample with small amount of labeled data sample to build the classifier.

In this paper, we propose a novel fast training algorithm called Constructive semi-supervised classification Approach (CS-SCA) for neural network construction. The proposed method is implemented using two processes, first is clustering and second is labeling. We illustrate the advantages of CS-SCA by using it in classification problems. There are various features of CS-SCA like it is a semi-supervised constructive approach. Sample reordering is allowed in proposed classifier and because of reordering, learning is fast in this approach. As we know that CS-SCA is a semi-supervised approach that's why it requires less human effort. This CS-SCA approach is tested with number of benchmark datasets and compared with SVM [6] based classifier.

The paper is organized as follows. Section 2 gives an overview of CS-SCA. Section 3 explains the method for CS-SCA in detail give algorithmic formulation of our methodology. In section 4, we give experimental results to demonstrate the usefulness of our approach; it also contains detail of data preparation. These experimental results include two well-known datasets [7], namely, Riply dataset and Wisconsin Breast cancer dataset. Finally, in section 5, we give concluding remarks.

2 Overview of CS-SCA

2.1 Basic Concept

Boolean functions have the geometrical property which makes it possible to transform non-linear representation to linear representation for each hidden neuron. We consider a Boolean function with n input and one output,

$$y = f(x_1, x_2, ..., x_n),$$

$$\text{Where } y \in (0,1) \text{ and } x_i \in (0,1), i = (1.....n).$$

These 2^n binary patterns $(0,1)^n$ can be considered as a $n-$dimensional unit hypercube.This ex-hypersphere is defined as the reference hypersphere (RHS) [5] as follows:

$$(x_1 - 1/2)^2 + (x_2 - 1/2)^2 + ... + (x_n - 1/2)^2 = n/4. \quad (1)$$

2.2 Network Construction

CS-SCA constructs a three-layered feed forward neural network with an input layer, a hidden layer and an output layer, as shown in Fig-1. We illustrate the advantages of CS-SCA by its implement in classification problems.

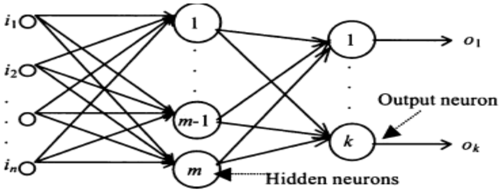


Fig. 1. Neural Network Structure by CS-SCA

3 Proposed Method: CS-SCA

CS-SCA begins with an empty hidden layer. To learn a sample, CS-SCA either adds a new hidden neuron to represent it or updates the parameters of some hidden neuron by expanding its corresponding hypersphere. This is done by clustering process and once clustering gets over by using the concept of majority voting labeling of hypersphere is done.

3.1 Clustering Process

CS-SCA constructs a three-layered feed forward neural network, of which first layer represent to the input data sample that will be in the binary coded format. Then input data samples will be grouped into various clusters. The middle layer of network architecture represents the hyperspheres(hidden neuron).A hidden neuron in Fig. 1 represents a corresponding hyper sphere with center c and radius r_1 . While constructing a hidden neuron, suppose that $\{x^1, x^2, ..., x^v\}$ are v (true) sample included in one hyper sphere (hidden neuron). In terms of these samples, the center is defined as the gravity center $c = (c_1, c_2, ..., c_n)$;

$$c_i = \sum_{k=1}^{v} \frac{x_i^k}{v} \tag{2}$$

The radius r_1 is defined as the minimal Euclidean distance such that all the v vertices are exactly in or on the surface of the corresponding hyper sphere.

$$r_1 = \min_{j=1}^{v} || x^j - c || = \min_{j=1}^{v} (\sum_{i=1}^{n} (x_i^k - c_i)^2)^{1/2}$$

Where n is the dimension of the input and ||*|| is the euclidean distance.

Given c and r_1 we can separate these v true sample from the remaining samples. In another words, this corresponding hypersphere represents these v true samples.

Two secondary central radii r_2 and r_3 are introduced to find compact cluster. Samples should be a compact cluster where $r_1 < r_2 < r_3$.

CS-SCA begins with an empty hidden layer. To construct the neural network, we examine whether a coming "true" sample can be covered by one of the existing hidden neurons. When the first sample x^1 comes, the hidden layer is empty and no hidden neuron covers this sample. A new hidden neuron, the first hidden neuron, is created to represent it. This new created hidden neuron represents a hyper sphere centered at x^1. Samples, which have been represented, are removed after parameter updating. The training process goes on. A coming sample x^k causes one of the following actions.

1. Update the parameters of some hidden neuron, and remove c;
2. Create a new hidden neuron to represent it, and remove x^k ;
3. Back up x^k to be learned in the next training circle.

Given a hidden neuron j with the center c^j and three radii r_1^j , r_2^j and r_3^j , we firstly compute the function for the hidden neuron j defined as:

$$f(w^j, x^k) = \sum_{i=1}^{n} w_i^j x_i^k \qquad (3)$$

Where w^j is ($w_1^j, w_2^j, ..., w_n^j$), the weight vector and x^k is the k^{th} vertex. The training process is continued as follows:

I. If $f(w^j, x^k) \geq t_1^j$, already covered, so nothing needs to be done.

II. If $t_1^j > f(w^j, x^k) \geq t_2^j$ the sample x is within the "claim region"; so to include it an immediate expansion is needed.

III. If $f(w^j, x^k) \geq t_3^j$ the sample is confusing sample so back up x^k to be dealt with in the next training circle.

IV. If for all j's, $f(w^j, x^k) < t_3^j$ then create a new hidden neuron and remove the sample x^k from the training set.

Thus the number of neurons generated is equal to the number of clusters. After this, the labeled samples are useful for labeling the clusters. The details are given next.

3.2 Labeling Process

In labeling process labels are assigned to hyperspheres formed after the clustering process by using the mechanism of Majority voting concept. Thus these labeled hypersphere can be represented as output neuron in the output layer of network architecture. After clustering when hyperspheres are identified, we assign labels to hyperspheres.

1. Repeat the step 2 and step 3 for each of the hyper sphere.
2. Perform majority voting by count number of samples belongs to one particular class.

3. Majority of samples of particular class in an individual hypersphere would decide the class of that hypersphere.
4. If a particular hypersphere is not covering any labeled data in that case merge this hypersphere with other which is closure to it.

4 Experimental Work

We used a Personal Computer (PC) with Pentium processor with 2.99 GHz speed and 1GB of RAM having windows XP operating system for testing. We used Matlab 7.0.1 for implementation.

Table 1.

Dataset	Dimensions, Number of classes	Training	Testing
Fisher's Iris	4, 3	123	27
Breast Cancer	9,2	683	68
Balance Scale	4,3	492	76
Riply	2,2	720	80

Each training samples $x^k = (x_1^k, x_2^k, \ldots, x_n^k)$ are normalized as follows:

$$x_i^k = x_i^k - \min(x_i) / \max(x_i) - \min(x_i) \tag{4}$$

After this transformation, each data sample is transformed in the range 0 and 1. CS-SCA requires binary form of input data therefore after normalization re-quantizes the data into eight levels as follows.

1. Apply each sample as an input in quantized function given in step 3.
2. Quantized value can be obtained by: $y = uencode\ (u, n, v)$
3. Repeat step 3, till the whole sample binary coded

After data preparation, for experimentation 80% of the original data taken as training data and rest 20% considered as testing samples. The datasets used for experimentation are given in table 1. Results are evaluated in terms of classification accuracy, training time, confusing samples and number of hyperspheres required. For different value of training parameter results for each dataset are getting change. After calculating the performance of CS-SCA, same datasets are applied in SVM based classifier [6], to compare the performance of both the classifiers, in terms of Classification accuracy, Training time. In SVM based classifier, training parameter α used in clustering process. Number of support vector in SVM based classifier depends on the value of training parameter α . Comparison results of both the classifier are displayed in Table3.

Table 2. For 10-fold cross validation results

Dataset	Average Accuracy
Wisconsin Beast Cancer	85.1 %
Riply	80.1 %

Table 3. Comparison with SVM

Dataset	Accuracy by CS-SCA	Accuracy in SVM	Training Time in CS-SCA	Training Time in SVM
Fisher's Iris	92.59 %	77 %	0.96 sec.	5.89 sec.
Balance Scale	80.26 %	77 %	2.18 sec.	65.9 sec.
Wisconsin breast cancer	85 %	70 %	4.29 sec.	1086 sec.
Riply	80.1 %	75 %	3.98 sec.	36.12 sec.

We give results for 10-fold cross validation on Wisconsin Breast Cancer and Riply dataset in table2 shown above. For 10-fold cross validation 90% of the data taken as training and rest 10% taken as testing data. From the results shown in above table3, it's clear that for each dataset CS-SCA is giving better accuracy and requires less training time compare to SVM based classifier.

5 Concluding Remarks

A binary neural network based Semi-supervised classifier is constructed using the concept of geometrical expansion, which classify semi-labeled data. The classification is performed using two processes, first is clustering and second is labeling. Various benchmark datasets used to demonstrate the performance of CS-SCA in terms of accuracy and number of hypersphere etc. After that same datasets is applied in SVM based classifier to compare its performance with developed classifier. It's found that CS-SCA gives better performance in terms of accuracy, training time etc.

References

1. Wang, D., Chaudhari, N.S.: A Constructive Unsupervised Learning Algorithm for Clustering Binary Patterns. In: Proceedings of International Joint Conference on Neural Networks (IJCNN 2004), Budapest, July 2004, vol. 2, pp. 1381–1386 (2004)
2. Wang, D., Chaudhari, N.S.: A Novel Training Algorithm for Boolean Neural Networks Based on Multi-Level Geometrical Expansion. Neurocomputing 57C, 455–461 (2004)
3. Kim, J.H., Park, S.K.: The geometrical learning of binary neural neworks. IEEE Transaction. Neural Networks 6, 237–247 (1995)
4. Joo Er, M., Wu, S., Yang, G.: Dynamic Fuzzy Neural Networks. McGraw-Hill, New York (2003)
5. Kwok, T.Y., Yeung, D.Y.: Constructive algorithms for structure learning in feedforward neural networks for regression problems. IEEE Trans. Neural Networks 8, 630–645 (1997)
6. Chaudhari, N.S., Tiwari, A., Thomus, J.: Performance Evaluation of SVM Based Semi-supervised Classification Algorithm. In: International Conference on Control, Automation, Robotics and Vision, Hanoi, Vietman, December 17-20 (2008)
7. `http://www.ics.uci.edu/kmlearn/MLRepository.html`

Novel Deterministic Heuristics for Building Minimum Spanning Trees with Constrained Diameter

C. Patvardhan[1] and V. Prem Prakash[2]

[1] Faculty of Engineering, Dayalbagh Educational Institute, Agra, India
[2] Department of Information Technology, Anand Engineering College, Agra, India
cpatvardhan@ieee.org, vpremprakash@acm.org

Abstract. Given a connected, weighted, undirected graph G with n vertices and a positive integer bound D, the problem of computing the lowest cost spanning tree from amongst all spanning trees of the graph containing paths with at most D edges is known to be NP-Hard for $4 \leq D < n\text{-}1$. This is termed as the Diameter Constrained, or Bounded Diameter Minimum Spanning Tree Problem (BDMST). A well known greedy heuristic for this problem is based on Prim's algorithm, and computes BDMSTs in $O(n^4)$ time. A modified version of this heuristic using a tree-center based approach runs an order faster. A greedy randomized heuristic for the problem runs in $O(n^3)$ time and produces better (lower cost) spanning trees on Euclidean benchmark problem instances when the diameter bound is small. This paper presents two novel heuristics that compute low cost diameter-constrained spanning trees in $O(n^3)$ and $O(n^2)$ time respectively. The performance of these heuristics vis-à-vis the extant heuristics is shown to be better on a wide range of Euclidean benchmark instances used in the literature for the BDMST Problem.

Keywords: Heuristics, Diameter Constrained Minimum Spanning Tree, Bounded Diameter, Greedy.

1 Introduction

Given a connected, undirected graph G = (V, E) on |V| vertices, an integer bound $D \geq 2$ and non-zero edge costs associated with each edge $e \in E$, a diameter constrained, or bounded diameter spanning tree (BDST) is defined as a spanning tree $T \in E$ on G with tree diameter no greater than D. The BDMST Problem finds a bounded diameter spanning tree of minimum edge cost, $w(T) = \Sigma_{\forall e \in T} w(e)$. Barring the special cases of D = 2, D = 3, D = n − 1, and all edge weights being the same, the BDMST Problem is known to be NP-Hard [5]. Furthermore, the problem is also known to be hard to approximate; it has been shown that no polynomial time approximation algorithm can be guaranteed to find a solution whose cost is within log (n) of the optimum [6]. The problem finds application in several domains, including distributed mutual exclusion [4], wire-based communication network design [2], ad-hoc wireless networks and data compression for information retrieval [3].

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 68–73, 2009.

Three branch-and-bound exact algorithms for solving the BDMST problem are given in [10], however, the exponential time complexity of these algorithms severely restricts their application to all but very small instances. Two greedy heuristics are proposed by Abdalla et al. [1], one of which starts by computing an unconstrained MST and then iteratively transforms it into a BDMST. This heuristic is computationally expensive and does not always give a BDMST for small values of n. The other heuristic is known as One-Time Tree Construction (OTTC), which modifies Prim's algorithm [8] for unconstrained MSTs to iteratively select the lowest cost non-tree vertex whose inclusion in the tree does not violate the diameter constraint. When run starting once from each node (and returning the lowest cost BDST thus obtained), the OTTC algorithm requires $O(n^4)$ running time. A faster center-based variant of OTTC, known as Center-Based Tree Construction (CBTC) and a better performing Randomized Tree Construction heuristic (RTC) for the BDMST problem are proposed in [7]. Center-Based Tree Construction grows a BDST from the tree's center, keeping track of the depth of each tree node and ensuring that no node depth exceeds $\lfloor D/2 \rfloor$. This improves the running time of OTTC by O(n). RTC is also a center-based heuristic based on Prim's algorithm, but with the proviso that the next node to be added to the MST is chosen at random and attached greedily. Greedily appending a randomly chosen node to the tree can be performed in linear time; building the BDST would require $O(n^2)$ time. This process is repeated n times and the lowest cost tree obtained is returned; thus the total running time incurred by the RTC heuristic is $O(n^3)$. The greedy heuristics given so far are extended by Singh and Gupta [11] with an $O(n^2)$ time post-processing step that re-attaches each node to the BDMST at a lower cost if possible. The performance of the OTTC, CBTC and RTC heuristics is presented on a wide range of benchmark problems in a recent work by Julstrom [9].

This paper presents a novel heuristic called the Least Sum-of-Costs Node Next (LSoC) heuristic that constructs the MST by greedily including the next node with the lowest mean cost to the remaining non-tree nodes. This heuristic requires $O(n^3)$ time, or totally $O(n^4)$ time when run starting from each node, and produces lower cost trees as compared to OTTC. An improved center-based version of the LSoC heuristic that requires $O(n^3)$ time and two variants of a faster heuristic designed specifically for Euclidean cases of the problem are also given. The performance of the proposed heuristics is shown to be better than existing heuristics on a wide range of Euclidean problem instances used in the literature for the BDMST problem. The rest of this paper is organized as follows: section 2 introduces the LSoC heuristic and presents the improved version of the LSoC heuristic and two fast heuristics that work well on Euclidean problems; section 3 discusses the results obtained on Euclidean benchmark instances and compares them with those obtained by the extant heuristics, and some conclusions are presented in section 4.

2 Proposed Heuristics

The Least Sum-of-Costs Node Next (LSoC) heuristic always selects as the next tree node, the non-tree vertex with the lowest total cost to the remaining non-tree vertices. This vertex is then appended to the tree via the lowest cost edge that does not violate the diameter bound. Diameter checking is essentially an O(n^2) breadth-first tree traversal that computes the maximum eccentricity between any two tree nodes; this check is performed n-1 times, resulting in a total running time of O(n^3). In order to obtain a low cost BDST, the heuristic is run starting from each graph vertex. This adds a factor of n to the running time, but gives consistently better trees.

A center-based variant of LSoC (abbreviated to CBLSoC) starts with each node as center when D is even or each node plus the node with lowest mean cost to the remaining non-tree nodes as centers when D is odd. Thereafter the next node to be added to the tree is simply the node with the lowest mean cost to the remaining non-tree nodes, and is appended greedily to the tree, while ensuring that no tree node had depth greater than $\lfloor D/2 \rfloor$. Appending a node to the tree does not affect the depth of any tree node. Choosing the next node to append to the tree, and selecting its parent are both linear time operations, and are repeated n-1 times. Thus each BDST is constructed in O(n^2) time; since this process is repeated starting from each of the n nodes, the total computation time required for CBLSoC is O(n^3).

As mentioned already, OTTC (and CBTC) greedily appends nodes to the tree, with the result that the edges that form the backbone of the growing tree are typically short edges. Thus (possibly several of) the remaining nodes have to be appended using long edges, which add to the total cost of the tree. The relatively less greedy LSoC heuristic mitigates this to some extent, as the results in section 3 show. Another approach to this problem, designed specifically for Euclidean benchmark problems, tries to overcome this problem by setting up the tree center and some of the initial tree nodes empirically, in effect building a backbone comprising of a small number of nodes appended to the tree via relatively longer edges, and then constructing the rest of the BDST either greedily or using the LSoC heuristic. These heuristics, termed as Nine Quadrant Centers-based Heuristics (NQC), start by choosing the node(s) closest to the center of the (Euclidean) data plane as the tree center, and appending to the tree, the node closest to each of the eight respective quadrant centers of a uniform 3x3 matrix of the unit square representing the plane of data points, with the exception of the central quadrant (which in any case already contains the tree center). The node closest to the center of the unit square of Euclidean points is initially fixed as the center if (the diameter) D is even. On the other hand, if D is odd, then the next closest node is selected and added to the tree via the lowest cost edge, and the center of the spanning tree is formed using both these nodes. For each of the nine quadrants of the data plane, the graph vertex closest to a quadrant center is appended to the tree center and designated as a node of depth 1. The remaining vertices are now appended to the tree in a greedy manner (called the NQC-Greedy heuristic), or using the LSoC heuristic in selecting each next vertex to append to the tree (termed as the NQC-LSoC heuristic). Setting up the backbone of the BDST in this manner requires constant time in both NQC variants. Appending each node to the BDST requires O(n^2) time in the greedy variant and linear time in the LSoC variant; thus the total running time for the NQC-Greedy heuristic works out to O(n^3), whereas in the case of NQC-LSoC it amounts to O(n^2). All three heuristics incorporate the Singh et al [11] post processing step to further improve the cost, which does not change the overall time complexity, and improves tree cost in several cases.

3 Performance on Benchmark Problems

Problem instances from the Euclidean Steiner Problem data set available in Beasley's OR-Library[1] have been predominantly used in the literature to benchmark the performance of heuristics and algorithms for the BDMST Problem. In a recent work, Julstrom [9] used a larger set of benchmark problems by combining the Beasley data sets with randomly generated Euclidean graphs, fifteen each of 100, 250, 500 and 1000 node instances, whose edge weights are the Euclidean distances between (randomly generated) points in the unit square. The heuristics presented in section 2 were implemented in C on a PIV 3-GHz processor with 2 GB RAM running Fedora 9, and tested on thirty instances of 100, 250, 500 and 1000 node graphs from this augmented test suite, totaling 120 graphs, and the mean (X) and standard deviation (SD) of tree costs, and mean CPU times were obtained for each node size. The mean CPU times (t) given in tables 1 and 2 for the existing and proposed heuristics were obtained on computing systems with different configurations, and are hence not directly comparable. However, the lesser computational complexity of the proposed algorithms is very clearly reflected in the times shown. The CBLSoC heuristic is compared with OTTC and CBTC in table 1.

Table 1. Results for the OTTC, CBTC and CBLSoC Heuristics on up to 1000 node graphs

Instances		OTTC			CBTC			CBLSoC		
n	D	X	SD	t	X	SD	t	X	SD	t
100	5	29.38	1.71	0.07	26.48	1.51	0.01	**22.34**	1.43	0.05
	10	18.43	1.86	0.07	15.59	1.28	0.01	**12.53**	0.97	0.04
	15	12.84	1.47	0.07	10.95	1.00	0.01	**9.15**	0.56	0.05
	25	8.06	0.59	0.07	7.69	0.34	0.02	**7.32**	0.24	0.04
250	10	58.20	5.38	1.18	49.07	2.99	0.16	**31.10**	1.95	0.28
	15	41.59	4.03	1.22	36.44	3.15	0.19	**21.83**	1.65	0.31
	20	32.55	3.86	1.26	26.17	2.36	0.22	**14.64**	1.26	0.31
	40	14.43	2.11	1.36	12.51	0.83	0.33	**11.47**	0.20	0.31
500	15	106.87	5.03	12.69	94.38	5.56	1.63	**44.60**	3.19	2.92
	30	58.52	7.10	14.65	46.51	4.13	2.80	**25.32**	2.03	3.25
	45	32.23	5.66	16.66	25.63	2.18	4.25	**17.82**	0.78	3.14
	60	20.33	3.46	17.64	17.72	0.92	5.41	**16.03**	0.29	3.24
1000	20	217.71	9.48	150.06	195.96	7.97	13.18	**51.96**	1.81	41.30
	40	124.21	17.33	167.91	99.57	7.86	22.61	**33.52**	2.95	51.69
	60	69.83	12.20	183.21	50.97	5.24	33.28	**27.75**	1.72	54.70
	100	28.95	4.14	189.33	23.41	0.78	55.34	**22.58**	0.19	55.63

The mean BDST tree costs obtained over a wide range of Euclidean instances for the heuristics clearly show that the CBLSoC heuristic outperforms the other two heuristics on all instances. The results obtained also indicate lower standard deviation as compared to the other two heuristics.

[1] Maintained by J.E. Beasley, Department of Mathematical Sciences, Brunel University, UK. (http://people.brunel.ac.uk/~mastjjb/orlib/files)

Comparing the mean BDST costs obtained in table 1 and those given for the RTC heuristic in table 2, it can be seen that on small diameter bounds, the RTC heuristic performs significantly better than the deterministic heuristics discussed above, and presented in table 1. However, as the diameter bound is increased, RTC by virtue of the fact that its edge selection is always random, fails to take advantage of the constraint relaxation and, as a result, is able to produce only slight improvements in tree costs. By contrast, the OTTC, CBTC and CBLSoC produce lower cost trees with larger diameter bounds; the proposed CBLSoC heuristic produces better trees than the other two on every instance size considered.

Table 2. Results for the RTC and NQC Heuristics on benchmarks graphs up to 1000 nodes

Instances		RTC			NQC-LSoC			NQC-Greedy		
n	D	X	SD	t	X	SD	t	X	SD	t
100	5	15.30	0.78	<0.01	**14.64**	0.53	0.0003	**14.64**	0.55	0.0033
	10	**9.38**	1.80	0.01	10.10	0.32	0.0003	9.95	0.32	0.0060
	15	**9.25**	0.28	0.01	9.65	0.32	0.0003	9.26	0.33	0.0060
	25	**9.19**	0.27	0.01	9.47	0.28	0.0003	**9.03**	0.30	0.0060
250	10	**16.80**	0.33	0.16	**16.76**	0.44	0.0020	17.01	0.54	0.0440
	15	15.30	0.23	0.25	**15.10**	0.35	0.0020	**14.70**	0.34	0.0547
	20	15.05	0.23	0.27	**14.36**	0.37	0.0063	**13.49**	0.23	0.0623
	40	14.98	0.21	0.27	**14.07**	0.19	0.0020	**13.34**	0.17	0.0530
500	15	**22.26**	0.37	4.47	23.01	0.59	0.0253	24.19	0.76	0.2763
	30	21.54	0.33	5.30	**19.87**	0.42	0.0090	**18.52**	0.35	0.3440
	45	21.43	0.33	5.43	**19.21**	0.55	0.0120	**17.83**	0.32	0.3233
	60	21.45	0.34	5.29	**18.83**	0.34	0.0270	**17.69**	0.29	0.3347
1000	20	**31.15**	0.24	52.28	32.04	0.88	0.0580	36.10	1.66	3.1853
	40	30.85	0.22	56.17	**27.50**	0.68	0.0883	**25.29**	0.48	3.9543
	60	30.84	0.24	56.82	**26.68**	0.66	0.0863	**24.33**	0.23	4.0297
	100	30.84	0.24	55.51	**25.49**	0.20	0.0780	**23.87**	0.18	4.0510

Table 2 also provides the results of the greedy and LSoC-based NQC Heuristic variants alongside those given for the RTC heuristic. As can be seen from the documented mean tree costs, the NQC heuristics produce lesser cost trees on average in several instances. The greedy variant obtains the best costs, in time that is $O(n^3)$, which is the same as the running time of the RTC heuristic. However, the LSoC-based variant also produces better trees in comparison to the RTC in several instances, especially on the higher sized instances and with increasing values of diameter constraints. Furthermore, the LSoC-based variant of the NQC heuristic runs in $O(n^2)$ time, which is an order of significance lesser than both the greedy variant and the extant RTC heuristic.

4 Conclusions

This work presents novel deterministic heuristics that produce good approximate BDMSTs in $O(n^3)$ and $O(n^2)$ time. These are compared with three existing heuristics for the problem over a large set of Euclidean benchmark instances. The Center-based

LSoC heuristic performs better than the relatively more greedy OTTC heuristic, and runs in time which is O(n) lesser. This heuristic also produces consistently better results in comparison to the CBTC heuristic. In the case of instances with small diameter bounds, all three heuristics produce inferior quality trees in comparison with the randomized RTC heuristic; however, they obtain better mean costs vis-à-vis RTC as the diameter bound is raised. The quadrant centers-based heuristics presented were developed keeping in mind the drawbacks encountered with the OTTC heuristic. They are aimed specifically at Euclidean problem instances; the nine quadrants-based approach was found to produce superior results as compared to other empirical approaches explored. The greedy variant of the NQC heuristic produces the lowest mean costs on several of the instances, however it runs in $O(n^3)$ time. Significantly, the LSoC-based variant of the NQC heuristic performs competitively in $O(n^2)$ time, obtaining mean costs comparable with the RTC and NQC-Greedy heuristics on small diameter bounds, and lower cost BDSTs in comparison with the OTTC, CBTC and CBLSoC heuristics on several instances with larger diameter bounds.

Acknowledgments. The authors are extremely grateful to the Revered Chairman, Advisory Committee on Education, Dayalbagh for the continued guidance. The authors also thank Prof. Bryant A. Julstrom, St. Cloud University, Minnesota (USA) for providing the extended suite of problem instances for the BDMST problem.

References

1. Abdalla, A., Deo, N., Gupta, P.: Random-tree Diameter and the diameter constrained MST. In: Congressus Numerantium, vol. 144, pp. 161–182. Utilitas Mathematica (2000)
2. Bala, K., Petropoulos, K., Stern, T.E.: Multicasting in a linear lightwave network. In: IEEE INFOCOM 1993, pp. 1350–1358 (1993)
3. Bookstein, A., Klein, S.T.: Compression of correlated bit-vectors. Information Systems 16(4), 110–118 (1996)
4. Raymond, K.: A tree-based algorithm for distributed mutual exclusion. ACM Transactions on Computer Systems 7(1), 61–77 (1989)
5. Garey, M.R., Johnson, D.S.: Computers and Intractibility: A Guide to the Theory of NP-Completeness. W.H. Freeman, New York (1979)
6. Kortsarz, G., Peleg, D.: Approximating shallow-light trees. In: Proc. Eighth ACM-SIAM Symposium on Discrete Algorithms, pp. 103–110 (1997)
7. Julstrom, B.A., Raidl, G.R.: A permutation-coded EA for the BDMST problem. In: GECCO 2003 Workshops Proc., Workshop on Analysis & Design of Representations, pp. 2–7 (2003)
8. Prim, R.C.: Shortest connection networks and some generalizations. Bell System Technical Journal 36, 1389–1401 (1957)
9. Julstrom, B.A.: Greedy Heuristics for the Bounded Diameter Minimum Spanning Tree Problem. ACM J. Exp. Algor. 14, Article 1.1 (2009)
10. Achuthan, N.R., Caccetta, L., Cacetta, P., Geelen, J.F.: Algorithms for the minimum weight spanning tree with bounded diameter problem. Optimization: Techniques and Applications, 297–304 (1992)
11. Singh, A., Gupta, A.K.: Improved heuristics for the bounded diameter minimum spanning tree problem. Soft Computing 11(10), 911–921 (2007)

Multi-objective Evolutionary Feature Selection

Partha Pratim Kundu and Sushmita Mitra

Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700 108, India
{ppkundu_r,sushmita}@isical.ac.in

Abstract. A new method of evolutionary feature selection, using multi-objective optimization in terms of fuzzy proximity and feature set cardinality, is developed. Results on two datasets indicate selection of the correct feature subset.

Keywords: Feature selection, Multi-objective optimization, Proximity.

1 Introduction

Feature selection is helpful as a preprocessing step for selecting a subset of n' features from an original set of n features ($n' < n$), so that the feature space is optimally reduced according to an evaluation criterion. The use of soft computing is an interesting proposition along this direction [1]. We can utilize the search potential of genetic algorithms for efficiently traversing large search spaces. When there are two or more conflicting characteristics to be optimized, often the single-objective genetic algorithm (GA) [2] requires an appropriate formulation of the single fitness function in terms of an additive combination of the different criteria involved. In such cases *multi-objective* GAs (MOGAs) [3] provide an alternative, more efficient, approach to search for optimal solutions.

For a dataset with N input patterns we can define an $N \times N$ symmetry matrix, termed the proximity matrix P, whose (i, j)th entry represents the similarity (or dissimilarity) measure for the ith and jth patterns for $i, j = 1, \ldots, N$. Typically distance functions are used for the purpose. The proximity matrix is a pertinent construct that allows us to deal with structural information inherent in the data. In the fuzzy perspective the concept of similarity boils down to the membership values.

In this article we focus on the proximity relationship from the fuzzy viewpoint, and employ this as one of the fitness functions of the MOGA for evaluating the fitness of the feature subsets of varying cardinality. The use of fuzziness allows us to efficiently model uncertainties and ambiguities inherent in real life overlapping data. The proximity of a pair of patterns in the original feature space is compared with that in the reduced subspace of selected features. If they are similar then this implies that the eliminated feature(s) are not so relevant to the decision making process. The second criterion is the cardinality of the selected feature subset. This is sought to be minimized. A close observation reveals that these two criteria are of a conflicting nature. A smaller subset of features is likely to result in a reduced proximity, and hence reduced classification accuracy (as compared

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 74–79, 2009.

to the original feature space). This prompted us to formulate the problem in the framework of MOGA. The algorithm terminates when an optimal subset of features is obtained, according to the fitness criteria of the multi-objective genetic optimization.

2 Proximity-Based Feature Selection

Fuzzy c-means (FCM) clusters a set of N patterns $\{\boldsymbol{x}_j\}$ into c clusters by minimizing the objective function $J = \sum_{j=1}^{N} \sum_{i=1}^{c} (\mu_{ij})^{m'} ||\boldsymbol{x}_j - \boldsymbol{m}_i||^2$, where $1 \leq m' < \infty$ is the fuzzifier, $\mu_{ij} \in [0, 1]$ is the membership of the jth pattern to the ith mean $\boldsymbol{m}_i$, and $||.||$ is the distance norm, such that $\boldsymbol{m}_i = \frac{\sum_{j=1}^{N} (\mu_{ij})^{m'} \boldsymbol{x}_j}{\sum_{j=1}^{N} (\mu_{ij})^{m'}}$ and $\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ji}}{d_{jk}}\right)^{\frac{2}{m'-1}}}$, $\forall i$, with $d_{ji} = ||\boldsymbol{x}_j - \boldsymbol{m}_i||_2$, subject to $\sum_{i=1}^{c} \mu_{ik} = 1$, $\forall k$, and $0 < \sum_{k=1}^{N} \mu_{ik} < N$, $\forall i$. The fuzzy partitions computed by FCM, are directly related to the proximity relation. We have

$$p(k_1, k_2) = \sum_{i=1}^{c} (\mu_{ik_1} \wedge \mu_{ik_2}), \tag{1}$$

where $\wedge$ indicates the minimum operation, $p(k_1, k_2) \in [0, 1]$, and $k_1, k_2 = 1, \ldots, N$. Evidently, $p(k_1, k_2) = 1$ for $k_1 = k_2$ and $p(k_1, k_2) = p(k_2, k_1)$.

Let there be K subsets of data located in different feature subspaces, with the number of patterns in each subspace being equal to N. We form a $c \times N$ partition matrix Ω consisting of fuzzy membership values μ_{ij}. Note that the dimensionality n of the patterns in each subset could be different. However, the distance of a pattern is computed from the fuzzy cluster prototypes over the same subset of features.

The partition matrix is used to evaluate proximity p, which measures the extent to which a pair of patterns are regarded as similar or dissimilar in different subspaces [4]. The proximity matrix P contains the proximity results for all possible pairs of patterns.

2.1 Proximity between Feature Spaces

Let the cardinality of the original and reduced feature spaces be n and n', respectively. Let the proximity matrices in these two spaces be denoted by P and P' respectively. The similarity between the two matrices is represented by a scalar value

$$P_s = \sum_{k_1=1, k_2=1, k_2>k_1}^{N} [p(k_1, k_2) \wedge p'(k_1, k_2)], \tag{2}$$

where $p'(k_1, k_2)$ is computed by eqn. (1) in the reduced feature space. In order to enhance contrast we use

$$P_{s_0} = \sum_{k_1=1, k_2=1, k_2>k_1}^{N} [(p(k_1, k_2) \geq 0.5) \wedge (p'(k_1, k_2) \geq 0.5)]. \tag{3}$$

2.2 Use of MOGA

We encode the problem as a real-coded string of length L, with the first n bits corresponding to the n features in the original space. In the bit representation, here a “1” implies that the corresponding attribute is present while “0” indicates that it is not. Let the size of a chromosome be $L = n + c \times n = n \times (c + 1)$. The c cluster centers are encoded in real form in the subsequent $c \times n$ bits. Only those features of the centers in the second part of the string, corresponding to a “1” in the first part, are considered during clustering. Fig. 1 depicts such an encoding in a chromosome, representing a sample set of cluster prototypes in a feature subspace. Initially all the bits are set randomly.

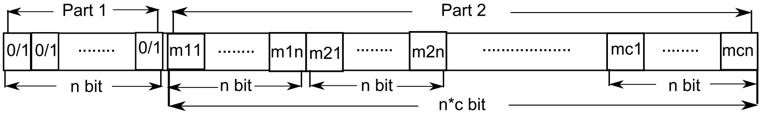


Fig. 1. An encoded chromosome representing a feature subspace and cluster prototypes

The objective is to optimize a conflicting set of requirements *i.e.*, select a minimal number of features that enable us to arrive at an acceptable (classification) decision. We employ MOGA with P_{s_0} of eqn. (3) as the fitness function $f_1 = P_{s_0}$. The second fitness function corresponds to the cardinality of the feature set under consideration, and is defined as $f_2 = n'$. While f_2 gives credit to a candidate string containing less attributes, the function f_1 determines the extent to which the set of all pairs of patterns belong to the same cluster in the two feature spaces, *viz.*, original and reduced subspace. These two fitness functions are optimized in the framework of MOGA. Clustering is done by FCM to update the prototypes m_i, in the different subspaces.

The performance of the selected feature subset is validated in terms of its accuracy of performance, as measured by the well-known k-nearest neighbour (k-NN) classifier [5].

2.3 The Algorithm

1. Initialize the population randomly.
2. Select a pair of chromosomes randomly, along with a random bit location for the initiation of single-point crossover.
3. Perform two-point mutation on the two parts of the string with the corresponding real values being randomly altered. In the first part, the value of the randomly chosen bit (signifying presence or absence of the corresponding attribute) is flipped. In case of the second part, the value $m_{ik_{old}}$ corresponds to a randomly chosen attribute , k of a i^{th} cluster center; this is mutated as $m_{ik_{new}} = \sigma \times x + m_{ik_{old}}$, where the random variable $x(\sim N(0, 1)$ is drawn from a Gaussian distribution, the variance σ^2 determines the magnitude

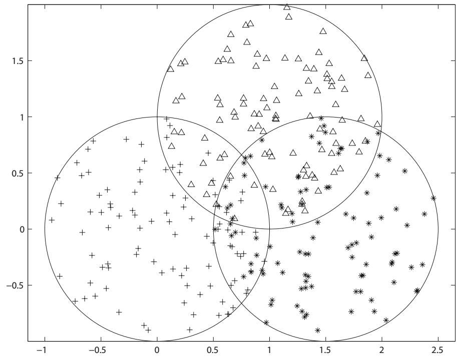


Fig. 2. Synthetic data

of this perturbation at position $m_{ik_{old}}$, and $m_{ik_{new}}$ is its new value after mutation.

4. Calculate the fitness values using f_1 and f_2.
5. Rank the population using dominance criteria.
6. Calculate the crowding distance of the chromosome.
7. Combine parent and offspring population. Replace the parent population by the best members of the combined population.

Note that the cluster centers are set randomly. During crossover and mutation the centers get modified. Their effect is reflected through the proximity function P_{so} into the fitness evaluation. The features present in a chromosome, as indicated by the "1" s in the first part, determine the feature subspace. They affect the computation of proximity in terms of cluster prototypes, using eqns. (1).

3 Experimental Results

The performance of the algorithm was tested on a synthetic data set and Iris flower data. *Iris* contains 150 instances, with four features and three classes of iris flower. The synthetic data contains three clusters, each with 100 randomly generated patterns. The two-dimensional scatter plot of Fig. 2 depicts the patterns lying within circles of unit radius, each having different centers. A lot of overlapping is artificially introduced. We introduced a third attribute having completely random values, to test the effectiveness of the algorithm in identifying the significance of the first two features.

The performance of the algorithm for strings generated in the non-dominated Pareto front, for the two datasets, are presented in Table 1. The second column indicates the selected attributes, marked by a "1". The third and fourth columns denote the corresponding two fitness functions, as evaluated by f_1 and f_2. The last four columns provide the classification accuracy (for the three classes and the total) for the corresponding feature subspace, corresponding to different values of k.

Table 1. Performance of selected features in Pareto-optimal front

Dataset	Feature subspace	f_1 Proximity	f_2 Cardinality	k-NN performance				
				$k =$	accuracy (%)			
					C1	C2	C3	NET
Iris	0111	5801.68	3	1	100.00	94.00	90.00	94.66
				3	100.00	92.00	98.00	94.00
				5	100.00	94.00	88.00	94.00
				7	100.00	94.00	94.00	96.00
	0010	3401.15	1	1	100.00	94.00	68.00	88.00
				3	100.00	88.00	86.00	91.33
				5	100.00	92.00	80.00	90.67
				7	100.00	92.00	86.00	92.67
	0011	3552.53	2	1	100.00	94.00	90.00	**94.66**
				3	100.00	94.00	92.00	**95.33**
				5	100.00	96.00	92.00	**96.00**
				7	100.00	96.00	92.00	**96.00**
Synthetic	111	21074.38	3	1	80.00	77.00	74.00	77.00
				3	78.00	76.00	72.00	75.33
				5	83.00	72.00	75.00	76.67
				7	83.00	75.00	70.00	76.00
	100	7444.99	1	1	75.00	66.00	52.00	64.33
				3	77.00	55.00	41.00	57.67
				5	81.00	60.00	38.00	59.67
				7	87.00	64.00	35.00	62.00
	110	15255.10	2	1	83.00	78.00	79.00	**80.00**
				3	81.00	78.00	79.00	**79.33**
				5	84.00	81.00	82.00	**82.33**
				7	87.00	82.00	78.00	**82.33**

Table 2. Comparative study on *Iris* data

Algorithm	PR	DK	PC	IM	$R*$
Selected Features	{3, 4}	{3, 4}	{3, 4}	{3, 4}	{3, 4}

In case of *Iris* data, it is observed that the choice of feature 3 occurs in all the three cases, with feature 4 being selected next most frequently. Together they result in the second highest proximity and second lowest cardinality. The accuracy in classification is also the best over all values of k. It is well-known that these are the two features most important for discriminating between the classes in this benchmark data. It is to be noted that the accuracy of performance over the original feature space (highest proximity and highest cardinality) turned out to be 95.33%, 95.33%, 96.00%, 96.00%, respectively, for $k = 1, 3, 5, 7$. Interestingly, with $k = 3, 5, 7$, the performance in the reduced space is found to be equally good – inspite of the elimination of two features.

The result obtained by the proposed algorithm (model PR) for *Iris* data was compared with some of the existing techniques, considered as benchmarks in this study. These are the statistical method of Devijver and Kittler [5] (model DK), the fuzzy entropy based method of Pal and Chakraborty [6] (model PC), the neural network based method of Ruck and Rogers [7] (model R*), and that of Ishibuchi [8] (model IM). Table 2 demonstrates a comparative study of the feature subsets selected by different algorithms for the *Iris* data. The overall study shows that the results tally with each other. The features 4 and 3 are found to be more important than the features 1 and 2 for classifying *Iris* data.

We know that the synthetic data is represented with the first two attributes, with the third feature being inserted randomly. As evident from the results, the selection of the first two features (only) results in an improved accuracy, over all values of k, due to the elimination of the unimportant third feature.

4 Conclusion

A new multi-objective feature selection algorithm has been developed. Fuzzy proximity was used to evaluate the similarity between the original and reduced feature subspaces. The cardinality of the feature subset was simultaneously minimized. Experimental results demonstrated the effectiveness of the developed method.

References

1. Banerjee, M., Mitra, S., Banka, H.: Evolutionary-rough feature selection in gene expression data. IEEE Trans. Syst. Man Cybern Part C 37, 622–632 (2007)
2. Goldberg, D.E.: Genetic Alogorithm in Search Optimization and Machine Learning. Addison-Wesley, Reading (1989)
3. Deb, K.: Multi-Objective Optimiztion using Evolutionary Algorithms. John Wiley & Sons, Chichester (2002)
4. Pedrycz, W., Loia, V., Senatore, S.: P-FCM: A proximity-based fuzzy clustering. Fuzzy Sets and Systems 148, 21–41 (2004)
5. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice-Hall, Englewood Cliffs (1982)
6. Pal, S.K., Chakraborty, B.: Fuzzy set theoretic measure for automatic feature evaluation. IEEE Transactions on Systems, Man, and Cybernetics 16, 754–760 (1986)
7. Ruck, D.W., Rogers, S.K., Kabrisky, M.: Feature selection using a multilayer perceptron. Neural Network Computing 20, 40–48 (1990)
8. Ishibuchi, H., Miyazaki, A.: Determination of inspection order for classifying new samples by neural networks. In: Proceedings of IEEE - International Conference on Neural Networks, Orlando, USA, pp. 2907–2910 (1994)

A Fast Supervised Method of Feature Ranking and Selection for Pattern Classification

Suranjana Samanta and Sukhendu Das

Dept. of CSE, IIT Madras, Chennai, India
ssamanta@cse.iitm.ac.in, sdas@iitm.ac.in

Abstract. This paper describes a fast, non-parametric algorithm for feature ranking and selection for better classification accuracy. In real world cases, some of the features are noisy or redundant, which leads to the question - which features must be selected to obtain the best classification accuracy? We propose a supervised feature selection method, where features forming distinct class-wise distributions are given preference. Number of features selected for final classification is adaptive, but depends on the dataset used for training. We validate our proposed method by comparing with an existing method using real world datasets.

1 Introduction

Feature selection is an important pre-processing step for classification of any large dimensional dataset. The noisy features confuse the classifier and often mislead the classification task. In order to obtain better accuracy, the redundant, inappropriate and noisy features should be removed before the process of classification. Given a dataset of D dimension, the aim of feature selection is to select a subset of best M features ($M < D$), which produces better result than that of the entire set. A lot of work has been done on feature ranking and selection [1], [2]. Feature selection methods can be broadly classified into two types [2], filter and wrapper methods. In filter method, features are first ranked based on a specific criteria (mostly statistical evaluation), and then the best 'M' features are considered for classification. Huan Liu and Lei Yu [2] have explained the feature selection algorithm into four different steps, namely, subset generation, subset evaluation, stopping criteria and result validation. In most of the work done earlier adaptiveness in the final number of selected features is missing, where the input is either a pre-determined number of features to be selected or a pre-determined threshold [1], which when exceeded stops the selection algorithm.

We propose a new feature ranking method and an adaptive feature selection technique. First, we obtain a score (quality measure) using any of the three proposed measures for each feature (dimension) based on the class-wise scatter of the training samples. Next, we rank the features using a greedy search method. Finally, we determine the number of top ranked features to be used for final classification, using Fisher's discriminant criteria. Results obtained over real world datasets show significant improvement in both performance and speed

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 80–85, 2009.

when compared with one of the current state of the art [1]. The next few sections give a detail discussion of the different steps of the algorithm.

2 Feature Selection and Ranking

The overall process consists of two steps - feature ranking and selecting the appropriate number of features in the final set. For speedup in computation a greedy method has been used, where we assume the top M ranked features to give better classification result than the total set.

2.1 Feature Ranking

A good feature subset should have low within-class scatter and large between-class scatter. In an ideal case, instances of a good feature or a subset of features belonging to a particular class, should form a compact non-overlapping cluster. In such a case, we expect C non-overlapping clusters Cr_j, $\forall j = 1, 2, ..., C$, where C is the number of classes. We obtain a class label, $\psi(Cr_j)$, for each of the C clusters formed, where $j = 1, 2, ..., C$. Let $class(i)$ denotes the class label of the i^{th} training sample. To calculate the $\psi(Cr_j)$ for a cluster Cr_j $(j = 1, 2, ..., C)$, *we find the class label from training data, which has the highest number of instances in* Cr_j. In this way all clusters are approximated by the distribution made by instances of training samples belonging to each of the distinct classes.

The method for computing $\psi(Cr_j)$, $\forall j = 1, 2, ..., C$, depends of the clustering algorithm used, which has been implemented in two ways. One way of clustering is to fit a 1-D Gaussian function separately on the data instances belonging to each of the classes using GMM (Gaussian Mixture Model). The class assignment for a cluster is the class label for instances of the training data used to fit the Gaussian curve. The other way of clustering is FCM (Fuzzy C-Means), an unsupervised algorithm, for C clusters. To compute the $\psi(.)$ for C clusters in this case, we form a confusion matrix (CM), where we calculate the number of instances of different classes (as given in training data) present in each cluster. The rows of CM correspond to the C clusters formed using FCM and the columns of CM corresponds to the C classes of the training dataset. Computation of ψ involves: *(i)* Convert the elements of $CM(C \times C)$ to a probability, using number of samples in each class; *(ii)* Obtain a class label ψ for each cluster, having maximum probability (row-wise) in a cluster.

After obtaining the ψ for each cluster, we determine the probability (soft class labels) of each instance belonging to each of the C clusters given by $p(k|j)$, $k = 1, 2, ..., N$ and $j = 1, 2, ..., C$; to incorporate fuzziness in the ranking process. Based on this clustering result, we calculate $P(k) = max(p(k|j))$, $\forall k = 1, 2, ..., N$, where N is the total number of training samples. We find a score for each feature (dimension) based on its class-wise distribution in each cluster. We assign a penalty for the occurrence of all instances i belonging to Cr_j whenever $class(i) \neq \psi(Cr_j)$. Class information is used to obtain how many instances of all other classes exist in a cluster. We calculate a probabilistic score based on the concept of Cohen-Kappa (κ) measure [3], which is used to measure inter-classifier agreement. Though we are not dealing with multiple classifiers here, we calculate the agreement between original class label of the training data with the class labels (ψ) assigned by clustering, using the Cohen-Kappa measure. This measure gives an indicator of how well the clusters are formed according to the original class-labels. So, we calculate the first measure: S_1, as,

$$S_1 = (c_1 - c_2)/(N - c_2) \tag{1}$$

where, $c_1 = |\{k|class(k) = \psi(Cr_j),\, k \in Cr_j\, \forall j = 1, 2, ..., C\}|$ and $c_2 = |\{k|class(k) \neq \psi(Cr_j),\, k \in Cr_j\, \forall j = 1, 2, ..., C\}|$. We extend this idea to calculate another fuzzy measure: S_2, as,

$$S_2 = (\phi_r - \phi_c)/(N - \phi_c) \tag{2}$$

where, ϕ_r and ϕ_c are similar to relative-agreement and chance-by-agreement of Cohen-Kappa [3] which is given by, $\phi_r = \sum_{j=1}^{C} \sum_{k \in Cr_j} P(k)|_{class(k)=\psi(Cr_j)}$ and $\phi_c = \sum_{j=1}^{C} \sum_{k \in Cr_j} P(k)|_{class(k)\neq\psi(Cr_j)}$.

The third measure: S_3 is based on the principle that the matrix CM should be diagonal in an ideal case which can be expressed as,

$$S_3 = (1 + [sum(CM) - trace(CM)])^{-1} \tag{3}$$

The last stage of the feature ranking process involves sorting the feature dimension sequentially based on any one of the measures (S_1, S_2 or S_3). At any stage of iteration, let F_s be the set of selected features and F_{us} be the remaining set of features. Once a feature is ranked, it is removed from F_{us} and is added to Fs. The feature with the highest score is ranked first. The next ranked feature should not only have good score but also should be distinct or non-correlated with the already selected features in order to contribute to better classification accuracy. Hence, for obtaining the next feature to be ranked, we consider the correlation between all pairs of features f_t and f_s, where $f_t \in F_{us}$ and $f_s \in F_s$. For obtaining the next ranked features from F_{us}, we compute the difference in the score calculated of the feature and the mean of the correlation of that feature with other set of features selected in F_s. The feature with the maximum value is chosen as the next ranked feature. This process goes on until $|F_s| = D$ and $|F_{us}| = NULL$. The total method has been described in Algorithm 1.

2.2 Feature Selection

After the process of feature ranking, the best M features must be selected. We choose the popular Fisher's discriminant criteria [4], for selecting a subset of M features from D rank-ordered features ($M < D$), which is given by $trace(S_w^{-1} \times S_b)$ where, S_b and S_w are the between-class and within-class scatter matrix [4]. Fisher's criteria is computed for the top ranked features, at each iteration with increasing dimension. This criteria when observed with increasing number of features in the dataset, shows an abrupt change in the gradient at a point. The

Algorithm 1. Algorithm for ranking the feature set.
INPUT: Training dataset $'data'$ along with their class-labels.
OUTPUT: Rank-ordered dataset F_s.

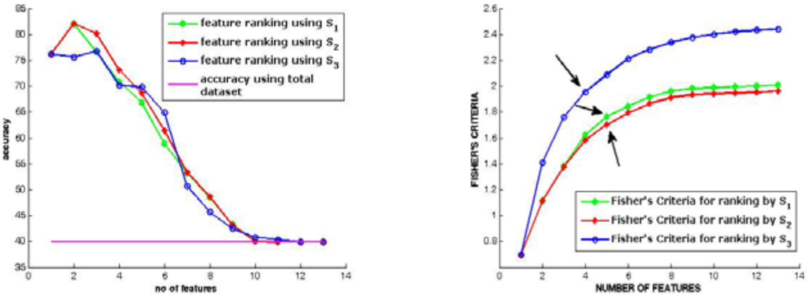
```
for $i = 1$ to D do
    Form $C$ clusters $(Cr)$ with the $i^{th}$ feature dimension in $'data'$, using FCM or GMM.
    Form confusion matrix $(CM^{C\times C})$ where, $cm_{jk} \in CM$ denotes the number of samples belonging to $j^{th}$ class and present in $k^{th}$ cluster $(Cr_k)$.
    Compute $\psi$ (class label) for each cluster, using $CM$ in case of FCM or class labels in training data in case of GMM.
    Compute $S_i$ using any one of the measures given in Eqns. 1, 2 or 3.
end for
$F_s \leftarrow data[:, \arg\max_i(S_i)]$ /* select the best dimension */
$F_{us} \leftarrow [data - F_s]$
while $|F_{us}| \neq NULL$ do
    for $i = 1$ to $|F_{us}|$ do
        $h = \sum_{i=1}^{|F_{us}|} \sum_{j=1}^{|F_s|} corr(f_i, f_j)$ where, $f_i \in F_{us}$, $f_j \in F_s$
        $r_i = S_i - h/|F_s|$
    end for
    $k = \arg\max_i [r_i]$
    Remove $k^{th}$ feature from $F_{us}$ and insert as last dimension in $F_s$
end while
```

features which are ranked after this point of transition are generally redundant and confusing, and hence removed from the final dataset.

Fig. 1(a) shows the change in the classification accuracy of Wine dataset obtained from UCI repository [5]. Fig. 1(b) shows the plot of Fisher's criteria with increasing number of rank-ordered features in the subset. The arrow marks points to the 'knee point' of the curves, for each of the scores, where there is a significant change in the gradient of the curve. In Fig. 1(b), after considering the 4^{th} rank-ordered feature the gradient of the curve does not show any appreciable change, and hence we select the first 4 features, which is consistent with the drop in accuracy in Fig. 1(a).

3 Experimental Results

Experiments done on real world datasets obtained from UCI Repository [5], show significant improvement of classification accuracy after feature ranking and selection. One-third of the training samples are used for feature ranking and selection and the rest are used for testing. Results of accuracy in classification obtained using SVM with Gaussian kernel have been reported after ten fold cross validation. Fig. 2(a) and (b) show four results each, obtained using FCM and GMM clustering method respectively using S_1 (green curve), S_2 (red curve),



(a) (b)

Fig. 1. Plot of (a) classification accuracy & (b) Fisher's criteria with increasing number of rank-ordered features. Wine Dataset has been used from UCI Repository [5].

Table 1. Accuracy and computational time of the proposed feature ranking and selection method using both FCM & GMM clustering. (best result in bold)

Dataset	Total %	Accuracy %						Time (in secs.)		
		FCM			GMM			FCM	GMM	M.I. [1]
		S_1	S_2	S_3	S_1	S_2	S_3			
Cancer	86.77	92.70	**93.08**	92.90	92.00	91.54	92.53	**0.20**	1.37	14.84
Lymph	57.45	67.98	**78.51**	72.55	70.21	66.06	67.55	**0.39**	0.76	3.23
Metadata	29.09	69.83	**100**	**100**	**100**	**100**	**100**	14.20	**3.35**	542.76
Spectf	52.88	63.58	63.65	64.81	64.81	**66.54**	65.38	**3.60**	4.19	90.64
Vehicle	26.19	29.66	29.66	31.17	33.63	35.20	**37.14**	**1.44**	6.81	1669
Zoo	80.31	89.84	94.84	**96.72**	95.94	95.47	92.81	**0.55**	0.61	2.07

S_3 (blue curve) and mutual information (mustard curve) [1] as feature ranking measures. The magenta horizontal line shows the classification accuracy using total dataset (baseline). Table 1 gives a comparative study of the classification accuracy obtained using two different clustering methods along with the baseline accuracy using total dataset. It also shows the average time taken (in seconds) to compute the feature ranking and selection using any of the proposed measures (S_1, S_2 or S_3), with respect to the method using Mutual Information [1].

4 Conclusion

We have observed the performance of the feature selection algorithm on real world datasets. Most of the existing feature selection algorithms are time consuming and parametric. To overcome this problem, we propose a fast nonparametric selection algorithm which improves the classification accuracy. Results discussed in the paper show that the proposed feature ranking method is better than in [1].

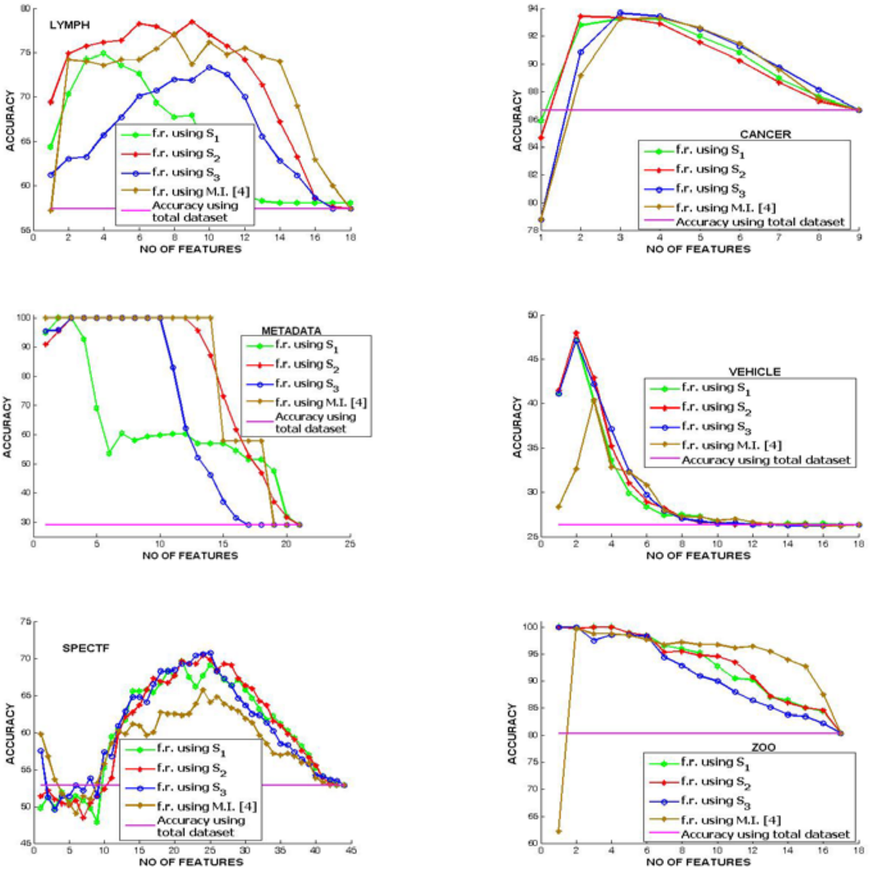


Fig. 2. Comparison of performance of the proposed method with [1], using (a) FCM- and (b) GMM based clustering. * f.r.: feature ranking; M.I.: mutual information [1].

References

1. Guoa, B., Damper, R., Gunna, S.R., Nelsona, J.: A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. Pattern Recognition 41, 1653–1662 (2007)
2. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering 17, 491–502 (2005)
3. Sim, J., Wright, C.C.: The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy 85, 257–268 (2005)
4. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Computer Science and Scientific Computing Series. Academic Press, London (1990)
5. Asuncion, A., Newman, D.: UCI machine learning repository (2007)

Clustering in Concept Association Networks

Arun R., V. Suresh, and C.E. Veni Madhavan

Department of Computer Science and Automation, Indian Institute of Science,
Bangalore 560 012, India
{arun_r,vsuresh,cevm}@csa.iisc.ernet.in

Abstract. We view association of concepts as a complex network and present a heuristic for clustering concepts by taking into account the underlying network structure of their associations. Clusters generated from our approach are qualitatively better than clusters generated from the conventional spectral clustering mechanism used for graph partitioning.

Keywords: complex-networks; clustering; concept associations; clustering coefficient; power-law distribution; scale-free networks; small-world networks.

1 Introduction

Studies on complex networks –networks arising from real-world interactions– have shed light towards understanding universal principles behind the dynamics of these diverse systems [1]. We view human concepts as a system of interacting entities and study their network structure. We observe that concepts and their associations have network properties that are similar to other well established networks like WWW and social networks. We then show that understanding the network structures of concepts is useful in grouping similar concepts and that these groupings are better than those generated by a conventional clustering method like spectral clustering [7]. This work is organized as following sections: In the next section we give an overview of some developments in complex network models and related work. This is followed by our empirical observations on concept association networks. We then describe our clustering heuristic and compare it with an existing approach following which we present our concluding remarks.

2 Complex Networks: Models and Related Work

The first model to explain successfully the properties of social networks is the Small-World Network Model [4] by introducing the notion of clustering coefficient to describe the common contacts shared in human friendships. Following this Albert and Barabási [3] explained the structure and evolution of WWW and other networks by accounting for the presence of hubs (nodes that interact with a large number of other nodes in the network) through the preferential attachment

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 86–91, 2009.

model. Networks obeying such growth models are known as power law or scale-free networks; their degree distributions obey the relation $p(k) \sim k^{-\gamma}$, where $p(k)$ is the probability of a node to have degree k and γ controls the number of hubs.

Word Web –a network of interacting words based on their co-occurrence in a sentence– is seen to have power-law degree distribution and small-world characteristics [5]. WordNet[1] and word associations from USF Free Association Norm[2] –the same database which we use in this work– are shown to be scale-free, small-worlds and possess high clustering coefficients [6]. Dorogovtsev and Mendes [2] study the evolution of languages by using a preferential growth model in conjunction with evolving links among existing nodes.

3 Concept Association Networks(CAN)

Concept associations can be intuitively understood as thoughts that occur in conjunction with each other. In this work individual words are considered to be 'concepts'; words that occur in the mind as a response to a cue word are considered to be concepts that are *cognitively associated* with the it. Thus a concept graph $G = (V, E)$, is a set of labelled nodes V representing concepts, and a pair of them are connected by an edge if they have a cognitive association. The set of edges are collectively represented as E. For our study we use word associations available from USF Free Association Norm.

It is evident that these networks, from the way they are built, are directed and weighted –MILK might evoke CAT, but CAT may evoke RAT more than MILK; however, we choose to treat CAN as an undirected, unweighted graph as it facilitates an easy first-cut analysis. Insights gained from this study would be put to use in the analysis of the more generic graph representations. An illustration of concept graph studied in this work is given in Figure 1. We note here that CAN is fundamentally different from word co-occurrence networks and semantic networks like WordNet. The latter are as a result of the structure of the language which may be studied and understood through the network properties. However they may not shed light on cognitive association of concepts which are interactions that are fundamentally independent of the structure of the language. In the rest of the section we give some important structural properties of concept association networks by treating the above mentioned USF Free Association Norm as their representative. A list of these properties are presented below:
nodes: 10618; **edges:** 63788; **avg degree:** 12.01; **max degree:** 332; **diameter:** 7; **power-law exponent:** $\gamma \sim 2.6$; **CC:** 0.1871

Diameter. Distance of the longest of all the shortest paths in the graph. Typically this is log order of the size of the graph. This means that the number of concepts can grow exponentially without changing the diameter much. However this is not unique for concept associations, this is true of any random graph

[1] http://wordnet.princeton.edu

[2] http://w3.usf.edu/FreeAssociation

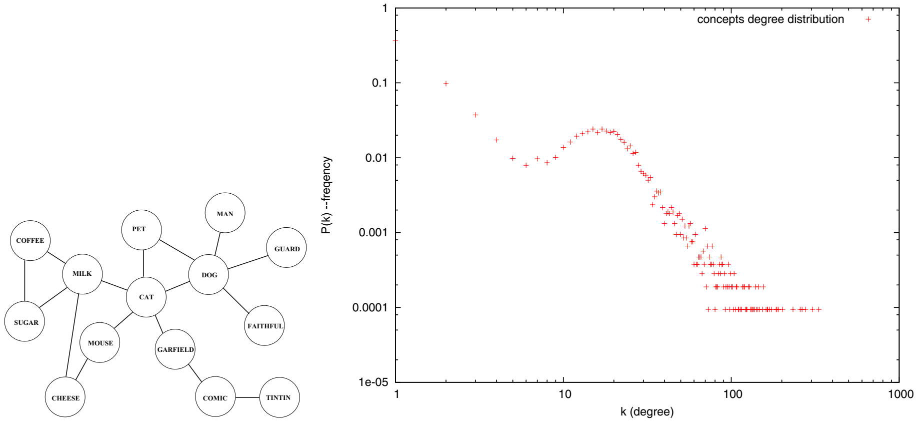


Fig. 1. An illustration of unweighted and undirected concept associations and its degree distribution

wherein the connections are unmotivated. However, unlike random graphs, short diameters arise from the hubs nodes and their effect on the overall topology of the graph as described below.

Degree Distribution. Concept association networks are scale-free and that they have a power-law distribution with the approximate value of γ being 2.6. This is shown in Figure 1 as a log-log plot. This could imply that a preferential attachment model as in other scale free networks like (eg: WWW) is responsible for the structure and evolution of concept associations. The high degree nodes –hubs– are some of the salient concepts associated with human life: FOOD, MAN, GOD, MONEY, MARRIAGE etc. Scale-free behaviour of concept associations imply that humans arrive that few thoughts more often than most other thoughts. This could imply that we tend to associate new concepts in terms of these salient concepts and use them as landmarks to navigate the mind-scape. Another important consequence of of the presence of hubs in the graph is a hierarchical organization of the network. This is the case if the subgraph induced by percolating from a hub node in turn has a scale-free distribution. This property is explored in the next section in the context of clustering.

Clustering Coefficient(CC). This property of nodes ensures that local communities are densely connected while still being able to communicate with the rest of the graph. CC of a node ranges from 0 to 1. In natural language, this is a measure of how many of our friends know each other. High CC may not necessarily mean isolated communities and hence longer diameters. Presence of a few random links across such local communities are enough to ensure small diameters.

For CAN we observe high CC for relatively small degree nodes while hub nodes have low CC indicating that the neighborhood of hubs are sparse. This indicates that the hub nodes mainly serve as switches to keep the network connected rather than ensure similarity of nodes. The average CC observed is 0.18. This implies that in a concept neighborhood almost 20% of the concepts are associated with each other; this is a fairly high value. We rationalize this in the following way: there is an incentive to associate concepts as dense neighbourhoods as it adds to subtle variations in terms of associations –subtilty is directly related to richness of concepts. Long range links across dense neighborhoods are the ones that make the concept graph navigable with relatively short path lengths and hence help in relating distant concepts. Thus there is a cohesion between these two different forces –*abstraction* and *detail* provided respectively by hubs and densely connected neighbourhoods– resulting in forging associations that are rich and diverse.

4 Clustering of Concepts

In the following, we present our heuristic for clustering by taking into account the scale-free property and the high CC of this network and compare properties of clusters generated with those that result from spectral clustering.

4.1 Our Clustering Heuristic

We consider hub nodes as the starting point of this heuristic. To begin with n hub nodes are labelled as belonging to its own cluster $\mathcal{C}_i$. For each unlabelled node u in the graph, we explore its neighborhood and find the highest degree node v. If it is labelled $\mathcal{C}_i$, u is assigned the same label. If v is unlabelled and if degree(v) $\geq$degree(u) we proceed to explore the neighborhood of v in order to assign it a label. Another possibility is v is unlabelled and degree(v) $<$degree(u). In this case we assign u to a default cluster called $\mathcal{C}_0$. After the nodes are exhausted, this cluster is handled separately. Ideally we expect very few nodes to be in $\mathcal{C}_0$. If not, a possible reason could be that some hub nodes that were not considered as as separate clusters have ended up in $\mathcal{C}_i$. Another reason could be the presence of small components that are detached from the main giant cluster. These can be rectified easily by including these hubs as separate clusters. Another way to resolve the nodes in the default cluster would be by labelling them nodes based on the shortest path to one of the hubs of the initialization step. In our experiments we encountered only few nodes in the default cluster and used this rule to resolve them.

The logic behind our heuristic is the following: As mentioned in the context of clustering coefficient, nodes with similar degrees in the neighborhood tend to give variations to a common theme whereas higher degree nodes tend to represent abstraction of concepts. Our heuristic ensures that label in a hub node percolates down to lower degree nodes in accordance with the above hypothesis.

4.2 Comparison between Our Heuristic and Spectral Clustering

A comparison between spectral clustering [7] and our heuristic is shown in Figure 2 as log-log plots of cluster degree distribution. For our discussions we have taken the

$n = 10$. As mentioned above, hubs correspond to concepts that are indicative of high recall. We are interested in 1. identifying a small number of leading concepts which serves as abstractions for other concepts, and 2. illustrating the nature of such clusters. In line with the hierarchical nature of concept associations, rather than having too many clusters, we prefer to cluster concepts into small number of clusters which lend themselves to sub-clustering. This apart there is no significance to our choice of n.

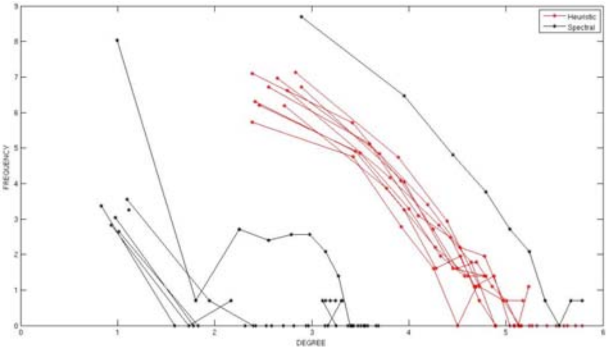


Fig. 2. Degree distribution of clusters: heuristic Vs spectral. Closely occurring plots in the middle correspond to clusters generated by our heuristic.

It is clear from the figure that spectral clustering does not preserve scale-free characteristics within clusters. Moreover, the sizes of clusters are uneven. A giant cluster is formed out of approximately 6000 nodes and another large cluster with around 3000 nodes; remaining 1000 odd nodes form 8 small clusters. Scale-free property is preserved only for the giant cluster[3] for the simple reason that it is almost the entire graph. Unlike this, our heuristic splits the original graph into roughly equal sized clusters and each one has scale-free distribution with the same power-law exponent that applies for the whole graph. Thus the clusters from our approach are self-similar to the whole network. One could argue that a 'cognitive structure' is preserved in the clusters generated by our approach. Further our heuristic lends itself to power-law preserving recursive sub-clustering thereby imparting a hierarchical view to the whole network.

4.3 Discussions

The difference in the cluster properties generated by our heuristic and spectral clustering may be as a result of the latter's nature. In essence spectral clustering is a series of random walks to estimate cluster boundaries –walks are contained within strongly connected components and rarely tend to take connecting

[3] Clustering this further with the same method results in unevenly fragmented clusters and do not conform to scale-free distribution.

bridges. One starts with various ‘seeds’ to begin the random walk and sees the nodes that are reached eventually and thereby identify clusters. We believe that the difference is due to the fact that random walks are an unnatural means to navigate cognitive associations. Further studies are required to quantify this and are part of our on-going efforts.

5 Conclusions

In this work we studied an important aspect of concept associations — grouping of similar concepts. We showed that the structural properties of CAN’s help group similar and related concepts better than a conventional mechanism like spectral clustering. Towards this end we proposed a clustering heuristic that accomplishes this task and explained the rationale behind its design. As concept associations are a manifestation of human cognition and thought process, we believe that further and deeper study of their network representations will enhance our understanding of cognition.

References

1. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Inc., Oxford (2003)
2. Dorogovtsev, S.N., Mendes, J.F.F.: Language as an Evolving Word Web. Proceedings of The Royal Society of London 268(1485), 2603–2606 (2001)
3. Albert, R., Barabási, A.-L.: Emergence of Scaling in Random Networks. Science 289, 509–512 (1999)
4. Watts, D., Strogatz, S.: Collective dynamics of ‘Small-World’ Networks. Nature 293, 440–442 (1998)
5. Cancho, R.F.I., Sole, R.V.: The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences 268, 2261–2266 (2001)
6. Steyvers, M., Tenenbaum, J.B.: The Large-scale Structure of Semantic Network: Statistical analyses and model of semantic growth. Cognitive Science 29(1), 41–78 (2005)
7. Andrew, Y., Ng, M.I.: Jordan and Yair Weiss: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 14, pp. 849–856. MIT Press, Cambridge (2001)

Interactive Rough-Granular Computing in Pattern Recognition

Andrzej Skowron[1], Jan Bazan[2], and Marcin Wojnarski[3]

[1] Institute of Mathematics, Warsaw Unvisersity
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl

[2] Chair of Computer Science, University of Rzeszów
Rejtana 16A, 35-310 Rzeszów, Poland
bazan@univ.rzeszow.pl

[3] Faculty of Mathematics, Informatics and Mechanics, Warsaw University
Banacha 2, 02-097 Warsaw, Poland
mwojnars@ns.onet.pl

Abstract. We discuss the role of generalized approximation spaces and operations on approximation spaces in searching for relevant patterns. The approach is based on interactive rough-granular computing (IRGC) in the WisTech program. We also present results on approximation of complex vague concepts in real-life projects from different domains using the approach based on ontology approximation. Software projects supporting IRGC are reported.

1 Introduction

Approximation spaces belong to the broad spectrum of basic subjects investigated in rough set theory (see, *e.g.*, [2,3,7,13,20,21,22]). Over the years different aspects of approximation spaces were investigated and many generalizations of the approach based on Z. Pawlak's indiscernibility equivalence relation [10,11] were proposed.

We discuss some aspects of approximation spaces in the framework of IRGC [12,23]. These aspects are important from an application point of view, e.g., in searching for approximation of complex concepts (see, e.g., [1,2] and the bibliography in [1]). Two kinds of interactions in searching for relevant approximation spaces are distinguished. The first one is related to attributes which are interpreted as sensors making it possible to store the results of interaction with the environments. The second ones are resulting in constructions of new approximation spaces from already constructed ones. This necessitates some refinements of basic concepts such as signatures of objects, attributes as well as the semantics of what are known as granular formulas. Signatures of objects are mapped to granular formulas constructed over some generic formulas. In this way, we obtain constructive descriptions of granular neighborhoods of objects defined by granular formulas. The constructive definitions of granular neighborhoods also make it possible to define neighborhood size in terms of the size of a granular

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 92–97, 2009.

formula that defines the granular neighborhood. We present a generalization of uncertainty functions assigning granular neighborhoods to objects. The neighborhoods considered so far [20] were treated as subsets of objects. However, in applications these neighborhoods consist of structural objects of different types, often of high order (see, *e.g.*, [6]). These types can be interpreted as structural schemes of perceived complex objects. For example, in the discovery of patterns in spatio-temporal reasoning from data, structural schemes may correspond to types of indiscernibility classes, time widows or their clusters, sequences of time widows or their clusters, sets of sequences of time widows representing processes and their interactions. Granular neighborhoods are considered in these more general cases as models of neighborhoods of higher order structural objects. We also generalize rough inclusion functions. The approximation space definition is included in Section 2. We consider approximation operations as components of approximation spaces. This is due to the fact that, in general, approximation operations may be defined in many different ways, e.g., in the case of coverings [3,11]. Together with approximation spaces we also consider the quality measure of approximation spaces. Approximation spaces are parameterized and their parameters are related to all components of approximation spaces including uncertainty functions, language of granular formulas, rough inclusion functions, approximation operations. Optimization of parameters is based on searching for (semi)optimal approximation space, i.e., for an approximation space for which the chosen quality measure takes (semi)optimal value on the set of objects under consideration. We propose to use the quality measures based on a variation of the minimal description length principle (MDL) [14]. The discussed issues are important in solving different data mining tasks, (*e.g.*, in searching for approximation of complex concepts [1] or in process mining [8]). Finally, software systems supporting IRGC in our projects are reported.

2 Approximation Spaces

Approximation spaces can be treated as granules used for concept approximation. They are some special parameterized relational structures. Tuning of parameters makes it possible to search for relevant approximation spaces relative to given concepts. In this section, we discuss a generalization of definition of approximation space from [20]) introduced in [22].

Lt us assume $P_\omega(U^*)=\bigcup_{i\geq 1} P^i(U^*)$, where $P^1(U^*)=P(U^*)$ and $P^{i+1}(U^*)=P(P^i(U^*)))$ for $i \geq 1$. If $X \in P_\omega(U^*)$ and $U \subseteq U^*$ then by $X \upharpoonright U$ we denote the set defined as follows (i) if $X \in P(U^*)$ then $X \upharpoonright U = X \cap U$ and (ii) for any $i \geq 1$ if $X \in P^{i+1}(U^*)$ then $X \upharpoonright U = \{Y \upharpoonright U : Y \in X\}$.

Definition 1. *An approximation space over a set of attributes A is a system*

$$AS = (U, L, I, \nu, LOW, UPP)\,,$$

where

- *U is a sample of objects with known signatures relative to a given set of attributes A,*

- *L is a language of granular formulas defined over atomic formulas corresponding to generic formulas from signatures* [22],
- *$I : U^* \rightarrow P_\omega(U^*)$ is an uncertainty function, where $U^* \supseteq U$ and the set U^* is such that for any object $u \in U^*$ the signature $Inf_A(u)$ of u relative to A can be obtained (as the result of sensory measurements on u); we assume that the granular neighborhood $I(u)$ is defined by $Inf_A(u)$, i.e., $I(u)$ is defined by a granular formula α selected from L by $Inf_A(u)$* [22],
- *$\nu : P_\omega(U^*) \times P_\omega(U^*) \rightarrow [0,1]$ is a rough inclusion function,*
- *LOW and UPP are the lower approximation operation and the upper approximation operation, respectively, defined on elements from $P_\omega(U^*)$ with values in $P_\omega(U^*)$ such that*
  1. *$\nu(LOW(X), UPP(X)) = 1$ for any $X \in P_\omega(U^*)$,*
  2. *$LOW(X) \upharpoonright U$ is included in $X \upharpoonright U$ to a degree at least deg, i.e., $\nu(LOW(X) \upharpoonright U, X \upharpoonright U)) \geq deg$ for any $X \in P_\omega(U^*)$,*
  3. *$X \upharpoonright U$ is included in $UPP(X) \upharpoonright U$ to a degree at least deg, i.e., $\nu(X \upharpoonright U, UPP(X) \upharpoonright U) \geq deg$ for any $X \in P_\omega(U^*)$,*

  where deg is a given threshold from the interval $[0,1]$.

Figure 1 illustrates this computation process of the granular formulas and the granular neighborhoods by means of granular formulas.

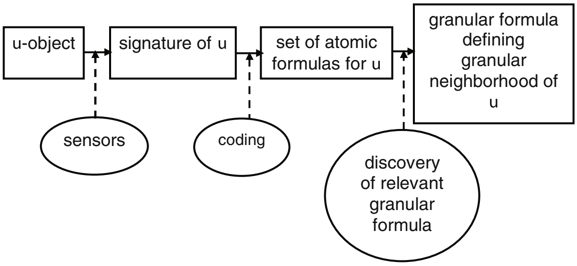

u-object
signature of u
set of atomic formulas for u
granular formula defining granular neighborhood of u
sensors
coding
discovery of relevant granular formula

Fig. 1. From objects to granular neighborhoods

Note that this satisfiability is defined for all objects from U^*, not only for objects from a sample $U \subseteq U^*$. Moreover, in computation of $I(u)$ for $u \in U^*$ can be used not only the signature of the object u currently perceived but also already known signatures of some other objects from a sample U as well as higher level knowledge previously induced or acquired from experts. The details explaining how granular formulas are hierarchically constructed will be explained elsewhere. Here, we only give an illustrative example. Let us assume that granular formulas on the hierarchical level j were defined. They can be treated as objects of an information system. For example, they can represent time windows. In this case, they are equal to the sets of pairs (α_i, α_{i+1}) of formulas, for $i = 1, \ldots, T-1$, where T is the time window length and each α_i is the conjunction of descriptors (atomic formulas) for the object observed at time corresponding to i. Next, formulas from a new language are used for defining properties of such granular formulas (e.g., analogously to indiscernibility classes of information systems).

In our example, with time windows, formulas from such a language describe properties of time windows. The defined sets are granular formulas on the level $j+1$. Let us consider some illustrative examples explaining how can be defined the semantics of granular formulas. If α is an atomic formula than $\|\alpha\|_{U^*} = \{u \in U^* : u \models \alpha\}$. If $\alpha \in L_0$ then $\|\alpha\|_{U^*}$ is defined using semantics of connectives selected for construction of formulas from L_0. If $\|\alpha\|_{U^*}$ is defined for $\alpha \in L$ for some language of granular formulas and $\beta \in P(L)$ then $\|\beta\|_{U^*} = \{\|\alpha\|_{U^*} : \alpha \in \beta\}$. However, one can also assume $\|\beta\|_{U^*} = \bigcup\{\|\alpha\|_{U^*} : \alpha \in \beta\}$. In the latter case, the defined granular neighborhood is from $P(U^*)$ but information on semantics of the granular neighborhood components are lost.

Approximation spaces are parameterized and their parameters are related to all components of approximation spaces including uncertainty functions, language of granular formulas, rough inclusion functions, approximation operations. In this way, we obtain a family $APPROX\_SPACES$ of approximation spaces corresponding to different values of parameters. Optimization of parameters is based on searching for the (semi)optimal approximation space, i.e., for an approximation space with the (semi)optimal value of the quality measure on the set of objects under consideration. For the given family $APPROX\_SPACES$, a quality measure Q is a function which assigns to any $AS \in APPROX\_SPACES$ and $X \in P_\omega(U^*)$ a nonnegative number $Q(AS, X)$ called an approximation quality of X in AS[1]. For a given triple (APPROX_SPACES, Q, X), where $X \subseteq U^*$, we consider the optimization problem. This is the searching problem for (semi)optimal approximation space $AS_0 \in APPROX\_SPACES$ such that $Q(AS_0, X) = inf\{Q(AS, X) : AS \in APPROX\_SPACES\}$. We use quality measures based on different versions of the minimal length principle [14].

3 Software Systems

In this section, we present a short information about two software platforms Rough Set Interactive Classification Engine (RoughICE) [15] and TunedIT [19] which are supporting our projects based on IRGC.

RoughICE is a software platform supporting the approximation of spatio-temporal complex concepts in the given concept ontology acquired in the dialogue with the user. RoughICE is freely available on the web side [15]. The underlying algorithmic methods, especially for generating reducts and rules, discretization and decomposition, are outgrows of our previous tools such as RSES [16] and RSESlib [17]. RoughICE software and underlying computational methods have been successfully applied in different data mining projects (e.g., in mining traffic data and medical data; for details see [1] and the literature cited in [1]).

TunedIT platform, launched recently by members of our research group, facilitates sharing, evaluation and comparison of data-mining and machine-learning algorithms. The resources used in our experiments – algorithms and datasets in particular – will be shared on TunedIT website. This website already contains

[1] In the case of classification $X \in P^2(U^*)$ and X is a partition of U^*.

many publicly available datasets and algorithms, as well as performance data for nearly 100 algorithms tested on numerous datasets - these include the algorithms from Weka, Rseslib libraries, and the datasets from UCI Machine Learning Repository. Everyone can contribute new resources and results. TunedIT is composed of three complementary modules: TunedTester, Repository and Knowledge Base. TunedIT may help researchers design repeatable experiments and generate reproducible results. It may be particularly useful when conducting experiments intended for publication, as reproducibility of experimental results is the essential factor that determines research value of the paper. TunedIT helps also in dissemination of new ideas and findings. Every researcher may upload his implementations, datasets and documents into Repository, so that other users can find them easily and employ in their own research.

For more details on RoughICE and TunedIT the reader is referred to [15] and [19], respectively.

4 Conclusions

We discussed a generalization of approximation spaces based on granular formulas and neighborhoods. Efficient searching strategies for relevant approximation spaces are crucial for application (e.g., in searching for approximation of complex concepts or in process mining). Some of such strategies based on ontology approximation were already elaborated and implemented for solving real-life problems (see, e.g., [1]). We are working on the project aiming at developing new strategies for hierarchical modeling as well as methods based on adaptive interactive granular computations [5] for selection and construction of features. These features are next used as relevant components of approximation spaces, in particular granular neighborhoods of approximation spaces.

Acknowledgements

The research has been partially supported by the grants N N516 368334 and N N516 077837 from Ministry of Science and Higher Education of the Republic of Poland.

References

1. Bazan, J.G.: Hierarchical classifiers for complex spatio-temporal concepts. In: Peters, J.F., Skowron, A., Rybiński, H. (eds.) Transactions on Rough Sets IX. LNCS, vol. 5390, pp. 474–750. Springer, Heidelberg (2008)
2. Bazan, J., Skowron, A., Swiniarski, R.: Rough sets and vague concept approximation: From sample approximation to adaptive learning. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets V. LNCS, vol. 4100, pp. 39–62. Springer, Heidelberg (2006)
3. Grzymała-Busse, J., Rząsa, W.: Local and global approximations for incomplete data. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets VIII. LNCS, vol. 5084, pp. 21–34. Springer, Heidelberg (2008)

4. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Heidelberg (2008)
5. Jankowski, A., Skowron, A.: Logic for artificial intelligence: The Rasiowa-Pawlak school perspective. In: Ehrenfeucht, A., Marek, V., Srebrny, M. (eds.) Andrzej Mostowski and Foundational Studies, pp. 106–143. IOS Press, Amsterdam (2008)
6. Ng, K.S., Lloyd, J.W., Uther, W.T.B.: Probabilistic modelling, inference and learning using logical theories. Annals of Mathematics and Artificial Intelligence 54(1-3), 159–205 (2008)
7. Nguyen, H.S.: Approximate Boolean Reasoning: Foundations and Applications in Data Mining. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets V. LNCS, vol. 4100, pp. 344–523. Springer, Heidelberg (2006)
8. Nguyen, H.S., Jankowski, A., Skowron, A., Stepaniuk, J., Szczuka, M.: Discovery of process models from data and domain knowledge: A rough-granular approach. In: Yao, J.T. (ed.) Novel Developments in Granular Computing: Applications for Advanced Human Reasoning and Soft Computation, IGI Global, Hershey (accepted)
9. Pal, S.K., Shankar, B.U., Mitra, P.: Granular computing, rough entropy and object extraction. Pattern Recognition Letters 26(16), 2509–2517 (2005)
10. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving, vol. 9. Kluwer Academic Publishers, Dordrecht (1991)
11. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences 177(1), 3–27 (2007); Rough sets: Some extensions. Information Sciences 177(1), 28–40
12. Pedrycz, W., Skowron, A., Kreinovich, V. (eds.): Handbook of Granular Computing. John Wiley & Sons, New York (2008)
13. Peters, J., Henry, C.: Reinforcement learning with approximation spaces. Fundamenta Informaticae 71(2,3), 323–349 (2006)
14. Rissanen, J.: Modeling by shortest data description. Automatica 14, 465–471 (1978)
15. The Rough Set Interactive Classificstion Engine (RoughICE), `http://logic.mimuw.edu.pl/~bazan/roughice`
16. The Rough Set Exploration System (RSES), `http://logic.mimuw.edu.pl/~rses`
17. The RSES-lib project, `http://rsproject.mimuw.edu.pl`
18. The road simulator, `http://logic.mimuw.edu.pl/~bazan/simulator`
19. The TunedIT platform, `http://tunedit.org/`
20. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae 27, 245–253 (1996)
21. Skowron, A., Stepaniuk, J., Peters, J., Swiniarski, R.: Calculi of approximation spaces. Fundamenta Informaticae 72(1-3), 363–378 (2006)
22. Skowron, A., Stepaniuk, J., Peters, J.F., Swiniarski, R.: Approximation spaces revisited. In: Proceedings of the Concurrency, Specification & Programming 2009 (CS&P 2009), Przegorzały, Kraków, Poland, September 28-30, pp. 538–549. Warsaw University (2009)
23. Skowron, A., Szczuka, M.: Toward interactive computations: A rough-granular approach. In: Koronacki, J., Wierzchon, S.T., Ras, Z.W., Kacprzyk, J. (eds.) Advances in Machine learning II, Dedicated to the memory of Ryszard Michalski. Studies in Computational Intelligence, vol. 263, pp. 1–20. Springer, Heidelberg (2009) (in print)
24. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences 46, 39–59 (1993)

Application of Neural Networks in Preform Design of Aluminium Upsetting Process Considering Different Interfacial Frictional Conditions

Ajay Kumar Kaviti[1,*], K.K. Pathak[2], and M.S. Hora[3]

[1] Department of Mechanical Engineering, SISTec, Bhopal (MP) India
ajaykaviti@yahoo.com
[2] Advanced Materials and Processes Research Institute (CSIR), Bhopal (MP) India
[3] Department of Applied Mechanics, MANIT, Bhopal (MP) India

Abstract. Design of the optimum preform for near net shape manufacturing is a crucial step in upsetting process design. In this study, the same is arrived at using artificial neural networks (ANN) considering different interfacial friction conditions between top and bottom die and billet interface. Back propagation neural networks is trained based on finite element analysis results considering ten different interfacial friction conditions and varying geometrical and processing parameters, to predict the optimum preform for commercial Aluminium. Neural network predictions are verified for three new problems of commercial aluminum and observed that these are in close match with their simulation counterparts.

Keywords: Artificial neural network; Preform; finite element; upsetting; deformation.

1 Introduction

Upsetting is an important metal forming operation. It is a class of bulk forming operation where large deformation is given to the material for shape and property modification. The major issue, which restricts imparting large deformation to the billet is the bulging induced tensile stress which later result in cracking. Bulge is also undesirable from near net shape manufacturing point of view as it will require secondary processing like trimming. To obtain the near net shape, preform design of the billets is a powerful solution. Considerable amount of literature are available on the preform design in forging process.

Roy et al. (1994) report application of neural networks in interpolation of preform shapes in plane strain forgings. Ranatunga et al. (1996) present preform designing techniques based on the upper bound elemental technique (UBET) with evidence of effective material usage and extended overall die-life. Lee et al. (1997) report application of an upper-bound elemental technique in preform design for asymmetric forging which is validated through experiments. Liu et al. (1998) present a preform design

[*] Corresponding Author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 98–105, 2009.

method which combines the FEM & upper bound based reverse simulation technique. The billet designed using this technique achieves a final forging with minimum flash. Ko et al. (1999) describe a new method of preform design in muti-stage metal forming processes considering workability limited by ductile fracture. Neural networks and Taguchi method are used for minimizing the objective function. Srikanth et al. (2000) present a continuum sensitivity analysis approach for the computation of the shape sensitivity, which is later used for the purpose of preform design and shape optimization in forging process. Chang et al. (2000) propose reverse simulation approach clubbed with finite element analyses for preform design. Bramley et al. (2001) report a new method named as tetrahedral upper bound analysis which enables a more realistic flow simulation to be achieved. Antonio et al. (2002) presents an inverse engineering formulation together with evolutionary search schemes for forging preform design. Shim et al. (2003) presents optimal preform design for 3D free forgings using sensitivity approach and FEM. Tomov et al. (2004) reports preform design of axisymmetric forging using FE software FORM-2D. Ou et al. (2004) reports finite element (FE) based forging simulation and optimization approach in order to achieve net-shape forging production for aero engine components. Effects of die-elastic deformation, thermal distortion and press-elasticity were considered. Poursina et al. (2004) proposes a FEM and GA based preform design procedure for axisymmetric forgings in view to achieve high quality products. Thiyagarajan et al. (2005) presents a 3-D preform shape optimization method for the forging process using the reduced basis technique. Repalle et al. (2005) presents reliability-based optimization method for preform shape design in the forging. Antonio et al. (2005), reports an inverse approach for preform design of forged components under minimal energy consumption using FEM and genetic algorithms. Park and Hwang et al. (2007) reports preform design for precision forging of rib type aerospace components using finite element analysis. Poshala et al. (2008) carried out formability analysis and its experimental validations for aluminum preforms using neural network. Haluk Tumer et al. (2008) optimised die and preform to minimize hardness distribution in back extrusion process using Nelder-Mead search algorithm integrated with the finite element model. Although substantial literature on preform design is available, they address it as individual problem considering one or few parameters.

The main objective of this study is to devise a generalized procedure of preform design considering various parameters. For this, neural network has been used for preform design of the upsetting process. In this study effect of critical factors including different preform shapes, interfacial friction conditions, and their effect on the final deformed profiles are studied using FE simulation. Ten cases of different interfacial friction conditions are considered for the same. Based on the simulation results, a back propagation neural network is trained to provide guidelines for selection of parameters to result in near net shape manufacturing. Neural network predictions are verified with three numerical examples for commercial Aluminium.

2 Artificial Neural Networks

Artificial neural network attempts to imitate the learning activities of the brain. In an artificial neural network (ANN), the artificial neuron or the processing unit may have several input paths corresponding to the dendrites in the biological neuron. The units combine usually, by a simple summation, the weighted values of these paths. The weighted value is passed to the neuron, where it is modified by threshold function. The modified value is directly presented to the next neuron. In Fig.1 a 3-4-2 feed forward back propagation artificial neural network is shown. The connections between various neurons are strengthened or weakened according to the experiences obtained during the training. The algorithm for training the back propagation neural network can be explained in the following steps-

Step1 – Select the number of hidden layers, number of iterations, tolerance of the mean square error and initialize the weights and bias functions.

Step2 – Present the normalized input –output pattern sets to the network. At each node of the network except the nodes on input layer, calculate the weighted sum of the inputs, add bias and apply sigmoid function

Step3-Calculate total mean error. If error is less than permissible limit, the training process is stopped. Otherwise,

Step4 –Change the weights and bias values based on generalized delta rule and repeat step 2.

The mathematical formulations of training the network can be found in Ref. 21.

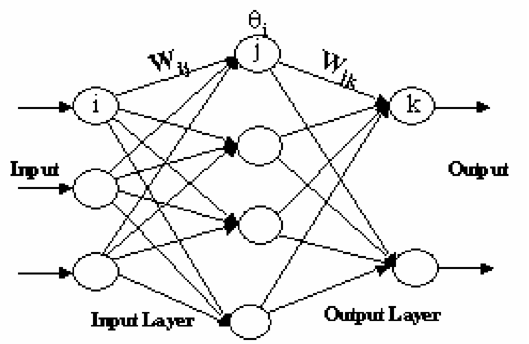


Fig. 1. Neural network

3 Methodology

In Fig.2, schematic undeformed and deformed billets are shown. Let top, middle and bottom diameters of these billets be a, b c, and a_1, b_1, c_1 respectively. Their diameter ratios with respect to top diameter, can be expressed as R_1=b/a, R_2=c/a and r_1=b_1/a_1, r_2=c_1/a_1. It is obvious that for near net shape manufacturing, r_1 and r_2 should be one. Since deformed profiles depend on geometrical and frictional conditions, large numbers of variation of these parameters are accounted. Ten sets of interfacial frictional parameters and 38 sets of geometrical conditions making total 380 combinations are considered in this study for commercial Aluminium. Finite element simulations of these cases are carried out to obtain the deformation behavior. Based on these results,

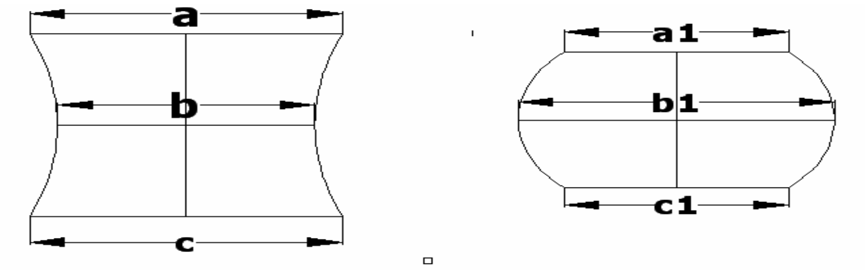


Fig. 2. Initial and final shapes of billet

back propagation neural networks are trained to predict desired preform for given f_t, and f_b values to result in near net shape upsetting.

4 Geometrical, Material and Processing Parameters

Cylindrical specimens of 40 mm top diameter and 40 mm height are used for simulation studies of commercial Aluminium. The central and bottom diameters are considered as 28, 30, 32, 34 36, 38 and 39 mm. In this way center and top diameter ratio and bottom and top diameter ratio (R_1 and R_2), also named as preform ratios, comes out to be 0.7, 0.75, 0.8, 0.85, 0.9, 0.95 and 0.975 respectively. Ten combinations of interfacial frictions, Coulomb friction, at top and bottom surfaces of billet and platens considered for simulation studies are given in Table 1.

Table 1. Frictional conditions at die and billet interface

S.No	f_t(Friction between top die and billet interface)	f_b(Friction between bottom die and billet interface)
1	0.1	0.1
2	0.2	0.1
3	0.2	0.2
4	0.3	0.1
5	0.3	0.2
6	0.3	0.3
7	0.4	0.1
8	0.4	0.2
9	0.4	0.3
10	0.4	0.4

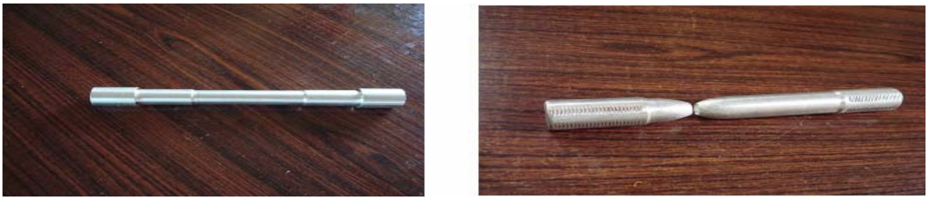


Fig. 3. Tensile specimens (Before & after test)

Table 2. Material Properties

Properties	Commercial Aluminium
Youngs modulus (E)MPa	$7x10^4$
Poisson's ratio (ν)	0.33
Strengthcoefficient (K)MPa	225.4
Hardening exponent (n)	0.095

The 38 cases of geometric parameters accounted in the study. Material properties of commercial Aluminium have been obtained by conducting tensile tests. Specimens of gauge length 80 mm, prepared as per ASTM standard, are tested in a Shimadzu make Universal Testing Machine (UTM). The test and tested specimens of commercial aluminum are shown in Fig.3. The engineering stress & strain are converted into their true counterparts using standard relationships (Kalpakjian and Schmid ,2004). Based on these results, material modeling is carried out. The post yielding behaviour is modeled using the power law equation (Meyers and Chawla, 1997):

$$\sigma = k\varepsilon^n$$

Where k is the strength coefficient and n is the hardening exponent. The material properties evaluated and adapted for FE simulation are given in Table 2.

5 FE Simulation

Finite element analyses of the upsetting process are carried out using MSC.Marc software (Ref 22). Curved profiles of specimens are modeled as arcs between top,

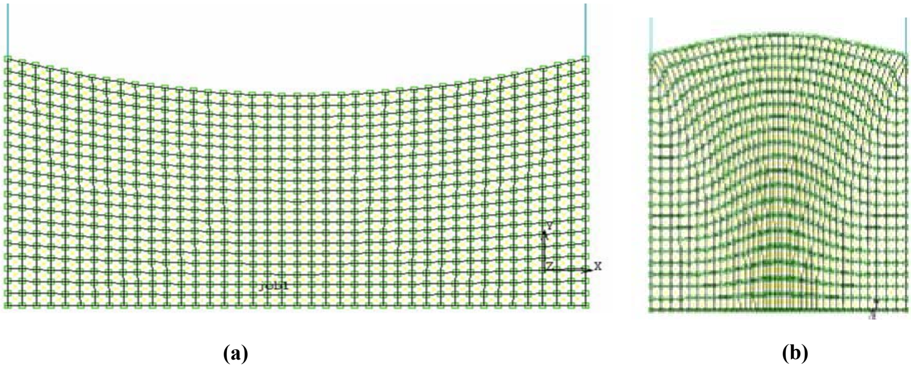
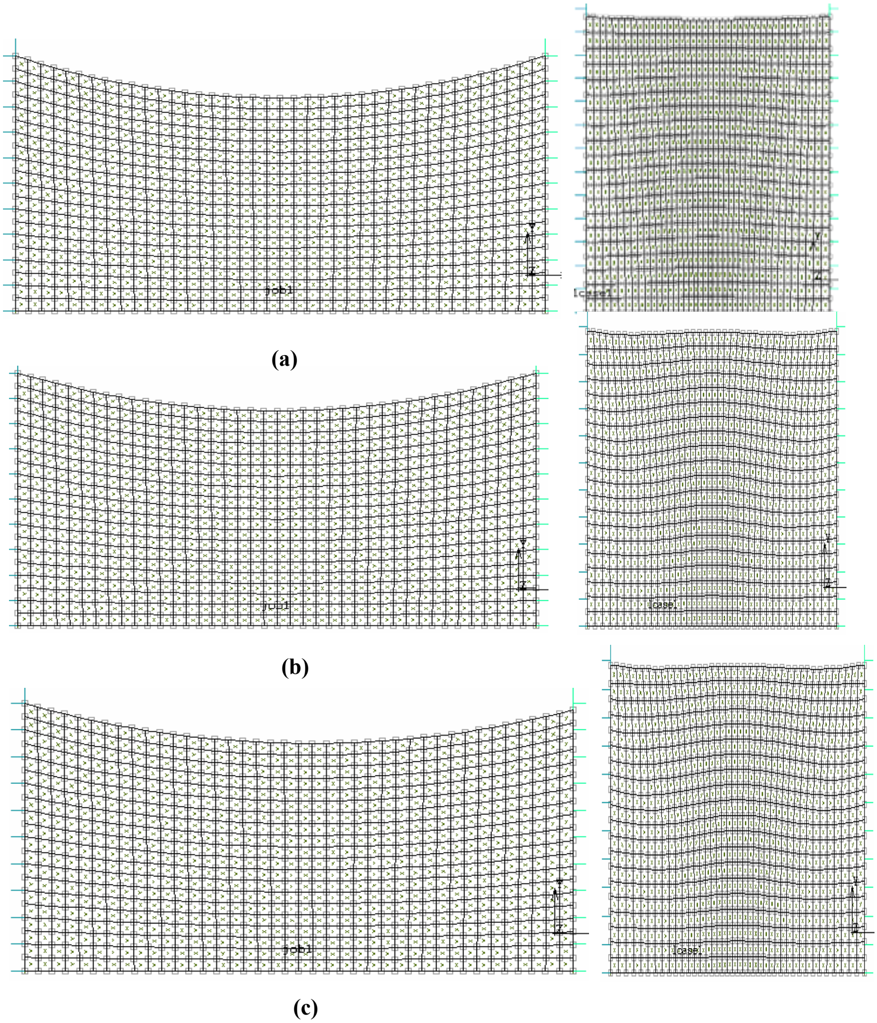


Fig. 4. FEM models (a) before deformation (b) after deformation

middle and bottom diameters using ARC command of the software. Taking advantage of the symmetrical conditions, axisymmetrical formulation is adopted. Four nodded quadrilateral elements are used for the FE modeling. There are 800 elements and 861 nodes in the model. Considering the variation in 38 geometrical cases and ten cases of frictional conditions, total 380 cases are simulated for commercial Aluminium. Punch and die are modeled as rigid bodies. Bottom die is fixed whereas punch is movable which is given the displacement boundary condition. All the commercial Aluminium billets are identically deformed to final height of 28 mm viz. 30 % reduction in height. A typical FE and deformed models are shown in Fig.4. Geometrical parameters of deformed and undeformed conditions for all the 380 cases are recorded separately for commercial aluminum.



(a)

(b)

(c)

Fig. 5. Initial and deformed shapes of Aluminium (a) f_t= 0.28, f_b= 0.28, R_1= 0.835, R_2=1 (b) f_t= 0.2, f_b=0.2, R_1=0.86, R_2=1 (c) f_t=0.35, f_b=0.25 R_1=0.83,R_2=0.975

6 Numerical Validation

FE results in terms of diameter ratios are used for training neural networks. One 4-6-2 back propagation neural network for commercial Aluminium has been used for the training. f_t, f_b, r_1, r_2 are input and R_1,and R_2 are output parameters. The error limit is 0.002 and it took 1447695 epochs to converge the desired limit. The trained network is tested for three new problems of commercial aluminum upsetting to show the efficacy of the neural network predictions. The input parameters for them are given in Table3. The predicted preforms (R_1 and R_2 values) are used for validation through FE simulation. The 'r_1 and r_2' values predicted are very close to the near net shape manufacturing. Maximum error is 1% which is very less. The initial and final deformed meshes for these cases are shown in Fig5. It can be observed that deformed profiles are close to the near net shapes of perfect cylinders.

Table 3. Numerical Validation of ANN for commercial Aluminium

S.No	f_t	f_b	R_1	R_2	$r_{1(Actual)}$	$r_{1(FEM)}$	%Error	$r_{2(Actual)}$	$r_{2(FEM)}$	%Error
1	0.28	0.28	0.835	1	1	0.99	1	1	0.999	0.1
2	0.20	0.20	0.86	1	1	0.998	0.2	1	0.999	0.1
3	0.35	0.25	0.825	0.975	1	0.995	0.5	1	1.003	0.3

7 Conclusion

In this study artificial neural networks have been used for the design of preforms for the cylindrical billet upsetting. Based on the results of 380 FE simulations a back propagation neural network is trained for commercial Aluminium. Trained networks are first verified with three numerical examples. It is found that simulation and network predictions are in close match. This study also demonstrates that ANN can be effectively used for preform design. It is hoped, this study will help design engineers in fast and reliable predictions of optimum preforms under different frictional conditions for net shape manufacturing.

References

1. Antonio, C.A.C., Dourado, N.M.: Metal-forming process optimisation by inverse evolutionary search. Journal of Materials Processing Technology 121(2-3), 403–413 (2002)
2. Antonio, C.C., Castro, C.F., Sousa, L.C.: Eliminating forging defects using genetic algorithms. Materials and Manufacturing Processes 20(3), 509–522 (2005)
3. Bramley, A.: UBET and TEUBA: fast methods for forging simulation and preform design. Journal of Materials Processing Technology 116(1), 62–66 (2001)
4. Chang, C.C., Bramley, A.N.: Forging preform design using a reverse simulation approach with the upper bound finite element procedure. Proceedings of The Institution of Mechanical Engineers Part C: Journal of Mechanical Engineering Science 214(1), 127–136 (2000)
5. Tumer, H., Sonmez, F.O.: Optimum shape design of die and preform for improved hardness distribution in cold forged parts. Journal of Materials Processing Technology (Article in the Press)

6. Ko, D.C., Kim, D.H., Kim, B.M.: Application of artificial neural network and Taguchi method to preform design in metal forming considering workability. International Journal of Machine Tools & Manufacture 39(5), 771–785 (1999)
7. Lee, J.H., Kim, Y.H., Bae, W.B.: An upper-bound elemental technique approach to the process design of asymmetric forgings. Journal of Materials Processing Technology 72(1), 141–151 (1997)
8. Liu, Q.B., Wu, S.C., Sun, S.: Preform design in axisymmetric forging by a new FEM-UBET method. Journal of Materials Processing Technology 74(1-3), 218–222 (1998)
9. Meyers, M.A., Chawla, K.K.: Mechanical Behaviour of Materials. Prentice-Hall, Englewood Cliffs (1999)
10. Ou, H., Lan, J., Armstrong, C.G.: An FE simulation and optimisation approach for the forging of aeroengine components. Journal of Materials Processing Technology 151(1-3), 208–216 (2004)
11. Park, J.J., Hwang, H.S.: Preform design for precision forging of an asymmetric rib-web type component. Journal of Materials Processing Technology 187, 595–599 (2007)
12. Poshala, G., Ganesan, P.: An analysis of formability of aluminium preforms using neural network. Journal of Materials Processing Technology 205(1-3), 272–282 (2008)
13. Poursina, M., Antonio, C.A.C., Castro, C.F.: Preform optimal design in metal forging using genetic algorithms. Engineering Computations 21(5-6), 631–650 (2004)
14. Ranatunga, V., Gunasekera, J.S.: UBET-based numerical modeling of bulk deformation processes. Journal of Materials Engineering and Performance 15(1), 47–52 (2006)
15. Repalle, J., Grandhi, R.V.: Reliability-based preform shape design in forging. Communications in Numerical Methods In Engineering 21(11), 607–617 (2005)
16. Roy, R., Chodnikiewicz, K., Balendra, R.: Interpolation of Forging preform shapes using neural networks. Journal of Materials Processing Technology 45(1-4), 695–702 (1994)
17. Thiyagarajan, N., Grandhi, R.V.: Multi-level design process for 3-D preform shape optimization in metal forming. Journal of Materials Processing Technology 170(1-2), 421–429 (2005)
18. Tomov, B.I., Gagov, V.I., Radev, R.H.: Numerical simulations of hot die forging processes using finite element method. Journal of Materials Processing Technology 153, 352–358 (2004)
19. Shim, H.: Optimal preform design for the free forging of 3D shapes by the sensitivity method. Journal of Materials Processing Technology 134(1), 99–107 (2003)
20. Srikanth, A., Zabaras, N.: Shape optimization and preform design in metal forming processes. Computer Methods in Applied Mechanics and Engineering 190(13-14), 1859–1901 (2000)
21. Hertz, J., Krogh, A.: Introduction to the Theory of Neural Networks: Addison- Wesley Publishing Company. Addison-Wesley Publishing Company, Reading (1991)
22. User's manual, MSC. Marc, MSC Software Corporation, Santa Ana, California 92707 USA (2005)

Case Indexing Using PSO and ANN in Real Time Strategy Games

Peng Huo, Simon Chi-Keung Shiu, HaiBo Wang, and Ben Niu

Department of Computing, The Hong Kong Polytechnic University, Kowloon,
Hong Kong, China
csphuo@comp.polyu.edu.hk

Abstract. This paper proposes a case indexing method using particle swarm optimization (PSO) and artificial neural network (ANN) in a defense-style real time strategy (RTS) game. PSO is employed to optimize the placement of cannons to defend the enemy attack. The execution time of PSO ($>$ 100 seconds) is unsatisfied for RTS game. The result of PSO is used as a case indexing of past experience to train ANN. After the training (approximately 30 seconds), ANN can obtain the best cannon placement within 0.05 second. Experimental results demonstrated that this case indexing method using PSO and ANN efficiently speeded up the whole process to satisfy the requirement in RTS game.

1 Introduction

In recent years, computer games have been developed rapidly and provide an ideal platform for artificial intelligence (AI) research. The application of AI improves the intelligence and attraction of games. Machine learning approach may be a better choice to introduce human-like behavior [1,2]. Previous studies used genetic algorithm (GA) in shooting game [3] and strategical game [4]. Louis [5] developed a case-injected GA in real time strategy (RTS) game.

The GA mimics the natural biological evolution and has been popular. However, previous studies have reported that the drawback of GA is its expensive computational cost, while one advantage of particle swarm optimization (PSO) is its algorithm simplicity [6,7,8].

PSO is similar to GA in a sense but employs different strategies and computational effort. PSO is a relatively recent heuristic search method whose mechanics are inspired by the swarming social behavior of species. PSO has been used to improve playing strategies in board games, such as game Seega [9], African Bao game [10], checker [11]. Our recent study compared the performance of PSO and GA in a tower defense game [12]. The result, however, indicated that the execution time of either PSO or GA is unsatisfied for RTS games.

Artificial neural network (ANN) represents the art of neural information processing. With the hybrid of PSO-ANN approach, we aim to design the games in which the computer players can acquire the knowledge automatically by learning from the previous cases iteratively and memorizing the successful experiences to handle future situations like human being behaves.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 106–115, 2009.

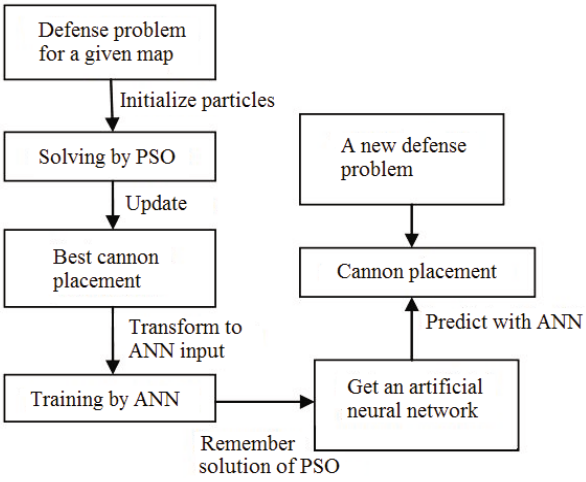


Fig. 1. Flowchart of game simulation

In this study, the optimum cannon placement in the battlefield obtained by PSO was used as a case indexing to train ANN. After training, ANN quickly obtained the optimum cannon placement in a new battlefield by consulting the case-base without going through the PSO process again. Figure 1 shows the flowchart of the game simulation.

2 Methodology

2.1 Scenario of Tower Defense Game

Similar to the concept of Forbus [13] spatial reasoning, battlefields can be divided into different small maps. In this study, the small maps are composed of defense base, barriers (mountains) and canyons. Canyon is the valley between hills or mountains and can be travelled through by the enemy. The number of canyons could be one, two, three, and four or determined randomly in the small maps as shown in Figure 2A-E. Dividing the battlefield into small maps gives two advantages. First, the fitness function can give a better solution easily. Second, it can increase the efficiency of PSO. Open area without canyons was not consider as it will not affect the solution and it will increase the run time as the search space increases.

Each battlefield map is digitalized as a 50×50 matrix. The points in barriers are set with a value of -1. In the areas of the barriers, the cannon cannot be placed. The points in the blank area (or the open area) are set with a value of 0, in which the cannons can be placed. When one cannon is placed at a point, the value of this point will be altered to 1.

A scenario of strategic defense game is chosen as a test bed (Figure 2F). The game description is as follows.

1. Two teams are created in the battlefield. One is the attack team (enemy). Another one is the defense team (player).

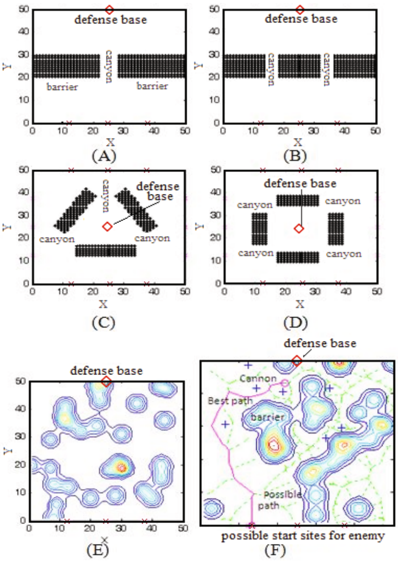


Fig. 2. Small battlefield maps with different canyons. (A) one canyon, (B) two canyons, (C) three canyons, (D) four canyons, (E) canyons determined randomly. (F) A base defense scenario. $\Diamond$: defense base, $\times$: possible start sites of enemy, +: cannon, solid line: best path for enemy to travel, dashed line: possible path for enemy to travel.

2. The enemy must start from one of the possible start sites and move to the base of the defense team along one possible path with a minimum casualty (damage).
3. The defense team is able to place a number of cannons (e.g. seven cannons in this study) in the battlefield to kill the enemy when the enemy is approaching the defense base. Each cannon has a circle cannon-shot range.
4. The goal is to search an optimum cannon placement with a maximum casualty to the enemy no matter which possible path they choose.

2.2 Learning Cases by Particle Swarm Optimization

PSO is inspired by the social behavior of species such as a flock of migrating birds or a school of fish trying to reach an unknown destination. In PSO, a 'bird' in the flock is considered as a 'particle' presenting a candidate solution in the problem space. The particles in the swarm evolve their position and move to the destination based on their own experience and the experience of other particles [14]. Each particle has a memory to be capable of remembering the best position ever visited by it. Its best position with the best fitness is known as personal best, 'pbest' and the overall best position in the swarm is called as global best, 'gbest'.

The process starts from the initialization of the swarm with N random particles. Suppose that the search space is d-dimensional and the i^{th} particle is represented by a d-dimensional vector X_i=(x_{i1}, x_{i2}, $\ldots$, x_{id}) and its velocity is denoted by V_i=(v_{i1}, v_{i2}, $\ldots$, v_{id}). And it can remember its best previously visited position P_{id}=($p_{i1}, p_{i2}, \ldots, p_{id}$).

If the g^{th} particle is the best particle among all particles, its position is P_{gd}. Then, each particle updates its velocity to catch up with the best particle g according to Eq. 1 and Eq. 2.

$$V_{id}^{t+1} = \omega V_{id}^{t} + c_1\gamma_1(P_{id}^{t} - X_{id}^{t}) + c_2\gamma_2(P_{gd}^{t} - X_{id}^{t}) \tag{1}$$

$$X_{id}^{t+1} = X_{id}^{t} + V_{id}^{t+1} \quad 1 \leq i \leq N \tag{2}$$

where ω is inertia weight. c_1 and c_2 are cognitive acceleration and social acceleration, respectively, usually c_1=c_2=2. γ_1 and γ_2 are random numbers uniformly distributed in the range [0,1].

It is noted that the velocity update (Eq. 1) consists of three components, the previous velocity, the cognitive component, and the social component [15]. The first component is inertia component serving as a memory of the previous movement direction. The second component quantifies the current position with its own best position. The third component compares the current position with the best position in the swarm. Without the first part, all the particles will tend to move toward the same position, that is, the search area is contracting through the generations. Only when the global optimum is within the initial search space, there is a chance for PSO to find the solution. On the other hand, adding the first part, the particles will have a tendency to expand the search space, that is, they have the ability to explore the new area. So they more likely have global search ability. Both the local and global search will benefit solving problems. For different problems, there should be different balances between the local search ability and the global search ability. Considering of this, inertia weight ω is proposed to decrease linearly with generations from 1.4 to 0.5 determined by Eq. 3 [14].

$$\omega_i = 0.5 + 0.9(G - i)/(G - 1) \quad i = 1, 2, \ldots, G \tag{3}$$

where *G* is the number of generations.

In this study, we used seven cannons as an example. The placement of the cannons is treated as a particle. Since the search space (battlefield) is 2-dimensional (X-Y coordinate), the position and velocity of the i^{th} particle are denoted by matrix X_{ijd} and V_{ijd}, respectively, as follows.

$$X_{ijd} = \{x_{ij1}, x_{ij2}, \ldots, x_{ijd}\} \quad j = 1, 2; d = 7$$

$$V_{ijd} = \{v_{ij1}, v_{ij2}, \ldots, v_{ijd}\} \quad j = 1, 2; d = 7$$

where *j* = 1 represents x direction and *j* = 2 represents y direction.

To control the change of particles' velocities, upper and lower boundary is limited to a user-specified value of V_{max} [16], V_{max}=10. In this study, the population size or the number of particles (*P*) equals 50 and the number of generation or iteration cycles (*G*) equals 100.

In our simulation, a limit number of cannons are placed to give maximum damage to the enemy whatever the travelling path they choose. Base on the cannon placement, the best path to travel by the enemy is the one with minimum damage. The enemy will receive a damage value within the shot range of each cannon. The total damage (*D*) is calculated by summing up all individual damages caused by different cannons (Eq. 4), and is used as the fitness value of this simulation. The higher the damage did to the enemy, the better the fitness of the cannon's positions. Cannon positions will be ranked by this fitness value to generate 'pbest' and 'gbest' particles.

$$D = \sum_{k=1}^{N} (\frac{d_i}{v} \times p_i) \quad i = 1, 2, \ldots, n \tag{4}$$

where d_i is the distance that the enemy travels through the cannon-shot range of the i^{th} cannon. v is the velocity of the enemy. p_i is cannon power of the i^{th} cannon. *n* is the number of cannons.

2.3 Case Indexing for ANN

In RTS games, one problem is that the long execution time of an optimization algorithm is unacceptable. To overcome this problem, we trained ANN using case indexing to speed up the whole process of optimization. Pal and Shiu [17] stated three principal advantages of using neural network to index cases compared with other traditional indexing methods, such as B+-tree, R-tree, and Bayesian Model. ANN first improves the case searching and retrieval efficiency. Secondly, it is very robust in handling noisy case data with incomplete or missing information which is the situation of strategy game. Finally, it is suitable for accumulating data in RTS games.

In our simulation, the PSO-obtained best solution of cannon placement in a given battlefield was used as a case input to ANN. It was supposed that every point of the best case had a circle range with a radius. The points in the circle were encoded as ”-1” or ”1”. Then an encoding string was formed, where ”-1” represents the barrier, ”1” represents the open area and the final digit of the code,'d' represents the distance between the central point and the defense base. For example (Figure 4), eight points around Point A in the circle will be encoded as ”-1” or ”1”. The 9-bit code string of Point A, [-1 1 -1 -1 -1 -1 1 -1 d], is used as the input of ANN. It is noted that Point A will not be counted into the code string.

The network used in this study consisted of three layers, i.e. input layer, hidden layer, and output layer. The code string was the input of ANN. The number of the hidden layer was calculated as the square root of the length of the code string in back-propagation (BP) model neural network structure. In the above example, the number of the hidden layer is $\sqrt{9}$=3. The output of ANN was the probability of the central point to set up cannon. The higher value means the better location for setting up cannon.

The equation of the best location function (Eq. 5) is defined based on the relationship among the current landscape, the cannon locations obtained by PSO, and the distance to the defense base.

$$f(A) = \sum_{k=1}^{N} e^{-\frac{(A_1 - K_{k1})^2 + (A_2 - K_{k2})^2}{r}} \tag{5}$$

where *f(A)*is the best location function of Point A. (A_1, A_2) is the coordinate of Point A. *N* is the total number of cannons. (K_{k1}, K_{k2}) is the coordinate of cannon *k*, and *r* is a power decay coefficient for controlling the spread of *f(A)* shown in Figure 3.

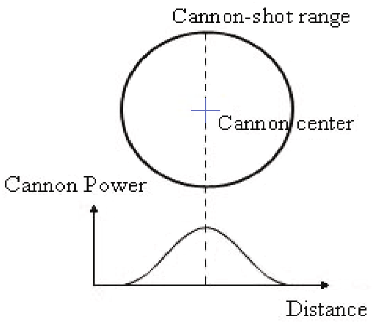


Fig. 3. Power decay coefficient

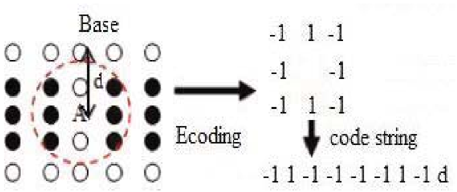


Fig. 4. Illustration of encoding of Point A for input to ANN

The code string and the best location value of each point in the PSO-obtained case were employed as input and output, respectively, to train the network. The ANN training was using BP and log-sigmoid output function. Then, the weight values of the hidden layer were settled. In a new battlefield map, the trained ANN quickly output the probability of each point to set up cannon when the surrounding terrain of the point was similar to the given battlefield in which the best case was obtained using PSO. Finally, the optimum cannon placement in the new battlefield map was obtained using ANN.

3 Resutls

Our simulations were performed using a computer with 2.1GHz CPU and 2 GB RAM under Windows XP. MathWorks Matlab 7.0 was used as the simulation tool.

3.1 Fitness in Different Generations

In our simulation, the evaluation function was based on the damage created on the enemy. Therefore, the fitness function is defined to be proportional to this damage value. Figure 5 shows the damage values as a function of generations measured in different maps. During the simulations, it was found that the damage value converged at approximately 50 generations. Figure 6 shows the final cannon placement in different maps after 100 generations. It is clearly seen that most of the cannons are placed in canyons, in which the cannons can make high damage to the enemy.

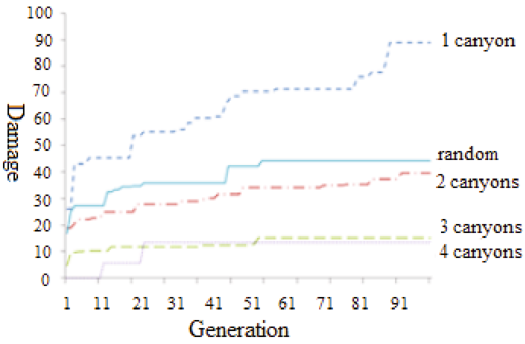


Fig. 5. Damage of the PSO-obtained cannons to the enemy as a function of generation in different maps

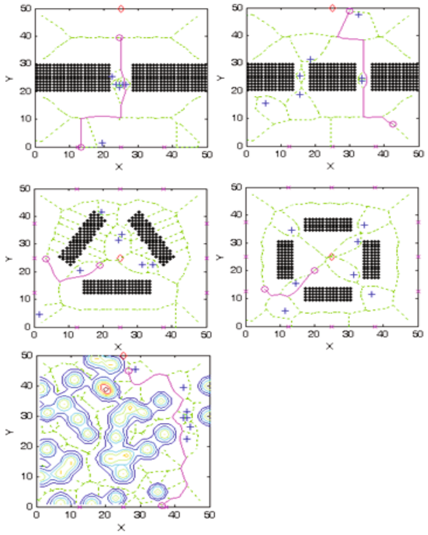


Fig. 6. PSO-obtained cannon placement in five battlefield maps

3.2 Execution Time of PSO and Training and Execution Time of ANN

Table 1 lists the execution time of PSO, the training time of ANN and the execution time of ANN for different maps. The maximum execution time of PSO is up to 360 seconds, which is obviously unacceptable for RTS games. With the application of case indexing for ANN, neural network could learn from the past experience (the PSO optimization result as case indexing) to speed up the searching process. After ANN training (approximately 30 seconds), our machine learning component became very useful. ANN could place the cannons in a new battlefield within less than 0.05 seconds that meets the requirement of RTS games. Figure 7A shows the points with the higher probability

Table 1. Execution time of PSO, training time of ANN, and execution time of ANN (Cannon number: 7, Population size: 50, Generation number: 100; ANN model: BP)

Map	Execution time of PSO (s)	Training time of ANN (s)	Execution time of ANN(s)
1 canyon	126.3	26.0	0.03
2 canyons	132.6	26.2	0.03
3 canyons	322.1	30.3	0.04
4 canyons	309.0	28.7	0.04
random	360.0	18.3	0.02

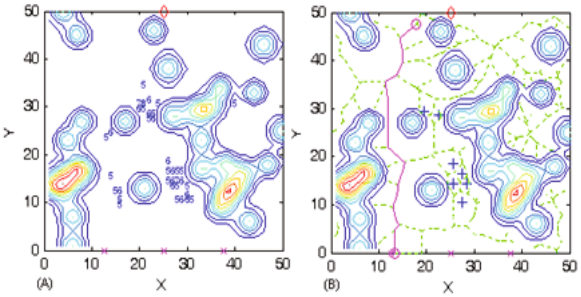


Fig. 7. (A) The points with higher probability (≥0.5) predicted by ANN. The value of probability displayed in the map equals to probability×10. (B) Final cannon placement obtained by ANN and the best path travelled by enemy with minimum damage.

Table 2. Training time and recalling time of BP and RB ANN models (Cannon number: 7, Population size: 50, Generation number: 100)

Neural network system	Training time of ANN (s)	Recalling time of ANN (s)
BP	21.5	0.02
RBF(NEWGRNN)	17.8	0.02
RBF(NEWRB)	10.9	0.04
RBF(NEWRBE)	12.4	10.2

of setting up cannons predicted by ANN. Figure 7B shows the final cannon placement and the best path that the enemy travels with minimum damage.

3.3 Training and Recalling Time of BP and RB networkN

BP and Radial Basis (RB) network models have been commonly used in neural network systems. In our MATLAB simulations, three RB functions (RBF), NEWGRNN, NEWRB and NEWRBE, were used to design RB network. Figure 8 shows similar

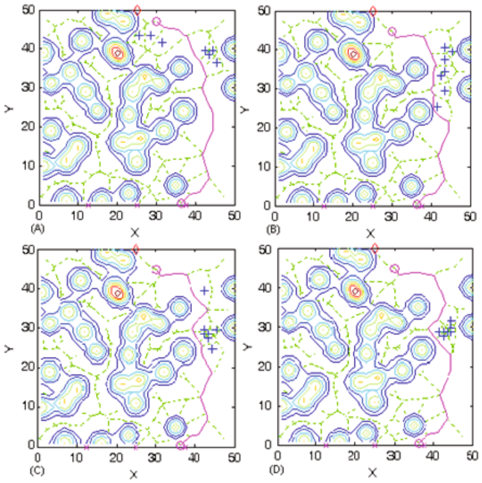


Fig. 8. Cannon placement obtained by ANN. (A) BP network; (B-D) RB networks using NEWGRNN (B), NEWRB (C) and NEWRBE (D) function.

results of cannon placement using BP and RBF. The main difference is in the training time and recalling time. In Table 2, RB network models show an improvement in the training time of ANN with a range from 17% to 49%. Previous study of Wang and He [18] reported a 10% improvement in Text Classification. It is shown that RB networks have a better improvement in comparison with BP. However, the recalling time of RB networks is averagely much more than that of BP. The difference in the recalling time may be a heavy workload as RTS games need to complete each game cycle in every 0.02 to 0.03 seconds. Therefore, we suggest using BP for RTS games due to its faster recalling time.

4 Conclusion

In this study, PSO is employed to optimize the placement of cannons to defend the enemy attack. The execution time of PSO ($>$ 100 seconds) is unsatisfied for RTS game. We proposed a hybrid PSO-ANN method using the result of PSO as a case indexing of past experience to train ANN. After the training (approximately 30 seconds), ANN can quickly obtain the cannon placement within 0.05 seconds. Experimental results demonstrated that this case indexing method using PSO and ANN efficiently speeded up the whole process to satisfy the requirement in RTS game. Compared with RB network models, BP model with a faster recalling time and similar performance is suggested in RTS games.

Acknowledgements

This work is supported by the Hong Kong Polytechnic University A-PA6N and RGLS PhD research grants.

References

1. Hsieh, J.L., Sun, C.T.: Building a player strategy model by analyzing replays of real-time strategy games. In: Neural Networks, IJCNN (2008)
2. David, M., Bourg, G.S.: AI for Game Developers. O'Reilly Media, Inc., Sebastopol (2004)
3. Cole, N., Louis, S.J.: Using a genetic algorithm to tune first-person shooter bots. In: Proceedings of IEEE Congress on Evolutionary Computation, Portland, OR, pp. 139–145 (2004)
4. Salge, C., Lipski, C.: Using genetically optimized artificial intelligence to improve gameplaying fun for strategical games. In: Proceedings of the 2008 ACM SIGGRAPH symposium on Video Games. ACM Press, New York (2008)
5. Louis, S.J., Miles, C.: Playing to learn: case-injected genetic algorithms for learning to play computer games. IEEE Transactions on Evolutionary Computation 9(4), 669–681 (2005)
6. Elbeltagi, E., Hegazy, T., Grierson, D.: Comparison among five evolutionary-based optimization algorithms. Advanced Engineering Informatics 19(1), 43–53 (2005)
7. Edwards, A., Engelbrecht, A.P.: Comparing Particle Swarm Optimisation and Genetic Algorithms for Nonlinear Mapping. In: Proceedings of IEEE Congress on Evolutionary Computation, Vancouver, BC, Canada, pp. 694–701 (2006)
8. Ou, C., Lin, W.X.: Comparison between PSO and GA for Parameters Optimization of PID Controller. In: Proceedings of the 2006 IEEE International Conference on Mechatronics and Automation, Luoyang, China, pp. 2471–2475 (2006)
9. Abdelbar, A.M., Ragab, S., Mitri, S.: Applying Co-Evolutionary Particle Swam Optimization to the Egyptian Board Game Seega. In: Proceedings of the First Asian-Pacific Workshop on Genetic Programming, Canberra, Australia, pp. 9–15 (2003)
10. Conradie, J., Engelbrecht, A.P.: Training Bao Game-Playing Agents using Coevolutionary Particle Swarm Optimization. In: IEEE Symposium on Computational Intelligence and Games, Reno/Lake Tahoe, USA, pp. 67–74 (2006)
11. Franken, N., Engelbrecht, A.: Comparing PSO structures to learn the game of checkers from zero knowledge. In: Proceedings of IEEE Congress on Evolutionary Computation, Canberra, Australia, pp. 234–241 (2003)
12. Huo, P., Shiu, S.C.K., Wang, H.B., Niu, B.: Application and Comparison of Particle Swarm Optimization and Genetic Algorithm in Strategy Defense Game. In: The 5th International Conference on Natural Computation, Tianjin, China, August 2009, pp. 14–16 (2009)
13. Forbus, K.D., Mahoney, J.V., Dill, K.: How qualitative spatial reasoning can improve strategy game AIs. IEEE Intelligent Systems 17(4), 25–30 (2002)
14. Shi, Y., Eberhart, R.: A modified particle swarm optimizer. In: Proceedings of IEEE Conference on Evolutionary Computation, Anchorage, AK, pp. 69–73 (1998)
15. Engelbrecht, A.P.: Fundamentals of Computational Swarm Intelligence. John Wiley-Sons, Ltd, Chichester (2005)
16. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Perth, Australia (1948)
17. Pal, S.K., Shiu, S.C.K.: Foundations of Soft Case-Based Reasoning. Wiley Interscience, Hoboken (2004)
18. Wang, Z., He, Y.: A Comparison among Three Neural Networks for Text Classification. In: Proceedings of 8th International Conference on Signal Processing (2006)

Construction of Fuzzy Relation by Closure Systems

Vladimír Janiš[1], Magdalena Renčova[1,⋆], Branimir Šešelja[2],
and Andreja Tepavčević[2,⋆⋆]

[1] Department of Mathematics, Faculty of Natural Sciences, Matej Bel University,
SK-974 01 Banská Bystrica, Slovak Republic
[2] Department of Mathematics and Informatics, University of Novi Sad, Serbia

Abstract. Starting with a collection of closure systems each of which is associated to an element of a given set X, we construct a lattice L and an L-fuzzy relation on X, such that its fuzzy blocks are precisely the given closure systems.

AMS Mathematics Subject Classification (2000): primary 03B52, 03E72; secondary 06A15.

Keywords and phrases: lattice-valued fuzzy set, lattice-valued fuzzy relation, block, cut.

1 Introduction

As it is known in the classical theory of relations, a block of a binary relation on a domain X is its subset associated to an element x of X: it contains all elements from the domain related to this particular element. The most known are blocks of an equivalence relation which split the domain into a quotient set; for an ordering relation, a block of an element is a principal filter generated by this element etc. In many applications of binary relations blocks are actually used; that is why they are very important. In addition, it is easy to (re)construct the relation if all blocks are known.

In the case of fuzzy (lattice valued) relations, the situation is similar, though much more complicated. Blocks of a fuzzy relation are (defined analogously as in the crisp case) fuzzy sets on the same domain. For particular fuzzy relations, blocks have been widely investigated by many authors, though under different names. Gottwald (see his book [5]) introduces a kind of blocks using the name *full image* of the fuzzy set under a fuzzy relation (these correspond to Zadeh's conditional fuzzy sets). In the book [6] by Klir and Yuan, one can find blocks of a fuzzy order,

[⋆] The research supported by grant no. 1/0539/08 Slovak Grant Agency VEGA and APVV no. SK-SRB-0008-07.

[⋆⋆] The research supported by Serbian Ministry of Science and Technological Development, Grant No. 144011. and by the Provincial Secretariat for Science and Technological Development, Autonomous Province of Vojvodina, Grant "Lattice methods and applications".

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 116–121, 2009.

called dominating classes. The best known are fuzzy partitions, arising from fuzzy equivalences (investigated by e.g., De Baets, Mesiar, Montes, Couso, Gil, [1,7,8], and by many others). For more papers about relational blocks, see references in the cited articles and books. Let us mention also the book [2] by Belohlavek, which provides a comprehensive introduction to fuzzy (lattice valued) relations.

In the present paper we deal with lattice-valued binary fuzzy relations. Due to our cutworthy approach, the co-domain of all fuzzy structures here is a complete lattice. Our aim is to investigate fuzzy blocks of an arbitrary fuzzy relation. Being fuzzy sets on the same domain, these blocks uniquely determine a family of subsets of the domain - its cuts. This family is a closure system under set inclusion.

In dealing with the opposite construction, we are motivated by the following real life problem. Suppose that each member of some group of persons is, by some preference, associated to several groups of other members of the same group. It is obviously a closure system on this group of people. Hence, there are as much closure systems on the group, as there are people in it. Now the problem is to determine a lattice, and a fuzzy (lattice valued) relation which connects all these closure systems. In addition, the cuts of the blocks of this relation should be the starting closure systems.

This problem is solved by the construction presented in the present article.

2 Preliminaries

If ρ is a binary relation on a set X, $\rho \subseteq X^2$, then for $x \in X$ we denote by $\rho[x]$ the subset of X defined by

$$\rho[x] := \{y \in X \mid (x,y) \in \rho\}. \tag{1}$$

The set $\rho[x]$ is called the x**-block** of ρ.

The following lemma is straightforward.

Lemma 1. *If* $\rho = \bigcap_{i \in I} \rho_i$, *then* $\rho[x] = \bigcap_{i \in I} \rho_i[x]$.

A **complete lattice** is a partially ordered set $(L, \leq)$ whose every subset has a least upper bound (join, supremum) and a greatest lower bound (meet, infimum) under $\leq$. A complete lattice has the top and the bottom element, denoted respectively by 1 and 0.

Lattices and related topics are presented e.g., in book [3].

A **fuzzy set** $\mu : X \to L$ is a mapping from a non-empty set X (domain) into a complete lattice L (co-domain). According to the original definition L is the unit interval $[0,1]$ of real numbers (which is a complete lattice under $\leqslant$). However, we consider a complete lattice L in a more general setting and sometimes we use the term L**-fuzzy set**, or **lattice-valued** (**fuzzy**) **set** [4].

A mapping $R : X^2 \to L$ (a fuzzy set on X^2) is a **fuzzy** (L**-fuzzy, lattice-valued**) **relation** on X.

If $\mu : X \to L$ is a fuzzy set on a set X then for $p \in L$, the set

$$\mu_p := \{x \in X \mid \mu(x) \geqslant p\}$$

is a p**-cut**, or a **cut set**, (**cut**) of μ. More about cutworthy approach to lattice-valued fuzzy structures can be found in [9,10,11].

3 Results

Let $R : X^2 \to L$ be an L-valued binary relation on a set X. For every $x \in X$, the **fuzzy x-block** (x-block) of R is the L-valued set $R[x] : X \to L$, defined by

$$R[x](y) := R(x, y), \quad \text{for each } y \in X. \tag{2}$$

Observe that we use the same name for the crisp x-block of a crisp relation (given by (1) above); still it is clear from the context which of these notions is used. We refer to *blocks* of R, meaning x-blocks, where x runs over all $x \in X$. Being an L-valued set on X, every x-block of R determines a collection of crisp subsets of X, its cuts.

Our main result is the following.

Theorem 1. *Let X be a nonempty set and for each $x \in X$ let $\mathcal{R}_x$ be a collection of nonempty subsets of X closed under set intersection and containing X. Then there is a lattice L and an L-valued relation R on X, such that for every $x \in X$, $\mathcal{R}_x$ is the collection of nonempty cuts of a relational block $R[x]$.*

First we describe a construction of the lattice L and of the corresponding L-valued relation R, and then we prove the theorem by showing that the constructed relation fulfills the requirements.

Let X be a nonempty set. By assumption of the theorem, for every $x \in X$,

$$\mathcal{R}_x = \{U_i^x \mid i \in I\}$$

is a family of subsets of X which is closed under set intersection and which contains X as a member. Observe that the cardinality of each family is the same, $\operatorname{card} I$; in addition, some sets in the same family may be equal.

The construction

A Starting with $\mathcal{R}_x$, for each $x \in X$ define the following collection of subsets of X^2:

$$\overline{\mathcal{R}}_x := \{\overline{U_i^x} \subseteq X^2 \mid i \in I\} \text{ where for every } i \in I,\ \overline{U_i^x} := \{(x, u) \mid u \in U_i^x\}.$$

In other words, $\overline{\mathcal{R}}_x$ is obtained in such a way that each element u of every set in $\mathcal{R}_x$ is replaced by the ordered pair (x, u).

B Construct a collection of ordinary (crisp) relations $\{\rho_i \mid i \in I\}$, so that for every $i \in I$

$$\rho_i := \bigcup_{x \in X} \{\overline{U_i^x} \mid \overline{U_i^x} \in \overline{\mathcal{R}}_x\}.$$

C Construction of the lattice L:

$$L := (\mathcal{C}, \leqslant),$$

where $\mathcal{C} := \widehat{\{\rho_i \mid i \in I\}}$, and the "hat" means that $\{\rho_i \mid i \in I\}$ is completed by adding to this collection all missing set intersections of its relations; the order $\leqslant$ is the dual of set inclusion.[1]

[1] In lattice-theoretic terms, Dedekind-McNeille completion of the poset $(\{\rho_i \mid i \in I\}, \subseteq)$ is constructed, and its dual is taken to be the lattice L.

D Define the relation $R : X^2 \to L$ by

$$R(x,y) := \bigcap\{\rho \in \mathcal{C} \mid (x,y) \in \rho\}.$$

In the following lemma, we assume that all objects defined above, from X to $\mathcal{C}$ and R are given.

Lemma 2. *Let $\rho \in \mathcal{C}$ and $x \in X$. Then the following holds:*

(i) The x-block of ρ, if it is not empty, belongs to the family $\mathcal{R}_x$, namely: $\rho[x] \in \mathcal{R}_x$.

(ii) $\{u \in X \mid \bigcap\{\sigma \in \mathcal{C} \mid (x,u) \in \sigma\} \subseteq \rho\} = \rho[x]$.

Proof. (i) Observe that either $\rho = \rho_i$ for some $i \in I$, or $\rho = \bigcap\{\rho_j \mid j \in J \subseteq I\}$.

In the first case,

$$\rho = \rho_i := \bigcup_{x \in X}\{\overline{U_i^x} \mid \overline{U_i^x} \in \overline{\mathcal{R}}_x\}.$$

$$\rho_i[x] = \{u \in X \mid (x,u) \in \rho_i\}.$$

Further, $\rho_i[x] = \{u \in X \mid (x,u) \in \overline{U_i^x}\}$.

$(x,u) \in \overline{U_i^x}$ is equivalent with $u \in U_i^x$.

Therefore,

$$\rho_i[x] = U_i^x \tag{3}$$

and $\rho[x] = \rho_i[x] \in \mathcal{R}_x$.

Now, suppose that $\rho = \bigcap\{\rho_j \mid j \in J \subseteq I\}$. By Lemma 1, $\rho[x] = \bigcap\{\rho_j[x] \mid j \in J \subseteq I\}$. Now we have, $\rho[x] = \bigcap\{U_j^x \mid j \in J \subseteq I\}$. Since, for every x, $\mathcal{R}_x$ is a family closed under intersection, then $\rho[x] \in \mathcal{R}_x$.

(ii) If $y \in \{u \in X \mid \bigcap\{\sigma \in \mathcal{C} \mid (x,u) \in \sigma\} \subseteq \rho$, then obviously $\bigcap\{\sigma \in \mathcal{C} \mid (x,y) \in \sigma\} \subseteq \rho$, therefore $(x,y) \in \rho$, hence $y \in \rho[x]$. Conversely, if $y \in \rho[x]$, then $(x,y) \in \rho$, therefore ρ belongs to the collection of relations that contain (x,y), and thus $\bigcap\{\sigma \in \mathcal{C} \mid (x,y) \in \sigma\} \subseteq \rho$. Hence, $y \in \{u \in X \mid \bigcap\{\sigma \in \mathcal{C} \mid (x,u) \in \sigma\} \subseteq \rho\}$. □

Proof of Theorem 1. As defined by (2), $\{R[x] \mid x \in X\}$ is a collection of L-valued sets on X which are the blocks of R; recall that for $x \in X$ and for every $y \in X$, we have $R[x](y) = R(x,y)$.

To prove the theorem, we show that for each $x \in X$, the collection of nonempty cuts of $R[x]$ coincides with the family $\mathcal{R}_x$ by which we started.

Let $x \in X$, and let $R[x]_\rho$ be the cut of the block $R[x]$, for some $\rho \in \mathcal{C}$. Now,

$R[x]_\rho = \{u \in X \mid R(x,u) \geqslant \rho\} = \{u \in X \mid R(x,u) \subseteq \rho\} =$

$\{u \in X \mid \bigcap\{\sigma \in \mathcal{C} \mid (x,u) \in \sigma\} \subseteq \rho\} = \rho[x] \in \mathcal{R}_x$,

by Lemma 2 (ii) and (i), respectively.

Conversely, let $U_i^x \in \mathcal{R}_x$. By the fact (3) proved within Lemma 2 (i), $\rho_i[x] = U_i^x$. By Lemma 2 (ii),

$\rho_i[x] = \{u \in X \mid \bigcap\{\sigma \in \mathcal{C} \mid (x,u) \in \sigma\} \subseteq \rho_i\} =$

$\{u \in X \mid R(x,u) \geq \rho_i\} = R[x]_{\rho_i}$.

Hence $U_i^x = R[x]_{\rho_i}$. □

Example. Let $X = \{a, b, c\}$.

We start with three collections of subsets of X which are closed under intersections:

$\{\{a\}, \{b\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$, $\{\{a\}, \{a, b\}, \{a, b, c\}\}$, $\{\{c\}, \{a, b, c\}\}$.

In order to have families of the same cardinality (five members), we consider the sets in above collections to be ordered as they are listed, and we repeat the set $\{a, b, c\}$ necessary number of times (two and three times in the second and the third collection, respectively).

$\mathcal{R}_a = (\{a\}, \{b\}, \{a, b\}, \{b, c\}, \{a, b, c\})$,
$\mathcal{R}_b = (\{a\}, \{a, b\}, \{a, b, c\}, \{a, b, c\}, \{a, b, c\})$,
$\mathcal{R}_c = (\{c\}, \{a, b, c\}, \{a, b, c\}, \{a, b, c\}, \{a, b, c\})$.

Relations $\overline{\mathcal{R}}_a$, $\overline{\mathcal{R}}_b$ and $\overline{\mathcal{R}}_c$, step **A**:

$\overline{\mathcal{R}}_a = (\{(a, a)\}, \{(a, b)\}, \{(a, a), (a, b)\}, \{(a, b), (a, c)\}, \{(a, a), (a, b), (a, c)\})$
$\overline{\mathcal{R}}_b = (\{(b, a)\}, \{(b, a), (b, b)\}, \{(b, a), (b, b), (b, c)\}, \{(b, a), (b, b), (b, c)\},$
$\{(b, a), (b, b), (b, c)\})$
$\overline{\mathcal{R}}_c = (\{(c, c)\}, \{(c, a), (c, b), (c, c)\}, \{(c, a), (c, b), (c, c)\}, \{(c, a), (c, b), (c, c)\},$
$\{(c, a), (c, b), (c, c)\})$.

Next we construct relations $\rho_1, \ldots, \rho_5$, as defined in step **B**: $\rho_1 = \{(a, a)\} \cup \{(b, a)\} \cup \{(c, c)\}$; $\rho_2 = \{(a, b)\} \cup \{(b, a), (b, b)\} \cup \{(c, a), (c, b), (c, c)\}$; etc. :

$$\begin{array}{c|ccc} \rho_1 & a & b & c \\ \hline a & 1 & 0 & 0 \\ b & 1 & 0 & 0 \\ c & 0 & 0 & 1 \end{array} \quad \begin{array}{c|ccc} \rho_2 & a & b & c \\ \hline a & 0 & 1 & 0 \\ b & 1 & 1 & 0 \\ c & 1 & 1 & 1 \end{array} \quad \begin{array}{c|ccc} \rho_3 & a & b & c \\ \hline a & 1 & 1 & 0 \\ b & 1 & 1 & 1 \\ c & 1 & 1 & 1 \end{array} \quad \begin{array}{c|ccc} \rho_4 & a & b & c \\ \hline a & 0 & 1 & 1 \\ b & 1 & 1 & 1 \\ c & 1 & 1 & 1 \end{array} \quad \begin{array}{c|ccc} \rho_5 & a & b & c \\ \hline a & 1 & 1 & 1 \\ b & 1 & 1 & 1 \\ c & 1 & 1 & 1 \end{array}$$

Now the missing set intersections of the above relations:

$\rho_6 = \rho_1 \cap \rho_2$; $\rho_7 = \rho_3 \cap \rho_4$

$$\begin{array}{c|ccc} \rho_6 & a & b & c \\ \hline a & 0 & 0 & 0 \\ b & 1 & 0 & 0 \\ c & 0 & 0 & 1 \end{array} \quad \begin{array}{c|ccc} \rho_7 & a & b & c \\ \hline a & 0 & 1 & 0 \\ b & 1 & 1 & 1 \\ c & 1 & 1 & 1 \end{array}$$

$\mathcal{C} = \{\rho_1, \ldots, \rho_7\}$, the lattice L (step **C**) is in Figure 1.

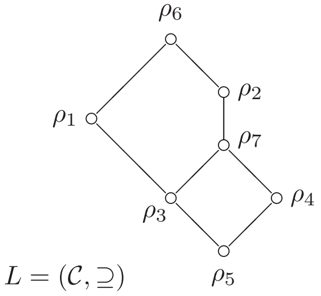


Fig. 1.

Construction of fuzzy relation $R : X^2 \to L$ (step **D**):
$R(a,a) = \bigcap\{\rho \in \mathcal{C} \mid (a,a) \in \rho\} = \rho_1 \cap \rho_3 \cap \rho_5 = \rho_1$;
$R(b,c) = \bigcap\{\rho \in \mathcal{C} \mid (b,c) \in \rho\} = \rho_3 \cap \rho_4 \cap \rho_5 \cap \rho_7 = \rho_7$; etc. :

R	a	b	c
a	ρ_1	ρ_2	ρ_4
b	ρ_6	ρ_2	ρ_7
c	ρ_2	ρ_2	ρ_6

.

The blocks of R are defined by the rows of the above table:

$$R[a] = \begin{pmatrix} a & b & c \\ \rho_1 & \rho_2 & \rho_4 \end{pmatrix},$$

and the corresponding cut sets are: $R[a]_{\rho_1} = \{a\}$, $R[a]_{\rho_2} = R[a]_{\rho_7} = \{b\}$, $R[a]_{\rho_3} = \{a,b\}$, $R[a]_{\rho_4} = \{b,c\}$, $R[a]_{\rho_5} = \{a,b,c\}$, $R[a]_{\rho_6} = \emptyset$. Hence, the nonempty cuts of $R[a]$ are the sets in $\mathcal{R}_a$.

Analogously, one can check that the collection of nonempty sets in $R[b]$ coincides with the family $\mathcal{R}_b$; the same holds for the cuts in $R[c]$ and the sets in $\mathcal{R}_c$. □

4 Conclusion

We have solved the problem of the construction of a relation, if collections of cuts of it blocks are given. As mentioned, this construction could be widely applied. Our next task is to investigate properties which should be fulfilled by the given closure systems, in order that the obtained fuzzy relation possesses some particular properties, like reflexivity, symmetry, antisymmetry, transitivity etc.

References

1. De Baets, B., Mesiar, R.: T-partitions. Fuzzy Sets and Systems 97, 211–223 (1998)
2. Bělohlávek, R. (ed.): Fuzzy Relational Systems: Foundations and Principles. Kluwer Academic/Plenum Publishers, New York (2002)
3. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order. Cambridge University Press, Cambridge (1992)
4. Goguen, J.A.: L-fuzzy Sets. J. Math. Anal. Appl. 18, 145–174 (1967)
5. Gottwald, S.: Fuzzy sets and Fuzzy Logic. Vieweg (1993)
6. Klir, G., Yuan, B.: Fuzzy sets and fuzzy logic. Prentice-Hall, Englewood Cliffs (1995)
7. Montes, S., Couso, I., Gil, P.: Fuzzy $\delta - \varepsilon$*-partition*. Information Sciences 152, 267–285 (2003)
8. Murali, V.: Fuzzy equivalence relations. Fuzzy Sets and Systems 30, 155–163 (1989)
9. Šeselja, B., Tepavčević, A.: Representing Ordered Structures by Fuzzy Sets, An Overview. Fuzzy Sets and Systems 136, 21–39 (2003)
10. Šešelja, B., Tepavčević, A.: Completion of ordered structures by cuts of fuzzy sets: An overview. Fuzzy Sets and Systems 136, 1–19 (2003)
11. Tepavčević, A., Vujić, A.: On an application of fuzzy relations in biogeography. Information Sciences 89(1-2), 77–94 (1996)

Incorporating Fuzziness to CLARANS

Sampreeti Ghosh and Sushmita Mitra

Center for Soft Computing, Indian Statistical Institute, Kolkata 700108, India
{sampreeti_t,sushmita}@isical.ac.in

Abstract. In this paper we propose a way of handling fuzziness while mining large data. Clustering Large Applications based on RANdomized Search (CLARANS) is enhanced to incorporate the fuzzy component. A new scalable approximation to the maximum number of neighbours, explored at a node, is developed. The goodness of the generated clusters is evaluated in terms of validity indices. Experimental results on various data sets is run to converge to the optimal number of partitions.

Keywords: Data mining, CLARANS, medoid, fuzzy sets, clustering.

1 Introduction

Clustering LARge Applications (CLARA) [1] incorporates sampling in the framework of PAM [1] to make it scalable for handling large data. Again, the results here remain dependent on the quality of the sampling process undertaken. An efficient variation of CLARA is Clustering Large Applications based on RANdomized Search (CLARANS) [2], which partitions large data. It is found to outperform both PAM and CLARA [3] in terms of accuracy and computational complexity, and can handle outliers. However, since all the three clustering algorithms are designed to generate crisp clusters, they fare poorly when modeling overlapping clusters.

Fuzzy sets [4] constitute the oldest and most reported soft computing paradigm [5]. These provide soft decision by taking into account characteristics like tractability, robustness, low cost, etc., and have close resemblance to human decision making. They are well-suited to modeling different forms of uncertainties and ambiguities, often encountered in real life. The concept of fuzzy membership μ, lying in $[0, 1]$, allows the simultaneous finite belongingness of a pattern to two or more overlapping clusters.

In this article we propose a novel fuzzy clustering algorithm, Fuzzy CLARANS (FCLARANS), that performs efficiently on large data. It incorporates the concept of fuzzy membership onto the framework of CLARANS for manoeuvering uncertainty in the context of data mining. Cluster validity indices, like Davies-Bouldin (DB) [6] and Xie-Beni (XB) [7], are used to evaluate the goodness of the generated partitions. Note that, unlike DB, the index XB incorporates fuzzy membership in its computations. The clustering algorithms compared here are hard c -means (HCM), fuzzy c - means (FCM), fuzzy c - medoid (FCMd), and CLARANS.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 122–127, 2009.

The performance on various data sets demonstrates that the proposed Fuzzy CLARANS always converges to the lowest value for both the indices DB and XB at an optimal number of clusters. The cost function of CLARANS is fuzzified using membership values. A new scalable approximation is developed to compute the maximum number of neighbours being explored at each node. It also helps to eliminate user-defined parameters in the expression.

The rest of the paper is organized as follows. Section 2 describes the preliminaries, like algorithms CLARANS and the clustering validity indices DB and XB. Then we move to the proposed Fuzzy CLARANS in Section 3. The experimental results are presented in Section 4. Finally, Section 5 concludes the article.

2 Preliminaries

In this section we describe some of the basic concepts like algorithms CLARANS [2],and clustering validity indices.

2.1 CLARANS

Large datasets require the application of scalable algorithms. CLARANS [2] draws a sample of the large data, with some randomness, at each stage of the search. Each cluster is represented by its medoid. Multiple scans of the database are required by the algorithm. Here the clustering process searches through a graph G, where node v^q is represented by a set of c medoids (or centroids) $\{m_1^q, \ldots, m_c^q\}$. Two nodes are termed as neighbors if they differ by only one medoid, and are connected by an edge. More formally, two nodes $v^1 = \{m_1^1, \ldots, m_c^1\}$ and $v^2 = \{m_1^2, \ldots, m_c^2\}$ are termed neighbors if and only if the cardinality of the intersection of v^1 and v^2 is given as $card(v^1 \bigcap v^2) = c - 1$. Hence each node in the graph has $c * (N - c)$ neighbors. For each node v^q we assign a cost function

$$J_c^q = \sum_{x_j \varepsilon U_i} \sum_{i=1}^{c} d_{ji}^q, \tag{1}$$

where d_{ji}^q denotes the dissimilarity measure of the jth object x_j from the ith cluster medoid m_i^q in the qth node. The aim is to determine that set of c-medoids $\{m_1^0, \ldots, m_c^0\}$ at node v^0, for which the corresponding cost is the minimum as compared to all other nodes in the tree.

Note that *the maximum number of neighbors* is computed as

$$neigh = p\% \;\text{ of }\; \{c * (N - c)\}, \tag{2}$$

with p being provided as input by the user. Typically, $1.25 \leq p \leq 1.5$ [2].

2.2 Validity Indices

There exist validity indices [8] to evaluate the goodness of clustering, corresponding to a given value of c. Two of the commonly used measures include the Davies-Bouldin (DB) [6] and the Xie-Beni (XB) [7] indices.

The DB index is a function of the ratio of sum of within-cluster distance to between-cluster separation. It is expressed as

$$DB = \frac{1}{c}\sum_{i=1}^{c} max_{k \neq i} \frac{diam(U_i) + diam(U_j)}{d'(U_i, U_j)}, \tag{3}$$

where the diameter of cluster U_i is $diam(U_i) = \frac{1}{|U|_i}\sum_{x_j \in U_i} ||x_j - m_i||^2$. The inter-cluster distance between cluster pair U_i, U_j is expressed as $d'(U_i, U_j) = ||m_i - m_j||^2$. DB is minimized when searching for the optimal number of clusters c_0.

The XB index is defined as

$$XB = \frac{\sum_{j=1}^{N}\sum_{i=1}^{c} \mu_{ij}^{m'} d_{ji}}{N * \min_{i,j} d'(U_i, U_j)^2}, \tag{4}$$

where μ_{ij} is the membership of pattern x_j to cluster U_i. Minimization of XB is indicative of better clustering, particularly in case of fuzzy data. Note that for crisp clustering the membership component μ_{ij} boils down to zero or one.

3 Fuzzy CLARANS

In this section we describe the proposed algorithm Fuzzy CLARANS (FCLARANS). Here fuzzy membership is incorporated in the framework of CLARANS. This enables appropriate modeling of ambiguity among overlapping clusters. A pattern is allowed finite, non-zero membership $\mu_{ij} \in [0, 1]$ to two or more partitions. The distance component is weighted by the corresponding membership value, analogously to FCM and FCMd.

The hybridization allows the modeling of uncertainty in the domain of large data. Although the computational complexity is higher than that of CLARANS, yet the performance is found to be superior for the optimal partitioning, as evaluated in terms of clustering validity indices. It is interesting to observe that fuzzy clustering in FCLARANS boils down to the crisp version in CLARANS, when $\mu_{ij} \in \{0, 1\}$.

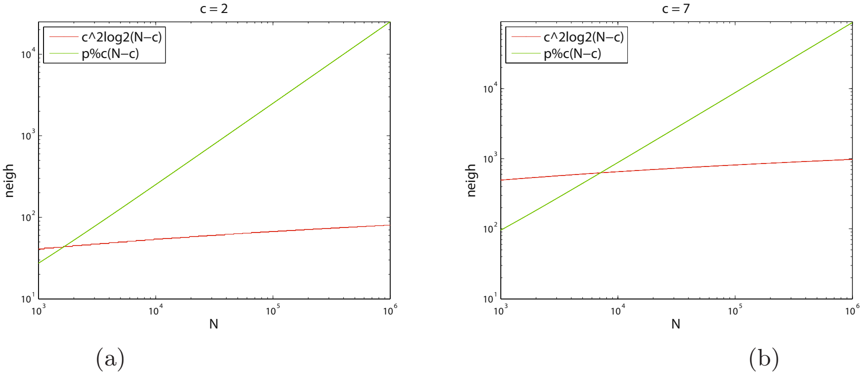
The cost of a node, as defined in eqn. (1), is now modified to

$$J_{fc}^{q} = \sum_{j=1}^{N}\sum_{i=1}^{c} (\mu_{ij}^{q})^{m'} d_{ji}^{q}. \tag{5}$$

We chose $m' = 2$ [9] after several experiments. The membership at node v^q is computed as $\mu_{ij}^{q} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ji}^{q}}{d_{jk}^{q}}\right)^{\frac{2}{m'-1}}}$.

We observed that the value of $neigh = p\%$ of $\{c*(N-c)\}$ [as in CLARANS, eqn. (2)] turns up to be very high for large data having $N \geq 10,000$. This increases the computational burden. We, therefore, propose a new scalable approximation expressed as

$$neigh = c^2 \log_2(N - c). \tag{6}$$



(a) (b)

Fig. 1. Variation of *neigh* (by the two expressions) with N, intersecting at N_{cross}, for cluster numbers (a) $c = 2$ and (b) $c = 7$

A study of its behaviour is provided in Fig. 1, with respect to the previous expression [eqn. (2)]. We are, as a result, able to eliminate the user-defined parameter p while also reducing computational time of the algorithm. The new eqn. (6) has been employed in this article for experiments involving large datasets ($N \geq 10,000$). Note Figs. 1(a)-(b),the expressions for *neigh* [by eqns. (2) and (6)] intersect each other.

4 Experimental Results

The performance of the algorithm Fuzzy CLARANS was tested on various data sets. The goodness of the partitions was evaluated in terms of cluster validity indices DB and XB. Comparative study was made with related clustering algorithms like HCM, FCM, FCMd, and CLARANS.

We used a synthetic dataset (*Set1*), and three real datasets *Magic gamma*, *Shuttle* and *Forest Cover*. Average results were reported over five runs.

4.1 Set 1

The data contains three clusters, each with 100 randomly generated patterns. The two-dimensional scatter plot of Fig. 2 depicts the patterns lying within circles of unit radius, each having different centers. A lot of overlapping is artificially introduced.

Table 1 establishes that DB and XB are minimum for Fuzzy CLARANS for the correct number of three partitions. Although FCM generates the globally least value for XB, yet the partitioning is incorrect at $c = 5$. On the other hand, FCMd also provides best result in terms of both the indices. However, algorithm FCLARANS is able to model the overlapping partitions in a better manner.

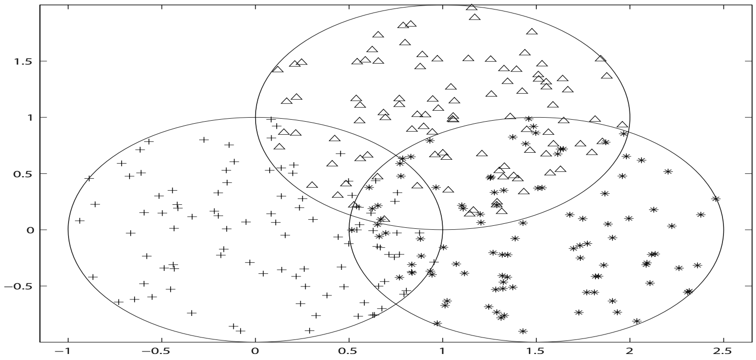


Fig. 2. Data *Set1*

Table 1. Average comparative performance on synthetic data

Algorithm	c=2, $neigh$=12	c=3, $neigh$=18	c=4, $neigh$=24	c=5, $neigh$=30
	DB, XB	DB, XB	DB, XB	DB, XB
HCM	0.70, 0.35	0.51, 0.28	0.49, 0.28	**0.45, 0.26**
FCM	0.76, 0.27	0.54, 0.20	0.56, 0.18	**0.48, 0.14**
FCMd	1.36, 0.40	**1.24, 0.31**	1.60, 0.32	1.64, 0.37
CLARANS	0.68, **0.34**	0.75, 0.42	0.83, 0.55	**0.55**, 0.42
FCLARANS	0.72, 0.26	**0.72, 0.23**	0.85, 0.28	0.76, 0.31

Table 2. Performance of FCLARANS on large datasets

Clusters	2	3	4	5	6	7	8
Dataset	DB,XB ($neigh$)	DB,XB ($neigh$)	DB,XB ($neigh$)	DB,XB ($neigh$)	DB,XB ($neigh$)	DB,XB ($neigh$)	DB,XB ($neigh$)
Magic gamma	**0.40, 1.44** (57)	0.70, 2.17 (128)	0.99, 3.14 (228)	1.10, 2.45 (356)	0.82, 1.69 (512)	0.81, 1.73 (697)	0.83, 2.04 (910)
Shuttle	1.61, 4.23 (64)	1.11, 2.71 (143)	1.84, 4.31 (254)	2.60, 4.97 (396)	2.51, 4.41 (570)	**1.10, 2.63** (776)	2.34, 4.11 (1013)
Forest cover	0.05, 1.73 (77)	0.07, 2.07 (173)	0.10, 2.34 (307)	0.15, 3.31 (479)	0.09, 1.60 (690)	0.08,**1.43** (939)	0.14, 2.54 (1226)

4.2 Large Data

The large data were taken from the *UCI Machine Learning Repository.*The *magic gamma* telescope data 2004 is made up of 19,020 instances, with ten features and two classes (gamma signal and hadron background). The *Shuttle* data (statlog version) consists of 58,000 measurements corresponding to seven classes, *viz.*, Rad flow, Fpv close, Fpv open, High, Bypass, Bpv close, Bpv open. There are nine numerical attributes. The *Forest cover* data consists of 5,81,012 instances. There are 10 numeric-valued attributes, with seven kinds of forest cover corresponding to spruce/fir, lodgepole pine, ponderosa pine, cottonwood/willow, aspen, douglas-fir and krummholz.

Table 2 provides the clustering results with FCLARANS on these three datasets. We observe that the minimum values for DB and XB always correspond to the correct number of clusters.

5 Conclusions

We have developed a new algorithm Fuzzy CLARANS, by incorporating fuzziness in CLARANS while clustering of large data. The cost function is weighted by fuzzy membership. A scalable approximation to the maximum number of neighbours, explored at a node, has been designed. It helps in reducing the computational time for large data, while eliminating the need for user-defined parameters. The goodness of the generated clusters has been evaluated in terms of the Davies Bouldin and Xie Beni validity indices. Results demonstrate the superiority of Fuzzy CLARANS in modeling overlaps, particularly in large data. The algorithm is found to converge to the optimal number of partitions.

Acknowledgements

The authors gratefully acknowledge Mr. Shivnath Shaw for his programming support and helpful discussions.

References

1. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York (1990)
2. Ng, R.T., Han, J.: Clarans: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering 14, 1003–1016 (2002)
3. Mitra, S., Acharya, T.: Data Mining: Multimedia, Soft Computing, and Bioinformatics. John Wiley, New York (2003)
4. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
5. Fuzzy logic, neural networks, and soft computing. Communications of the ACM 37, 77–84 (1994)
6. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1, 224–227 (1979)
7. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 841–847 (1991)
8. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, NJ (1988)
9. Yu, J.: General c-means clustering model. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 1197–1211 (2005)

Development of a Neuro-fuzzy MR Image Segmentation Approach Using Fuzzy C-Means and Recurrent Neural Network

Dipankar Ray[1] and D. Dutta Majumder[2,3]

[1] Department of CSE, Indian School of Mines, Dhanbad – 826 004, India
dpnkray@yahoo.co.in
[2] Institute of Cybernetics and Information Technology, Kolkata – 700 108, India
ddmdr@hotmail.com
[3] Indian Statistical Institute, Kolkata – 700 105, India

Abstract. A neuro-fuzzy clustering framework has been presented for a meaningful segmentation of Magnetic Resonance medical images. MR imaging provides detail soft tissue descriptions of the target body object and it has immense importance in today's non-invasive therapeutic planning and diagnosis methods. The unlabeled image data has been classified using fuzzy c-means approach and then the data has been used for training of an Elman neural network. The trained neural net is then used as a ready-made tool for MRI segmentation.

Keywords: Medical Image Segmentation, Neuro-Fuzzy, Elman Recurrent Neural Network (ERNN).

1 Introduction

Medical image segmentation provides the basis for any kind of high-level image analysis and subsequent therapeutic and diagnostic planning. Medical images in general contain considerable uncertainty, unknown noise, limited spatial resolution, poor contrast and non-uniform intensity distribution leading to difficulties with segmentation [1]. Among them noise removal is a challenging task and makes segmentation approach unstable. Segmentation on the basis of spatial information can smooth out acquisition noise as well as reduce classification ambiguities. Here we have proposed a semiautomatic segmentation strategy incorporating cognitive computational ability of two well-known soft computing approaches, namely, fuzzy and neural network (NN) models, that is purely data-driven and self organized approach. It is done primarily by identifying the characteristic properties of the unlabeled image feature space composed of gray levels of each voxel; then by an artificial neural network (ANN) with the labeled feature space. The information obtained from the unsupervised fuzzy clustering is used to train the subsequent supervised clustering approach. This is achieved by applying a fuzzy c-means approach as unsupervised clustering (UC) and a recurrent neural network model as supervised clustering (SC). Dynamically

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 128–133, 2009.

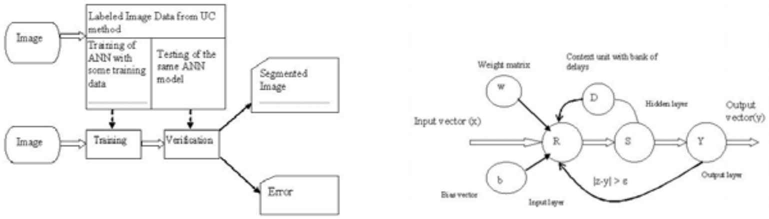


Fig. 1. (a) DFD and Process flow diagram of the segmentation approach and (b) State flow diagram of a recurrent neural network

driven recurrent neural structures can capture both the temporal as well as spatial information distribution patterns making it a suitable device to represent state-space configuration for dynamic nonlinear systems and adaptive prediction of function parameters. The Levenberg-Marquardt (LM) algorithm is used as the backpropagation (BP) optimization strategy of the used recurrent network. The method is basically a nonlinear least square minimization algorithm with a blend of gradient descent and Gauss-Newton iterative methods. The outcomes of the segmentation were visually verified and compared with predefined pathologically tested segmented data by practising radiologists.

2 Fuzzy Segmentation of MR Images

In a fuzzy segmentation, each pixel is assigned a membership value for each of the c regions. This forms a matrix U that partitions the image. Considering the membership values of each pixel we often obtain more accurate estimates of region properties. But this approach assigns pixels to more than one region and regions do not preserve the connectivity property of segmentation which "tear up tissue classes" into medically incorrect regions. This problem can also be solved either by hardening each column of U with suitable hardening function [2] or by assigning fuzzy colors to each pixel by mixing c basic colors in proportion to their memberships. Colors can be selected as per the subject matter of the image or by some pre-established expectations in the medical community.

3 Unsupervised Fuzzy Clustering

Fuzzy clustering can handle the uncertainty and imprecision better than the deterministic or statistical approaches. Among the various fuzzy soft computing approaches, fuzzy c-means (FCM) model and algorithm family is most popular and well-developed least square model. A generalized model of optimization can be defined in terms of an objective function J$_m$ as [2]:

$$\underbrace{min}_{(u,v)}\left\{J_m\left(U,V,w\right)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^{m'}\,d_{ik}^{2}+\sum_{i=1}^{c}w_i\sum_{k=1}^{n}\left(1-u_{ik}\right)^{m'}\right\} \qquad (1)$$

Where U=[u$_{ik}$], a $c\times\ n$ matrix, is the c-partition of the object data $X = \{x_{1,}\ldots,x_n\} \subset R^P$, m$^{'}$ ≥ 1 is a measure of fuzziness, d$_{ik}$ is a Euclidian distance function and $u_{ik}^{m^{'}}$ is the membership measure of kth data object to ith cluster. $V = (v_1, v_2, \cdots v_c) \ \in \ R^{cp}$; $v_i \ \in \ R^p$, is the centroid vector of the clusters and $w = (w_1, w_2, \cdots w_c)^T$; $w_i \ \in \ R^+$, is the weighting vector and for FCM $w_i = 0\ \forall i$. The distance between x_k and cluster centre v_i can be expressed as:

$$d_{ik} = d\left(x_k - v_i\right) = \left(\sum_{j=1}^{m}\left(x_{kj} - v_{ij}\right)^2\right)^{1/2} \tag{2}$$

in m space. For FCM, membership value u_{ik} and the centre v_{ij} can be expressed as:

$$u_{ik}^{(r+1)} = \left(\sum_{j=1}^{c}\left(\frac{d_{ik}^{(r)}}{d_{jk}^{(r)}}\right)^{\frac{2}{\left(m^{'}-1\right)}}\right)^{(-1)} \quad \forall i,k\ for\ I_k = \Phi,\ u_{ik}^{(r+1)} = 0\ \forall i,\ i\epsilon I_k \tag{3}$$

$$and\ \ v_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m^{'}} x_k^j}{\sum_{k=1}^{n} u_{ik}^{m^{'}}} \forall i, m^{'} > 1, \sum_{i\epsilon I_k} u_{ik}^{(r+1)} = 1 \tag{4}$$

With a suitable guess of the initial partition matrix and a convergence limit ϵ_L the above expressions of u and v can be solved using iterative optimization technique as suggested by [3]. To monitor the singularity of the denominator, $d_{jk}^{(r)}$, of the membership function $u_{ik}^{(r+1)}$ the method proposes a bookkeeping system I^k and $\overline{I^k}$ to keep a track of null values of $d_{jk}^{(r)}$. $I_k = \{i|2 \leq c < n; d_{ik}^{r} = 0\}$ is the set of null distance from the kth feature point to the cluster centers and $\overline{I_k} = \{1, 2, \cdots c\} - I_k$.

4 The Recurrent Neural Network

A schematic representation of the state-space model of a batch recurrent neural network (RNN) model is depicted in Fig. 1(b). The hidden layer neurons constitute the state of the network and the output of the hidden layer is fed back to the input layer through context unit. The input of the model consists of inputs from the given pattern and the context units. Following this general configuration of dynamic behaviour of recurrent network, with R^m and R^p as input and output space, may be described by the following two equations:

$$a(t+1) = f\left(w_h^a x_t + w_c^b a_t + b_h\right) \tag{5}$$

$$y(t) = g\left(w_y^a a_t + b_y\right) \tag{6}$$

where $w_h^a : (m \times n)$ is a synaptic weight matrix of the hidden layer $S_{(n\times 1)}$ and input layer $R_{(m\times 1)} : X = \{x_i : i = 1 \cdots m\}$. Bias vectors $b_h : (n \times 1)$ and b_y :

$(p \times 1)$ are connected to the hidden layer and the output layer respectively. Input layer $R_{(m \times 1)}$ provides X_m input in conjunction with $a_{(t-1)}$ from the context unit to the hidden layer. $w_c^b : (m \times n)$ is the synaptic weight matrix of n neurons of the hidden layer and is connected to the feedback signals from the context unit. Linear layer $Y_{(p \times 1)}$ with synaptic matrix $w_y^a : (n \times p)$ and bias b_y generates desired output y(t). The process function $f : (R^m \rightarrow S^n)$ is a diagonal mapping of state-space into itself for some memoryless component-wise nonlinearity $f : (R \rightarrow R)$. The measurement function $g : (S^n \rightarrow Y^p)$ is a linear mapping of S^n space to output space Y^p.

4.1 Training Procedure

The training procedure of the ERNN [4] is implemented by transfer functions (f, g), free parameters (w,b) and backpropagation optimization strategy. All of these form the basis for the approximation of a mapping function f : R^m→R^p, and x→y= f(x) for a training data set $X_m = \{(x_i, y_i), i = 1 \cdots m\}$, where n is the number of sample pattern. While selecting these parameters, the generalization of the trained network with respect to the unknown test data was kept in mind and selected accordingly. As a transfer function we have used hyperbolic tangent sigmoid function (f) and linear function (g). The tangent sigmoid function can accept any value between plus and minus infinity and can transform them into the range –1 to +1. To have a fast and memory efficient error back propagation training method we have used Levenberg-Marquardt nonlinear least square minimization algorithm with a memory reduction factor of magnitude 2. Given a set of training samples $\{x(k), y(k)\, 1 \leq k \leq n\}$ the error backpropagation training begins by feeding all n inputs through the multilayer neural network model and computing the corresponding output $\{z(k)\, 1 \leq k \leq n\}$. Then a sum of square error can be written as:

$$S(w) = \sum_{k-1}^{k} (y(k) - z(k))^2 \tag{7}$$

$$= \sum_{k-1}^{k} (y(k) - (f(x(k)w_h(k)b_h(k)) + g(h(k)w_y(k)b_y(k))))^2 \tag{8}$$

where w_h and w_y are weightage vectors corresponding to the kth input vector. In each iteration w_h and w_y are replaced by a new estimation $(w_i + \delta) : i = h, y$ and δ is approximated by Levenberg-Marquardt optimization method.

4.2 Optimization Method Using LM Algorithm

As an optimization technique we have used the Levenberg-Marquardt algorithm as it outperforms simple gradient descent and other conjugate gradient methods in a wide variety of problems. It can be seen that simple gradient descent and Gauss-Newton optimization are complementary in the advantages they provide. Levenberg algorithm is based on this observation and is written as:

$$w_{i+1} = w_i - (H - \mu I)^{-1} J^T e \tag{9}$$

where Hessian matrix H is evaluated at w_i and e is the vector of network errors. The Jacobian matrix contains first derivatives of the network errors with respect to the weights and biases:

$$J(w) = \frac{\delta e_j}{\delta w_i} \;\; where \;\; 1 \leq j \leq p \;\; and \;\; 1 \leq i \leq m \tag{10}$$

for a performance function f: $\mathrm{R}^m \rightarrow \mathrm{R}^p$ defined as $f(x) = 1/2||e(x)||^2$. When μ is zero, the algorithm is just a newton's method and when large, this shows gradient descent behaviour with very small step size. To speed up the convergence in the direction of small gradient Marquardt replaced the identity matrix with the diagonal of H (Eq. 11). This incorporates a large movement along the directions of smaller gradient.

$$w_{i+1} = w_i - (H + \mu \; diag|H|)^{-1} J^T e \tag{11}$$

where μ is the learning rate. Since H is proportional to the curvature, w_i will have a large step in the direction with low curvature and small step in the direction with high curvature [5]. The Levenberg-Marquardt function is used with a memory reduction factor of two by dividing Hessian as a linear combination of two Jacobians J_1 and J_2:

$$H = J_1^T J_1 + J_2^T J_2 \tag{12}$$

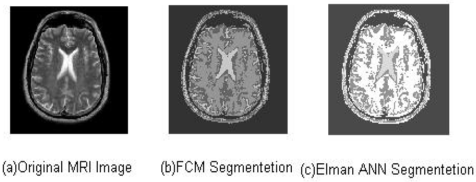

(a)Original MRI Image (b)FCM Segmentetion (c)Elman ANN Segmentetion

Fig. 2. Original MRI (a) with FCM (b) and Elman NN (c) segmented images

5 Experiments and Results

The study has been done on transaxial MR-T2 images (Fig. 2(a)). The image is a 256×256×3 RGB image with voxel dimension 1×1×3 mm^3. The image has been taken from the Harvard Medical School "the whole brain image atlas" image database. Initially, image pixels are (visually) classified into four clusters namely GM, WM, CSF and non-brain matter (cranium and mandible). At the outset we have randomly selected a number of pixels and divided them into training data set and checking data set. The training data set pixels are first classified into four classes by an unsupervised fuzzy clustering (UC) technique. Then the data points are divided into four regions or classes and assigned appropriate membership values corresponding to each of the four classes. Model is trained using a randomly selected n number of training and membership values to each class. This simulation process continues by adjusting different NN parameters till the model reaches the required given accuracy level. After successful convergence, the trained NN engine is verified with the check-data set and their corresponding classes. Once the NN is ready, it can be used to determine the membership values of any input data from the given test images. Visually, the accuracy can be verified by comparing the membership graph of the NN model and the UC model but to confirm the desired accuracy level, a regression analysis has been done between the network response and the corresponding targets. The correlation coefficient we get is almost a unity. In the comparative study of Elman, Back Propagation and Radial Basis Function [6] neural network models, we have found that in a normalized environment Elman model takes least number of epochs (olny 12 compared to 244 and 125 epochs of BP and RBF respectively) to reach the given goal of accuracy and the simulation is also within the acceptable range (Fig. 2(c)).

6 Conclusion

In this article a neuro-fuzzy model has been proposed to deal with the MR medical image segmentation. This model can provide a robust and stable segmentation solution over the intrinsic uncertainty of any MR image. The SC model performs segmentation on the basis of trained activation function with the help of labeled data as identified by the unsupervised method. Quantitative performance and visual comparisons clearly demonstrate the superiority of the combined method. Furthermore, we have found that this proposed segmentation model does not require image preprocessing as the supervised model trains up the activation function with a very large number of labeled data that minimizes the chance of influence of noise components and inhomogeneity, if any, in the image.

References

1. Ray, D., Dutta Majumder, D.: Studies on Some multimodal medical image registration approaches for diagnostics and therapeutic planning: With some case studies. Science Letters, NAS 28(5&6), 135–154 (2005)
2. Sutton, M.A., Bezdek, J.C., Cahoon, T.C.: Image Segmentation by Fuzzy Clustering: Methods and Issues. In: Bankman, I.N. (ed.) Handbook of Medical Imaging-Processing and Analysis, pp. 87–106. Academic Press, London (2000)
3. Bezdek, J.C., Hathaway, R.J., Sabin, M.J., Tucker, W.T.: Convergence theory for fuzzy C-means: counter examples and repairs. IEEE Trans. Syst. Man Cybern. 17873–17877 (1987)
4. Elman, J.L.: Finding structure in time. Cognitive Science 14, 179–221 (1990)
5. Hagan, M.T., Menhaj, M.: Training feedforward networks with the Marquardt algorithm. IEEE Transaction on Neural Networks 5(6), 989–993 (1994)
6. Powell, M.J.D.: Radial basis functions for multivariable interpolation: a review. In: Algorithms for Approximation, pp. 143–167. Clarendon Press, Oxford (1987)

Weak Fuzzy Equivalence and Equality Relations

Branimir Šešelja and Andreja Tepavčević*

Department of Mathematics and Informatics, University of Novi Sad, Serbia

Abstract. Weak fuzzy (lattice valued) equivalences and equalities are introduced by weakening the reflexivity property. Every weak fuzzy equivalence relation on a set determines a fuzzy set on the same domain. In addition, its cut relations are crisp equivalences on the corresponding cut subsets. Analogue properties of weak fuzzy equalities are presented. As an application, fuzzy weak congruence relations and fuzzy identities on algebraic structures are investigated.

AMS Mathematics Subject Classification (2000): primary 03B52, 03E72; secondary 06A15.

Keywords and phrases: lattice-valued fuzzy set, lattice-valued fuzzy relation, block, cut, fuzzy equivalence, fuzzy equality, fuzzy identity.

1 Introduction

Fuzzy equivalence relations belong to the most important notions of fuzzy structures. These were investigated from the beginning of fuzzy era. Instead of citing numerous relevant papers we refer to the book of Bělohlávek [2], in which an extensive list of references is provided. The notion of fuzzy equality was introduced by Höhle ([11]) and then used by many others. In several papers, see e.g. [7,8], Demirci considers particular algebraic structures equipped with fuzzy equality relation, see also [4]. Bělohlávek (see [2], references there, and recent paper with Vychodil [3]) introduces and investigates algebras with fuzzy equalities.

Our approach in the present paper is cutworthy in the sense that cut substructures preserve crisp fuzzified properties. Due to this reason, the co-domain of our structures and relations is a fixed complete lattice, without additional operations. Differences of our approach and the foregoing mentioned ones is commented in the concluding section.

We introduce the notion of weak equivalence relation by weakening the reflexivity property. In this way, we obtain relations which determine fuzzy subsets on the same domain. In addition, the cut relations are precisely fuzzy equivalences on the cuts of the mentioned fuzzy set. In the case of weak fuzzy equality, these cuts are crisp equalities on the corresponding cut subsets.

* The research supported by Serbian Ministry of Science and Technology, Grant No. 144011 and by the Provincial Secretariat for Sci. and Techn. Development, Autonomous Province of Vojvodina, Grant "Lattice methods and applications".

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 134–139, 2009.

Our applications are situated in fuzzy universal algebra. We define a notion of fuzzy weak congruence relation on an algebra. It is a fuzzy weak equivalence which fulfills a substitution property with operations. Then, fuzzy subsets are fuzzy subalgebras and cut relations are crisp congruence relations on these subalgebras. Applying the obtained results, we have got some new results on fuzzy identities.

Although our research is mostly theoretical, we see its importance in real life applications. It is well known that fuzzy equivalences better model classification of objects then crisp relations. In case of weak equivalences, these applications, particulary in pattern recognition, can be even more suitable. This would be the task of our future investigation.

2 Preliminaries

We present basic notions related to fuzzy (lattice valued) structures and relations.

Let $(L, \wedge, \vee, \leq)$ be a complete lattice (for the notions from lattice and order theory see e.g. [6]).

A fuzzy set μ on a nonempty set A is a function $\mu : A \rightarrow L$. For $p \in L$, a **cut set**, or a p-**cut** (sometimes simply a **cut**) of μ is a subset μ_p of A which is the inverse image of the principal ideal in L, generated by p:

$$\mu_p = \{x \in A \mid \mu(x) \geq p\}.$$

Other notions from the basics of fuzzy sets that we use are either defined in the sequel where they appear, or they can be found in any introductory text about fuzziness.

An L-valued relation R on A is
reflexive if $R(x,x) = 1$, for every $x \in A$;
symmetric: $R(x,y) = R(y,x)$, for all $x, y \in A$;
transitive: $R(x,y) \wedge R(y,z) \leq R(x,z)$, for all $x, y, z \in A$.

An L-valued relation R on A is a **lattice valued** (L**-valued**) **equivalence relation** on X if it is reflexive, symmetric and transitive. An L-valued equivalence relation R on A is an L**-valued equality relation** if it fulfills the following

$$R(x,y) = 1 \text{ if and only if } x = y.$$

Throughout the paper, L is supposed to be a fixed complete lattice.

3 Lattice Valued Weak Equivalence Relations

An L-valued relation R on A is **weakly reflexive** if for all $x, y \in A$,

$$R(x,x) \geq R(x,y).$$

An L-valued relation R on A is a **weak** L**-valued equivalence relation** on A if it is weakly reflexive, symmetric and transitive. In particular a **weak** L**-valued equality** R **on** A is an L-valued equivalence which fulfills also condition

$$\text{if } u \neq v, \text{ then } R(u,v) < \bigwedge_{x \in A} R(x,x).$$

In the following we demonstrate that each weak L-valued equivalence relation on a set determines a fuzzy subset on the same domain and we investigate the connection between the corresponding cut relations and cut subsets.

Theorem 1. *If R is a weak L-valued equivalence relation on a set A, then the mapping $\mu[R] : A \to L$, defined by*

$$\mu[R](x) := R(x,x)$$

is an L-valued subset of A. In addition, for every $p \in L$, the cut relation R_p is a crisp equivalence relation on the cut subset $\mu[R]_p$.

Proof. $\mu[R](x)$ is an L-valued subset of A by the definition. Let $p \in L$. $x \in \mu[R]_p$ if and only if $R(x,x) \geq p$, and $(x,x) \in R_p$. Hence, R_p is reflexive. Symmetry and transitivity of the cut R_p follow directly from the symmetry and transitivity of fuzzy relation R. □

Particular case of the above claim is obtained for fuzzy weak equalities.

Theorem 2. *If $R : A^2 \to L$ is a lattice valued weak equality on a set A, then for every $p \geq \bigwedge_{x\in A} R(x,x)$, the cut relation R_p is a crisp equality on the cut subset $\mu[R]_p$.*

Proof. Let $p \geq \bigwedge_{x\in A} R(x,x)$ and $x \in \mu[R]_p$. Then, $R(x,x) \geq p$ and $(x,x) \in R_p$. Let $(x,y) \in R_p$. Then $R(x,y) \geq p$. By the definition of weak L-valued equality, if $x \neq y$, then $R(x,y) < \bigwedge_{x\in A} R(x,x)$, hence, $x = y$, i.e., R_p is a crisp equality on $\mu[R]_p$. □

4 Application in Algebra

4.1 *L*-Valued Subalgebras

We advance some notions from fuzzy universal algebra, together with their relevant properties; more about crisp analogous properties can be found e.g., in the book [5].

Let $\mathcal{A} = (A, F)$ be an algebra and L a complete lattice. As it is known, an L**-valued** (**fuzzy**) **subalgebra** of $\mathcal{A}$ is any mapping $\mu : A \to L$ fulfilling the following:

For any operation f from F, $f : A^n \to A, n \in \mathbf{N}$, and all $x_1, \ldots, x_n \in A$, we have that

$$\bigwedge_{i=1}^{n} \mu(x_i) \leq \mu(f(x_1, \ldots, x_n)).$$

For a nullary operation (constant) $c \in F$, we require that $\mu(c) = 1$, where 1 is the top element in L.

Next, let $R : A^2 \to L$ be an L-valued relation on A (underlying set of the algebra $\mathcal{A}$).

R is said to be **compatible** with operations on $\mathcal{A}$ if for any (n-ary) $f \in F$ and all $x_1, \ldots, x_n, y_1, \ldots, y_n \in A$, we have that

$$\bigwedge_{i=1}^{n} R(x_i, y_i) \leq R(f(x_1, \ldots, x_n), f(y_1, \ldots, y_n)).$$

Let $E : A^2 \rightarrow L$ be a fuzzy equality relation on A, which is compatible with operations on $\mathcal{A}$. We call E a **compatible L-valued equality** on $\mathcal{A}$. Analogously, we define a **weak compatible L-valued equality** on $\mathcal{A}$ as a weak L-valued equality on A, which is also compatible with operations on $\mathcal{A}$.

The following facts about cuts of fuzzy subalgebras and of compatible fuzzy equalities are known.

Theorem 3. *Let $\mathcal{A}$ be an algebra, L a complete lattice, $\mu : A \rightarrow L$ a fuzzy subalgebra of $\mathcal{A}$, and E a compatible fuzzy equality on $\mathcal{A}$. Then for every $p \in L$,*

(i) the cut μ_p of μ is a subuniverse of $\mathcal{A}$, and

(ii) the cut E_p of E is a congruence relation on $\mathcal{A}$.

4.2 Weak Lattice Valued Congruences

Theorem 4. *If $R : A^2 \rightarrow L$ is a lattice valued weak congruence on an algebra $\mathcal{A}$, then the mapping $\mu[R] : A \rightarrow L$, defined by*

$$\mu[R](x) := R(x, x)$$

is an L-valued subalgebra of $\mathcal{A}$.

Proof. Let $\mathcal{A} = (A, F)$ be an algebra and f an n-ary operation from F, $f : A^n \rightarrow A, n \in \mathbf{N}$ and let $x_1, \ldots, x_n \in A$.

Then

$$\bigwedge_{i=1}^{n} \mu[R](x_i) = \bigwedge_{i=1}^{n} R(x_i, x_i) \leq R(f(x_1, \ldots, x_n), f(x_1, \ldots, x_n)) =$$
$$= \mu[R](f(x_1, \ldots, x_n)).$$
□

Theorem 5. *If $R : A^2 \rightarrow L$ is a lattice valued weak congruence on $\mathcal{A}$, then for every $p \in L$, the cut relation R_p is a congruence relation on the cut subalgebra $\mu[R]_p$.*

Proof. Let $p \in L$. It is known from already mentioned facts that the cut relation R_p is symmetric, transitive and compatible. To prove reflexivity, suppose that $x \in \mu[R]_p$. Then $R(x, x) \geq p$ and $(x, x) \in R_p$. □

As a straightforward consequence of Theorem 2, we present the following property of lattice valued compatible weak equalities.

Corollary 1. *Let $R : A^2 \rightarrow L$ be a lattice valued compatible weak equality on $\mathcal{A}$ and for $x \in A$, let $R(x, x) = p$. Then R_p is an equality relation on the cut subalgebra $\mu[R]_p$.* □

4.3 Fuzzy Identities

If E is a compatible L-valued equality on an algebra $\mathcal{A}$, and t_1, t_2 are terms in the language of $\mathcal{A}$, we consider the expression $E(t_1, t_2)$ as a **fuzzy identity with respect to** E, or (briefly) **fuzzy identity**. Suppose that $x_1, \ldots, x_n$ are variables appearing in terms t_1, t_2. We say that a fuzzy subalgebra μ of $\mathcal{A}$ **satisfies** a fuzzy identity $E(t_1, t_2)$ if for all $x_1, \ldots, x_n \in A$

$$\bigwedge_{i=1}^{n} \mu(x_i) \leq E(t_1, t_2). \qquad (1)$$

In the present investigation we additionally consider the case in which the relation E appearing in formula (1) is a *weak* compatible L-valued equality on $\mathcal{A}$. Then also we say that μ **satisfies** the fuzzy identity $E(t_1, t_2)$.

Proposition 1. [15] *Let $\mathcal{A}$ be an algebra satisfying a (crisp) identity $t_1 = t_2$ whose variables are $x_1, \ldots, x_n$. Let also L be a complete lattice, $\mu : A \to L$ a fuzzy subalgebra of $\mathcal{A}$, and E a compatible fuzzy equality on $\mathcal{A}$. Then, any fuzzy subalgebra $\mu : A \to L$ satisfies fuzzy identity $E(t_1, t_2)$.*

Theorem 6. [15] *Let $\mathcal{A}$ be an algebra, L a complete lattice, $\mu : A \to L$ a fuzzy subalgebra of $\mathcal{A}$, and E a compatible fuzzy equality on $\mathcal{A}$. Let also μ satisfies a fuzzy identity $E(t_1, t_2)$ in the sense of formula (1). Then for every $p \in L$, if μ_p is not empty then the crisp quotient algebra $\mu_p/E_p(\mu_p)$ satisfies the (crisp) identity $t_1 = t_2$.*

If we apply the above results to weak compatible fuzzy equalities, then we get that cut subalgebras (and not quotient algebras as above) fulfill the corresponding crisp identities, as demonstrated by the following theorem.

Theorem 7. *Let $\mathcal{A}$ be an algebra, L a complete lattice, $\mu : A \to L$ a fuzzy subalgebra of $\mathcal{A}$, and E a weak compatible fuzzy equality on $\mathcal{A}$. Let also μ satisfies a fuzzy identity $E(t_1, t_2)$ in the sense of formula (1). Then for every $p \geq \bigwedge_{x \in A} E(x, x)$, if μ_p is not empty then the crisp subalgebra μ_p of $\mathcal{A}$ satisfies the (crisp) identity $t_1 = t_2$.*

Proof. Let $p \geq \bigwedge_{x \in A} E(x, x)$. Let $x_1, ..., x_n$ be elements from μ_p. Then, $\mu(x_1) \geq p$,..., $\mu(x_n) \geq p$. Hence,

$$p \leq \bigwedge_{i=1}^{n} \mu(x_i) \leq E(t_1(x_1, ..., x_n), t_2(x_1, ..., x_n)),$$

since μ satisfies fuzzy identity E. On the other hand, by the definition of weak compatible fuzzy equality, we have that if $t_1(x_1, ..., x_2) \neq t_2(x_1, ..., x_n)$, then

$$E(t_1(x_1, ..., x_n), t_2(x_1, ..., x_n)) < \bigwedge_{x \in A} E(x, x) \leq p,$$

which is a contradiction with the fact that $E(t_1(x_1, ..., x_n), t_2(x_1, ..., x_n)) \geq p$. Hence, $t_1(x_1, ..., x_2) = t_2(x_1, ..., x_n)$, and μ_p satisfies the identity $t_1 = t_2$. □

5 Conclusions

The paper deals with weak fuzzy equivalences on a set and with applications in algebra. From the algebraic aspect, our approach differs from the one developed in [3] (see also [2]). First, due to our cutworthy framework, we use lattice theoretic operations, and not additional ones (existing in residuated lattices). Next, we start with a crisp algebra and use a fuzzy equality to introduce fuzzy identities. In addition, our fuzzy equality is weakly reflexive, which, due to compatibility, enables determination of fuzzy subalgebras by its diagonal.

Our next task is to introduce and investigate the corresponding (weak) fuzzy partitions. Apart from algebraic application, these could be used in pattern recognition. Indeed, in addition to crisp and fuzzy partitions, weak partitions could model not only properties of whole domains, but also their fuzzy sub-domains (observe that a weak fuzzy equivalence possesses the fuzzy diagonal instead of the constant).

References

1. De Baets, B., Mesiar, R.: T-partitions. Fuzzy Sets and Systems 97, 211–223 (1998)
2. Bělohlávek, R.: Fuzzy Relational Systems: Foundations and Principles. Kluwer Academic/Plenum Publishers, New York (2002)
3. Bělohlávek, R., Vychodil, V.: Algebras with fuzzy equalities. Fuzzy Sets and Systems 157, 161–201 (2006)
4. Bodenhofer, U., Demirci, M.: Strict Fuzzy Orderings with a given Context of Similarity. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 16(2), 147–178 (2008)
5. Burris, S., Sankappanavar, H.P.: A Course in Universal Algebra. Springer, Heidelberg (1981)
6. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order. Cambridge University Press, Cambridge (1992)
7. Demirci, M.: Vague Groups. J. Math. Anal. Appl. 230, 142–156 (1999)
8. Demirci, M.: Foundations of fuzzy functions and vague algebra based on many-valued equivalence relations part I, part II and part III. Int. J. General Systems 32(3), 123–155, 157–175, 177–201 (2003)
9. Di Nola, A., Gerla, G.: Lattice valued algebras. Stochastica 11, 137–150 (1987)
10. Goguen, J.A.: L-fuzzy Sets. J. Math. Anal. Appl. 18, 145–174 (1967)
11. Höhle, U.: Quotients with respect to similarity relations. Fuzzy Sets and Systems 27, 31–44 (1988)
12. Malik, J.N., Mordeson, D.S., Kuroki, N.: Fuzzy Semigroups. Springer, Heidelberg (2003)
13. Montes, S., Couso, I., Gil, P.: Fuzzy $\delta - \varepsilon$-partition. Information Sciences 152, 267–285 (2003)
14. Murali, V.: Fuzzy equivalence relations. Fuzzy Sets and Systems 30, 155–163 (1989)
15. Šešelja, B., Tepavčević, A.: Fuzzy identities. In: Proceedings of FUZZ-IEEE 2009, pp. 1660–1663 (2009)
16. Šešelja, B., Tepavčević, A.: On Generalizations of Fuzzy Algebras and Congruences. Fuzzy Sets and Systems 65, 85–94 (1994)
17. Šešelja, B., Tepavčević, A.: Fuzzy groups and collections of subgroups. Fuzzy Sets and Systems 83, 85–91 (1996)
18. Tepavčević, A., Vujić, A.: On an application of fuzzy relations in biogeography. Information Sciences 89(1-2), 77–94 (1996)

Estimation of Tailor-Welded Blank Parameters for Acceptable Tensile Behaviour Using ANN

Abhishek Dhumal[1], R. Ganesh Narayanan[1], and G. Saravana Kumar[2,*]

[1] Department of Mechanical Engineering, Indian Institute of Technology Guwahati, Guwahati, India
[2] Department of Engineering Design, Indian Institute of Technology Madras, Chennai, India
adhumal@iitg.ernet.in, ganu@iitg.ernet.in, gsaravana@iitm.ac.in

Abstract. The tensile and forming behavior of Tailor-welded blanks (TWB) is influenced by many parameters like thickness ratio, strength ratio, and weld conditions in a synergistic fashion. It is necessary to predict suitable TWB conditions for achieving better stamped product made of welded blanks. This work primarily aims at developing an artificial neural network (ANN) model to predict the TWB parameters for acceptable tensile behavior of welded blanks made of steel grade and aluminium alloy base materials. The important tensile characteristics of TWB like limit strains, failure location, minimum thickness, strain path are considered within chosen range of varied blank and weld condition. Through out the work, ABAQUS 6.7-2® finite element (FE) code is used to simulate the tensile behavior and to generate data required for training the ANN. Predicted results from ANN model are compared and validated with FE simulation. The predictions from ANN are with acceptable prediction errors.

Keywords: TWB Parameters, Inverse Modeling, Neural Networks, Parameter Estimation.

1 Introduction

Tailor-welded blanks (TWB) are blanks with sheets of similar or dissimilar thicknesses, materials, coatings welded in a single plane before forming. They are widely used in automotive industries because of their great benefits in reducing manufacturing costs, decrease vehicle weight, great flexibility in component design, reduction in fuel consumption, improved corrosion resistance and product quality etc. [1]. Formability characteristics of TWBs is affected synergistically by weld conditions such as weld properties, weld orientation and weld location, thickness difference and strength difference between the sheets and hence it is difficult to design the TWB conditions that can deliver a good stamped product with similar formability as that of un-welded blank. In this context, few research groups have aimed at predicting the formability of welded blanks by using different necking, limit strain theories and finite element (FE) simulations and compared with that of from experiments [2,3,4]. Developing decision support system based on soft computing techniques like artificial neural network

[*] Corresponding Author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 140–145, 2009.

(ANN), especially in fields like material forming and deformation behavior is of interest to manufacturing, design engineers and scientists for long time. In the earlier work [5], the authors have predicted the global TWB tensile behavior like yield strength, ultimate tensile strength, uniform elongation, strain hardening exponent and strength coefficient for a wide range of thickness and strength combinations, weld properties, orientation by ANN technique. However, there has not been any study on the TWB parameter estimation using ANN. In the present study an attempt has been made to find TWB parameters for better tensile behavior using ANN based inverse modeling.

2 Methodology

The flow chart describing the methodology followed in this work is shown in Fig. 1. In the present work, ANN models are developed to predict the TWB conditions viz., thickness difference, strength difference, weld properties, orientation etc. for good tensile behavior described by limit strains, failure location, minimum thickness, strain path of TWBs. Tensile test sample is used to simulate the tensile process using ABAQUS 6.7-2® FE code. The sheet base materials considered for the present work

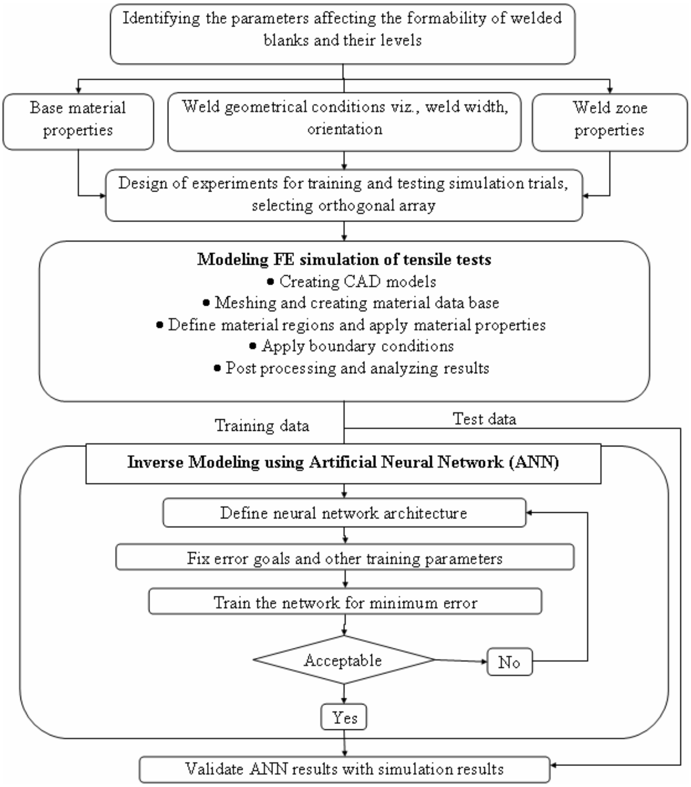

Identifying the parameters affecting the formability of welded blanks and their levels
Base material properties
Weld geometrical conditions viz., weld width, orientation
Weld zone properties
Design of experiments for training and testing simulation trials, selecting orthogonal array
Modeling FE simulation of tensile tests
• Creating CAD models
• Meshing and creating material data base
• Define material regions and apply material properties
• Apply boundary conditions
• Post processing and analyzing results
Training data
Test data
Inverse Modeling using Artificial Neural Network (ANN)
Define neural network architecture
Fix error goals and other training parameters
Train the network for minimum error
Acceptable
No
Yes
Validate ANN results with simulation results

Fig. 1. Overall methodology of FE simulation and ANN modeling

are a low carbon steel grade and formable aluminum alloy (Table 1). Initially, the number of TWB parameters and their levels were decided for the application of Taguchi design of experiment (DOE) methods for systematic analyses. For the present work an L_{27} orthogonal array is selected for conducting simulations at three levels for six factors. The six factors considered at three levels are shown in Table 2 for steel and aluminium TWB. These parameters are schematically represented in Fig. 2. The levels of parameters were chosen such that it covers practically all the combinations involved in typical experiments and TWB parts [1]. Tensile behavior for varied TWB conditions viz., thickness difference, strength difference, weld properties, orientation, was simulated using ABAQUS 6.7-2® and output parameters i.e. limit strains, failure location, minimum thickness, strain path are numerically computed. ANN models are developed using a large data set obtained from simulation trials that can predict the weld and blank conditions for a given tensile behavior of TWB within a chosen range of weld and blank conditions. Design of experiments technique is used to optimize the number of simulations. The accuracy of ANN prediction was validated with simulation results for chosen intermediate levels.

Table 1. Material properties of steel grade and aluminum alloy base materials used for TWB

Property	Steel sheet [4]		Aluminium alloy sheet	
	Base	Weld	Base	Weld
Young's modulus (E), GPa	210	210	77	77
Density (ρ), kg/m^3	7860	7860	2700	2700
Poisson's ratio (v)	0.3	0.3	0.3	0.3
Plastic r_0	1.21	1	0.7	1
strain r_{45}	1.08	1	0.6	1
ratios r_{90}	1.68	1	0.8	1
Strain hardening exponent (n)	0.27	--	0.172	--

Table 2. Material properties of steel grade and aluminum alloy base materials used for TWB

Parameters\levels	Steel grade base metal			Aluminum alloy base metal		
	1	2	3	1	2	3
Thickness ratio T_1/T_2, T_2 = 1.5 mm	0.5	0.75	1	0.5	0.75	1
Strength ratio YS_1/YS_2	0.5	0.75	1	0.5	0.75	1
	YS_2 = 300 MPa			YS_2 = 380 MPa		
Weld orientation(°)	0	45	90	0	45	90
Weld 'n' value (n_w)	0.1	0.15	0.2	0.1	0.13	0.15
Weld yield strength, (YS_w), MPa	125	250	500	150	300	400
Weld width (W), mm	2	5	10	2	5	10

2.1 Modeling Simulation of Tensile Test for Welded Blanks

Solid models of tensile specimen as per the ASTM E 646-98 specifications (Fig. 2) were modeled in Pro-E® and imported into ABAQUS 6.7-2® for preprocessing, performing simulations and post processing. To induce necking after applying displacement boundary condition in ABAQUS 6.7-2®, a small heterogeneity in the form of a 10 mm notch was introduced in the geometry of the specimen [6]. These models

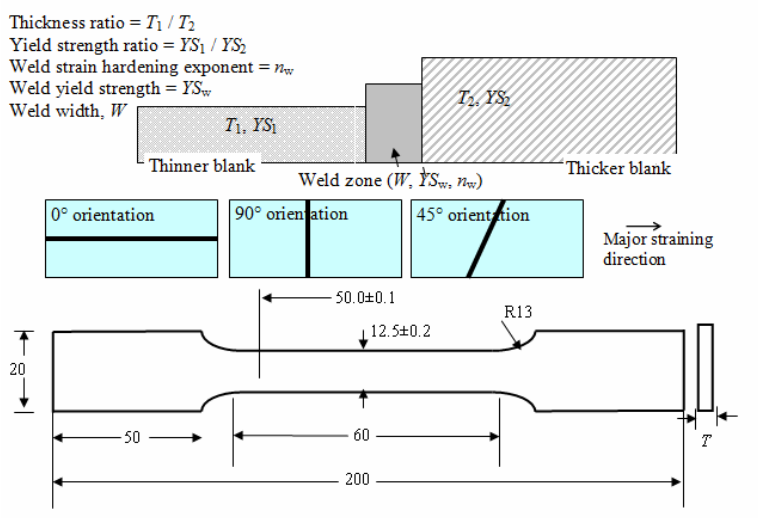

Thickness ratio = T_1 / T_2
Yield strength ratio = YS_1 / YS_2
Weld strain hardening exponent = n_w
Weld yield strength = YS_w
Weld width, W
T_1, YS_1
T_2, YS_2
Thinner blank
Weld zone (W, YS_w, n_w)
Thicker blank
0° orientation
90° orientation
45° orientation
Major straining direction
50.0±0.1
12.5±0.2
R13
20
50
60
200
T

Fig. 2. Schematic representation of controllable factors and ASTM E 646-98 standard tensile testing specimen. All dimensions in mm.

were meshed using 3D deformable quadrilateral shell element with 1mm size and was divided into three different regions viz., weld region, base material 1 and base material 2 to construct meshed models of TWB for varied weld orientations [4]. The material properties were assigned to weld zone and base metals according to the different parameter levels (Tables 1 and 2) in the orthogonal array. Displacement boundary conditions are applied to the tensile sample such that one end of the specimen is fixed and the other end is given some finite displacement with a velocity of 0.5 mm/min. In this work Swift law was used to describe the strain hardening behavior, thickness gradient criterion is used as necking or failure criterion and Hill's 1948 isotropic hardening yield criterion was used as the plasticity model for all materials [4]. The tensile response of the TWB was obtained and from this the following properties were evaluated:

1) **Limit strain**: As per thickness gradient theory, when the necking occurs, in that progression or increment the major and minor strain corresponding to thicker element is the limit strain.

2) **Failure location:** It is the distance from the fixed end to the thicker element in the progression where necking has occurred.

3) **Minimum thickness:** It is the minimum thickness of the element of specimen in the progression where necking has occurred.

4) **Strain path**: It is the plot between major and minor strain from the starting progression to the progression where necking has occurred.

These output tensile properties of TWB from 27 simulation trials each for steel grade and aluminium alloy base materials were used for ANN modeling.

2.2 TWB Parameter Estimation Using ANN

An inverse technique using ANN is described for estimating TWB parameters for good formability. ANN based inverse modeling has been used in various fields of modeling and has shown better performance when compared to analytical and numerical methods [7]. The inverse model of the TWB tensile test simulation gives the test input i.e. the TWB parameters for a given test output i.e. the tensile behaviour. In the present formulation, it is assumed that an inverse exists, i.e. for a given tensile behaviour, a set of TWB parameters can be obtained. The tensile behaviour obtained from FEM i.e. limit strain, failure location, minimum thickness and strain path are given as input and TWB conditions i.e. thickness ratio, strength ratio, weld yield strength, weld '*n*' value, weld orientation and weld width are predicted. The ANN models were constructed and the available simulation data sets, 27 data sets were used to train and two intermediate data sets were utilized for testing. A feed forward back propagation algorithm is selected to train the network using scaled conjugate gradient algorithm to minimize the error in Matlab®.

3 Results and Discussion

The results of tensile test simulation of the TWBs were used for training the ANN. As an example thickness contour after simulation of one such TWB with longitudinal weld is shown in Fig. 3. The thickness gradient criterion is satisfied in the location of failure. The figure also shows the thicker and thinner elements corresponding to the failure location. The various ANN parameters like number of hidden layers, neurons, and transfer functions were optimized based on many trials to predict the outputs within the normalized error limit of 10^{-4}. Various network structures with one and two hidden layers with varying number of neurons in each layer were examined. Out of all the networks tested, a network with two hidden layer reached the best performance goal when compared with other networks. The final optimized ANN architecture is 5 input neurons (corresponding to 5 outputs), a double hidden layer with 12 and 6 neurons and 6 output neurons (corresponding to 6 factors) with 'tan sigmoid' and 'pure linear' as transfer functions. Table 3 summarizes the average error statistics pertaining to ANN prediction for training and two intermediate test data for TWB made of steel and aluminium alloy base materials respectively. At the design stage for TWB parameter prediction an error range of 5-10% is considered acceptable between the ANN and simulation results. This is justified since the TWB parameters predicted will nevertheless need to be further modified considering the manufacturing process constraints and material availability. Also tensile behaviour is only a basic criterion for good formability. It can be seen that almost all the output parameters are predicted within acceptable error limits. The authors are currently working on extending this framework for predicting parameters for good formability of TWBs.

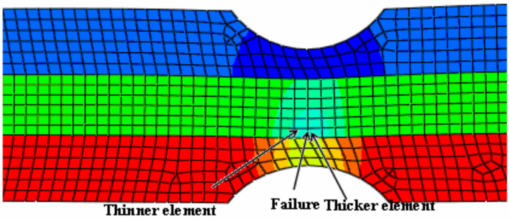


Fig. 3. Failure location in TWB with longitudinal weld

Table 3. Error values for testing and training data for inverse modeling using ANN

Parameters	Steel grade base metal				Aluminum alloy base metal			
	Training		Testing		Training		Testing	
Error (%)	μ	σ	μ	σ	μ	σ	μ	σ
Thickness ratio	0.01	0.43	6.10	1.86	0.09	0.00	1.70	3.51
Strength ratio	0.00	0.02	3.61	0.73	0.00	0.01	1.62	0.87
Weld orientation	0.12	0.76	6.20	1.23	0.01	0.56	5.25	1.82
Weld '*n*' value	0.17	0.13	4.20	0.09	0.02	0.03	4.21	2.64
Weld yield strength	0.36	1.74	8.00	2.11	0.88	0.41	7.83	2.70
Weld width	0.91	1.00	2.10	0.65	0.53	0.25	6.43	1.53

References

1. The Ultra Light Steel Auto body, `http://www.ulsab.org`
2. Jie, M., Cheng, C.H., Chan, L.C., Chow, C.L., Tang, C.Y.: Experimental and Theoretical Analysis on Formability of Aluminum Tailor-Welded Blanks. J. Engg. Mat. & Tech. 129, 151–158 (2007)
3. Anand, D., Boudreau, G., Andreychuk, P., Chen, D.L., Bhole, S.D.: Forming Behaviour of Tailor (Laser) Welded Blanks of Automotive Steel Sheet. Canadian J. Metallurgy and Mat. Sc. 45(2), 187–197 (2006)
4. Ganesh Narayanan, R., Narasimhan, K.: Predicting the Forming Limit Strains of Tailor Welded Blanks. J. Strain Analysis for Engg. Des. 43(7), 551–563 (2008)
5. Veerababu, K., Ganesh Narayanan, R., Saravana Kumar, G.: An Expert System based on Artificial Neural Network for Predicting the Tensile Behavior of Tailor Welded Blanks. Expert Sys. Appl. 36, 10683–10695 (2009)
6. Holmberg, S., Enquist, B., Thilderkvist, P.: Evaluation of Sheet Metal Formability by Tensile Tests. J. Mat. Processing Tech. 145, 72–83 (2000)
7. Saravana Kumar, G., Kalra, P.K., Dhande, S.G.: Parameter Estimation for B-spline Curve Fitting using Neural Networks. In: 2nd International Conference on CIRAS, Singapore, p. 127 (2003)

Identification of N-Glycosylation Sites with Sequence and Structural Features Employing Random Forests

Shreyas Karnik[1,3], Joydeep Mitra[1], Arunima Singh[1], B.D. Kulkarni[1],
V. Sundarajan[2], and V.K. Jayaraman[2,⋆]

[1] Chemical Engineering and Process Development Division, National Chemical Laboratory, Pune, India - 411008

[2] Center for Development of Advanced Computing, Pune University Campus, Pune, India - 411007
jayaramanv@cdac.in

[3] School of Informatics, Indiana University, Indianapolis, IN, USA, 46202

Abstract. N-Glycosylation plays a very important role in various processes like quality control of proteins produced in ER, transport of proteins and in disease control.The experimental elucidation of N-Glycosylation sites is expensive and laborious process. In this work we build models for identification of potential N-Glycosylation sites in proteins based on sequence and structural features.The best model has cross validation accuracy rate of 72.81%.

1 Introduction

Carbohydrates that are attached to proteins play various important roles in biological systems. One of the major functions of protein linked glycans is to provide additional epitopes for the sake of recognition by protein receptors. [1,2] This type of recognition events are involved in a myriad of processes such as initiation of inflammation, protein trafficking and defense mechanisms [3].N-glycosylation refers to the addition of oligomeric glycans to asparagine (ASN or N) residues on proteins. The sequence motif Asn(N)-Xaa-Ser(S)/Thr(T) has been defined are prerequisite for N-glycosylation. In the motif N-X-S/T, X position can be occupied by any amino acid except for proline[4]. In some cases of N-Glycosylation the motif N-X-C (Cystine) is also found but vary rarely. The presence of N-X-S/T motif is a necessary but not sufficient criterion for N-Glycosylation.

Experimental determination of N-glycosylation in proteins is a laborious and expensive process[5]. Although methods like X-ray crystallography provide direct and unambiguous evidence for the glycan occupied sites, evidence from X-ray crystallography for the unoccupancy of a site is more ambiguous because the absence of the glycan is only one reason for the absence of resolved electron density. Thus, there is need for computational approaches in determining the

⋆ Corresponding Author

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 146–151, 2009.

glycosylation of proteins from the available sequence and structural information. Machine learning methods are explored to provide reliable and cost effective predictive model based on the known data of the proteins several approaches that handle this problem.Gupta et al.[6] have used sequence features for the prediction of potential N-Glycosylation sites in humans using neural networks. In a recent study Caragea et. al[7] have used ensemble of Support Vector Machines (SVM) classifiers using string kernels to identify potential glycosylated sites. In their comprehensive study on protein environment of N-glycosylation sites Petrescu et al. and Ben-Dor et al.[5,8]have indicated that the sequence and structural properties play an important role in deciding the occupancy of the sequon. They came up with the following observations in the analysis of the N-Glycosylation sites :

1. Glycosylation sites occur on hydrophobic regions of the protein surface.
2. Marked preference for hydrophobic residues at both the sides on the glycosylated sites.
3. There is a clear preference for nonpolar amino acids in positions +2 to +4 and in particular for aromatic residues in position +2 and +1, small nonpolar amino acids in position +1, and bulky hydrophobic amino acids in positions +3 to +5.
4. There is a notable reduction in the probability of finding acidic residues immediately before the glycosylation site.
5. When secondary structure around these sites was analyzed they found a bias in favor of turns and bends.

Apart from the above mentioned factors, accessibility and conformational factors play an important role in the process of glycosylation. It is clear that identification of structural and sequence features that correlate with N-Glycosylation is important. In our study we have employed Random Forest(RF)[14] classifier for selection of the most important features and prediction of N-Glycosylation sites in proteins.

2 Materials and Methods

Collection of Data: “ASN GLYCOSYLATION” was used as keyword for advanced search at the RCSB PDB[9] server. 1671 hits were found which were filtered at 90% sequence similarity and the selected 465 structures were downloaded. Occupied glycosylation sites (NXS/T) in each structure were determined by the MODRES field in the PDB file while the unoccupied sites were found by scanning the structure for NXS/T sites other than the ones mentioned in the MODRES field. For the negative dataset we used PDB select 25 data and all the N-X-S/T sites, which are not glycosylated, were taken. From the positive and the negative sequences a window of fixed length (11,19 and 35) with(N-X-S/T) at the center flanked by equal number of amino acids on both sides were extracted. After the collection of the subsequences containing the glycosylated sites the task at hand was to quantify the innate patterns in the sequences into features that can be used as knowledge by the machine learning algorithm to identify glycosylation sites. We considered to use the amino acid frequencies to begin with but literature sources[5,8] indicated that the process of glycosylation has a interplay at the compositional and the structural levels so we incorporated this information in terms of sequence and structure based features described below. We have used Random Forest as the classification algorithm which also has the ability to select the most informative features from a large number of features, because of this capability of Random Forests we used a variety of features that can help us gain insights into the signals in the subsequences that are responsible for the sequon to get glycosylated. We have used features that represent sequence and structure related information of glycosylated subsequences.

The sequence and structural features comprises of the following features:

- Sequence Dependant Features
  - Amino acid and Dipeptide frequencies
  - Features based on physiochemical properties of amino acids by the PROFEAT Server[10]
- Structural Features
  - Secondary Structure assignment from DSSP[11]
  - Solvent Accessibility
  - Contact Number
  - Contact Order
  - Backbone Torsional Angles

The PROFEAT Server[10] calculates different autocorrelation descriptors these are defined on the basis of of the distribution of the different amino acid properties along the sequence. These autocorrelation descriptors for a particular property can provide insights about the periodicity of amino acid physiochemical properties. Amino acid properties employed for the computation of autocorrelation descriptors were taken from AAIndex [12]. A detailed list of the amino acid properties used for the calculation of autocorrelation descriptors is given in Table I.We have used Random Forest[13]to build a model that will aid in the identification of the N-Glycosylated sites from the two set of the features described above. Random Forest is an algorithm for classification and regression developed by Leo Breiman[14] that uses an ensemble of trees for classification and regression. Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of variables is a random subset of the variables (*mtry*). Thus, Random Forest uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners, and random variable selection for tree building. Each tree is unpruned (grown fully). The trees grown then vote on the class for a given input, the class that gets the maximum number of votes is then assigned to the input. Random forest has excellent performance in classification tasks. It has been successfully employed to various problems from the life sciences domain[15,16,17,18] it has several characteristics that make it ideal for these data sets. We used the Random Forest implementation in R[19] by Andy Liaw and Matthew Wiener[20] based on original code in FORTRAN by Breiman. The methodology applied for the selection of the best features is as follows:

1. Start with the full set of features.
2. Tune the *mtry* parameter to find the optimal value (usually in the vicinity of the square root of the total number of features).
3. Once an optimal value of *mtry* has been found build Random Forest model and get the variable importance measure for each feature. Also do 10-fold cross-validation in this step.
4. Select half the number of top scoring features and build Random Forest model again. Monitor Mathew's Correlation Coefficient (MCC) on the training dataset each time.
5. Repeat steps 2-4 till the highest MCC is obtained.

3 Results and Discussion

The study attempts to build a model for identification of N-Glycosylation sites using Random Forests Classifier. We used sequence and structure dependent features for building models that could identify the N-Glycosylation sites with good accuracy. We used the subsequence lengths of 11,19 and 35 for feature calculation. Models for the classification of N-Glycosylation sites using different subsequence lengths generated using the methodology described above. In all the models the model with subsequence length 35 and 28 features (using feature selection) performed best. This model has accuracy 72.81% and MCC 0.4426; summary of this model is given in Table 1.

Table 1. Summary of the Best Model

Model Description	Accuracy	TP[a]	TN	FP	FN	Sensitivity	Specificity	MCC
Model 1 [b]	72.81	591	590	212	211	0.716	0.735	0.442

[a] TP:-True Positives, TN:- True Negatives,FP:- False Positives, FN:- False Negatives.

[b] Best Model:- Comprises of top 28 structural and sequence features (subsequence length 35).

The window length 35 is giving good performance signifies that there are long term interactions that play a very important role in the Glycosylation process. It turns out that the in the sequence and the structural features *contact number*, *contact order*, *solvent accessibility*, *composition descriptor of hydrophobic amino acids* and *Geary autocorrelation descriptors*[10] around the glycosylated site play an important role in determining the occupancy of the sequon the list of top 10 features is given in Table 2.

The top scoring features of this model agree well with a comprehensive study that does the statistical analysis of the N-glycosylated sites[5,8]. There have been different significant approaches for the identification of N-Linked Glycosylated sites by Gupta et al. [6] who reported 76% accuracy for the identification of N-Glycosylation sites in humans and Caragea et al. [7] who employed ensembles of Support Vector Machines for the prediction of Glycosylation sites (N-linked,O-linked and C-mannosylation) in this work the accuracy for N-linked

Table 2. Top 10 features from the model incorporating Sequence and Structural Features

Feature
Contact Number
Contact Order
Accessibility at position 10
Composition descriptor of Hydrophobic Amino Acids (EDRK)
Accessibility at position 12
Composition descriptor of Pi Helix in the DSSP secondary structure assignment
Torsion angle of residue 11 in the backbone of sequon
GAC [a] for lag of 26 for Optimized propensity to form reverse turn (Oobatake et al., 1985)
GAC [a] for lag of 26 for Radius of gyration of side chain (Levitt, 1976)
Torsion angle of residue 22 in the backbone of sequon

[a] GAC:- Geary Autocorrelation Descriptor.

glycosylation with Single SVM is 94%, but contains less number of experimentally verified N-Glycosylation sites thus, a rational comparison of our approach with these methodologies is not possible owing to the difference in the datasets and methodologies.

4 Conclusion

The aim of this study is to provide a reliable model to predict the N-linked glycosylation of sequons employing Random Forest classifier. Apart from classification, Random Forest also provides heuristic information in terms of a ranking score. This facilitates us to select the top ranking features to obtain optimal classification performance. The latter provides valuable domain knowledge regarding the N-Glycosylated sites. Finally, the model with sequence and structural features has a subsequence length of 35 and a MCC of 0.4426 and accuracy 72.81 with sensitivity as 0.7165072 and specificity as 0.7356771. This type of study provides a high throughput screening mechanism in experiments where the role of N-Glycosylation is being investigated in some functional aspect, in this case this model will provide reliable first hand results even before the N-Glycosylation is experimentally verified.

Acknowledgements. VKJ gratefully acknowledes financial assistance from Department of Science and Technology New Delhi, INDIA. VKJ also acknowledges fruitful suggestions by Dr. Petety V. Balaji, IIT, Mumbai, INDIA.

References

1. Drickamer, K., Taylor, M.E.: Biology of animal lectins. Annual Review of Cell Biology 9(1), 237–264 (1993) PMID: 8280461
2. Lis, H., Sharon, N.: Lectins: Carbohydrate-specific proteins that mediate cellular recognition. Chemical Reviews 98(2), 637–674 (1998)

3. Crocker, P.R.: Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell-cell interactions and signalling. Curr. Opin. Struct. Biol. 12(5), 609–615 (2002)
4. Gavel, Y., Heijne, G.v.: Sequence differences between glycosylated and non- glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. Protein Eng. 3(5), 433–442 (1990)
5. Petrescu, A.J., Milac, A.L., Petrescu, S.M., Dwek, R.A., Wormald, M.R.: Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. Glycobiology 14(2), 103–114 (2004)
6. Gupta, R., Jung, E., Brunak, S.: Netnglyc 1.0 server (Unpublished)
7. Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., Honavar, V.: Glycosylation site prediction using ensembles of support vector machine classifiers. BMC Bioinformatics 8, 438–438 (2007)
8. Ben-Dor, S., Esterman, N., Rubin, E., Sharon, N.: Biases and complex patterns in the residues flanking protein N-glycosylation sites. Glycobiology 14(2), 95–101 (2004)
9. Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., Abola, E.E.: Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr. D. Biol. Crystallogr. 54, 1078–1084 (1998)
10. Li, Z.R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X., Chen, Y.Z.: PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucl. Acids Res. 34, W32–W37 (2006)
11. Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12), 2577–2637 (1983)
12. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M.: AAindex: amino acid index database, progress report. Nucl. Acids Res. 36, D202–D205 (2008)
13. Breiman, L.: Random forests. Machine Learning, 5–32 (2001)
14. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
15. Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., Van Eerdewegh, P.: Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 28(2), 171–182 (2005)
16. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1), 3 (2006)
17. Hamby, S., Hirst, J.: Prediction of glycosylation sites using random forests. BMC Bioinformatics 9, 500 (2008)
18. Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., Zhao, H.: Pathway analysis using random forests classification and regression. Bioinformatics (2006)
19. R Development Core Team: R: A Language and Environment for Statistical Computing. In: R. Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0
20. Liaw, A., Wiener, M.: Classification and regression by randomforest. R. News 2(3), 18–22 (2002)

Identification of Defensins Employing Recurrence Quantification Analysis and Random Forest Classifiers

Shreyas Karnik[1,3], Ajay Prasad[1], Alok Diwevedi[1], V. Sundararajan[2,⋆], and V.K. Jayaraman[2,⋆]

[1] Chemical Engineering and Process Development Division, National Chemical Laboratory, Pune, India - 411008
[2] Center for Development of Advanced Computing, Pune University Campus, Pune, India - 411007
vsundar@cdac.in, jayaramanv@cdac.in
[3] School of Informatics, Indiana University, Indianapolis, IN, USA, 46202

Abstract. Defensins represent a class of antimicrobial peptides synthesized in the body acting against various microbes. In this paper we study defensins using a non-linear signal analysis method Recurrence Quantication Analysis (RQA). We used the descriptors calculated employing RQA for the classification of defensins with Random Forest Classifier.The RQA descriptors were able to capture patterns peculiar to defensins leading to an accuracy rate of 78.12% using 10-fold cross validation.

1 Introduction

Defensins are a class of antimicrobial peptides usually rich in cystine residues. They can be defined as a family of potent antibiotics synthesized within the body by neutrophils (a type of white blood cell) and macrophages (cells that can engulf foreign particles). Defensins act against bacteria, viruses and fungi by binding to their membranes and kill cells by forming voltage regulated polymeric channels in the susceptible cell's membrane[1]. The defensins are classified into alpha-defensins, beta-defensins and theta-defensins on the basis of their sequence homology. Defensins, owing to their small and potent antimicrobial effects can be used effectively for development of new clinically applicable antibiotics. They are also known to harbor anti-tumor activity, mutagen activity and/or behave as signaling molecules [1]. A few of their characteristics that have made them preferred candidates as peptide drugs are their short length, fast and efficient action against microbes and low toxicity to mammals [1]. Protein sequences can be visualized as linear heteropolymers that are formed by non-periodic arrangement of 20 different amino acids. In spite of the non-linear arrangement of amino acids in a protein, biologically stable proteins and functional are formed only with certain arrangements of amino acids. Literature sources indicate that some of the amino acids are preserved during the

⋆ Corresponding authors.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 152–157, 2009.

course of evolution, the conservation of these sequences points that the signals for a particular functionality of proteins are preserved in the sequence. Advanced techniques that can be applicable for investigating protein structure functional relationship like Fourier analysis, Wavelet analysis, chaos based methods can explore the underlying correlations within the protein sequences. Proteins can be viewed as a time series, where the order of amino acids indicates the role of time. There are examples of protein sequence analysis using signal processing based techniques [2]. Techniques like recurrence plots are used to visualize time series in a meaningful manner and give idea about the nature of the time series. Recurrence quantification analysis (RQA) has earlier been applied to study structure activity relationships in various proteins[3,4].and for other protein science and engineering applications providing interesting insights[5].

In this work we utilize the recurrence property to study the protein sequences that are characterized as defensins which play important role in innate immunity. Further, we employ Random Forests as the machine learning algorithm to classify sequence as defensin based on the RQA descriptors. Literature sources[3] indicate that hydrophobic patterns govern the native structure of proteins, also influencing the activity of proteins so we decided to use the Kyte-Doolittle Hydrophobicity scale for the conversion of the protein sequences into numerical time series to calculate RQA descriptors.

In the following sections brief description of RQA and Random Forest will be given along with the methodology followed by the results and discussion.

2 Methodology

2.1 Dataset

Literature on defensins was searched using search engines like Pubmed, iHop, Google Scholar and HubMed. Uniprot was also queried in order to get defensin sequences. A dataset of 238 non-redundant sequences which were annotated as defensins was complied. The defensin data set constituted the positive training examples, for the negative examples Uniprot was randomly sampled for sequences which are not annotated as defensins having length less than 150 amino acids containing cystine residues.

2.2 Recurrence Quantification Analysis (RQA)

The concept of Recurrence Plots was introduced by Eckmann et al.[6] Recurrence Plots are used as a tool that can enable the visualization of n dimensional phase space trajectory as a 2-D or 3-D plot of the recurrenences. The notion of recurrence is based on two facts viz. similar situations evolve in a similar way and some situations occur quite often. Thus, the recurrent points represent a state at position i which recurs at position j. Both the axes of the recurrence plots represent the time series that is under consideration. Recurrence plots are based on the recurrence matrix which is calculated as follows:

$$R_{i,j} = \Theta(\epsilon - ||\boldsymbol{x_i} - \boldsymbol{x_j}||) \quad i,j = 1 \cdots N \tag{1}$$

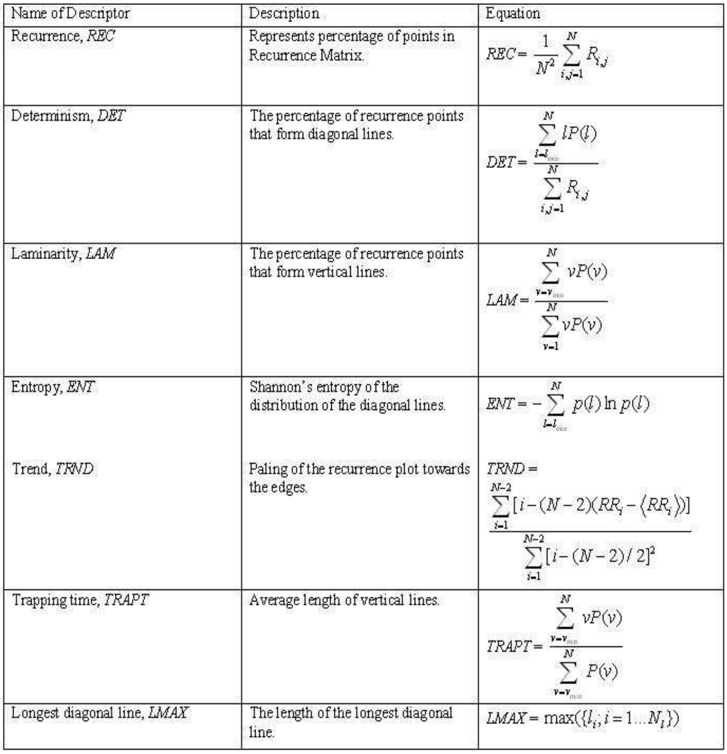
Where N is the number of states considered for analysis, ϵ is a threshold distance, $|| \; ||$ is the norm, and Θ is the Heavyside function. For given point to be recurrent, it must fall within the threshold distance, ϵ. The recurrence plot inspection gives us the idea about correlation in the states at higher dimensions visually. The quantification of the recurrence plots [7] provides a number of parameters that describe small scale structures in the plot such as dots, diagonal and horizontal lines. Following parameters obtained by RQA quantification reveal the small scale structures in the recurrence matrix:

– *REC*: Quantification parameter *REC* is a count of single isolated points in the plot which signify that these states are rare, and do not fluctuate or persist consistently. These points do not necessarily imply noise in the data.
– *DET*: Certain points in the time series are parallel to each other i.e. they have same values but are placed at a different interval. This is expressed mathematically as: $\mathrm{R}_{i+k,j+k} = 1$ (for k,$\cdots$,l; where l is the length of the stretch). These stretches can be visualized as diagonal lines in the recurrence matrix. These diagonal lines are called *deterministic*, because they represent a deterministic pattern in the series. *DET* is the count of all diagonal lines in a recurrence matrix. *DET*, thus, quantifies the deterministic local evolution of states in the series.
– *LAM*: Quantification parameter *LAM* represents the states which change or changes very slowly and as vertical lines in Recurrence Plot. Mathematically, this is $\mathrm{R}_{i,j+k} = 1$ (for k 1$\cdots$,v, where v is the length of the vertical line). A vertical line represents a state in which the series is trapped; and *LAM* is a count of all vertical lines in a matrix.

In addition to the above mentioned descriptors additional descriptors were suggested by Webber et. al [7] in order to describe the recurrence plots. These include *ENT* and *MAXLINE*, which are the Shannon's Entropy of the distribution of the diagonal lines and the maximum length of a diagonal line, respectively, and *TRAPT*, which is the average length of vertical lines. Complete list of the descriptors is given in Figure 1:

3 Classification Algorithms

Random Forest is an algorithm for classification and regression developed by Leo Breiman[8] that uses an ensemble of trees for classification and regression. Each of the classification trees is built using a bootstrap sample of the data, and at each split the candidate set of variables is a random subset of the variables (*mtry*). Thus, Random Forest uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners, and random variable selection for tree building. Each tree is unpruned (grown fully). The trees grown then vote on the class for a given input, the class that gets the maximum number of votes is then assigned to the input. Random forest exhibits excellent performance in classification tasks. It has been successfully employed to solve various problems from the life sciences domain[9,10,11]. It has several characteristics that make it



Name of Descriptor	Description	Equation
Recurrence, *REC*	Represents percentage of points in Recurrence Matrix.	$REC = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}$
Determinism, *DET*	The percentage of recurrence points that form diagonal lines.	$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{i,j=1}^{N} R_{i,j}}$
Laminarity, *LAM*	The percentage of recurrence points that form vertical lines.	$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)}$
Entropy, *ENT*	Shannon's entropy of the distribution of the diagonal lines.	$ENT = -\sum_{l=l_{min}}^{N} p(l) \ln p(l)$
Trend, *TRND*	Paling of the recurrence plot towards the edges.	$TRND = \frac{\sum_{i=1}^{N-2} [i-(N-2)(RR_i - \langle RR_i \rangle)]}{\sum_{i=1}^{N-2} [i-(N-2)/2]^2}$
Trapping time, *TRAPT*	Average length of vertical lines.	$TRAPT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)}$
Longest diagonal line, *LMAX*	The length of the longest diagonal line.	$LMAX = \max(\{l_i; i = 1 \ldots N_l\})$

Fig. 1. RQA Descriptors

ideal for these data sets. We used the Random Forest implementation in R[12] by Andy Liaw and Matthew Wiener[13] based on original code in FORTRAN by Breiman. Following methodology was adopted for classification:

1. All the protein sequences were converted to their numerical equivalents using Kyte-Doolittle Hydrophobicity scale.
2. Protein sequences were partitioned into train (80%) and test (20%) splits. The test split was used for the evaluation purpose.
3. Individual sequence which is now represented as a time series was subjected to RQA using parameters embedding dimension as 3, delay as 1, and radius as 6 using programs from Webber et al§.
4. RQA descriptors were calculated for all the protein sequences in the train and test splits of the data.
5. A model for classification was built using RQA descriptors and Random Forest algorithm(using best *mtry* parameter)on the training data and the model was evaluated using 10-fold cross validation to check the robustness. The performance of the model on test data was also evaluated.

[1] §http://homepages.luc.edu/∼cwebber/RQA131.EXE

Table 1. Summary of Performance on Test Set

Algorithm	Sensitivity	Specificity	MCC	Accuracy
Random Forest (with *mtry =3*)	0.736	0.812	0.588	79.16%

4 Results and Discussion

Performance of the model was evaluated on the basis of the cross validation accuracy and the performance on the test set in terms of Mathew‘s Correlation Coefficient(MCC) and other standard metrics used for evaluation of classification performance. The 10-fold cross validation accuracy of the model is 78.2%. The results on the test set are given in Table 1. Random Forest algorithm returns the variables that are most important in classification. We tried eliminating the features that have a low importance but it did not improve the classification performance thus only a list of ranking is presented. The order of importance of the RQA descriptors used for classification of defensins is as follows:

1. Recurrence *REC*
2. Trend *TRND*
3. Determinism *DET*
4. Laminarity *LAM*
5. Entropy *ENT*
6. L_{max}
7. Trapping Time *TRAPT*
8. V_{max}

Recurrence being the most important feature suggests that single isolated points representing rare states which do not fluctuate or persist consistently are important signals in that discriminate defensins. RQA descriptors such as *TRND*, *DET*, *LAM*, and *ENT* are also amongst the important features in terms of classification of defensins. In current study, classification of proteins as defensins based on features extracted from RQA could be achieved with 78.2% accuracy (cross validation accuracy)this suggests that RQA based on the representation of proteins in terms of their numeric equivalent (Kyte-Doolittle Hydrophobicity Scale in this work) captures the essential signals characteristic of defensins and can hence be used as an effective tool for exploring sequence function relationships and classification.

Acknowledgements. VKJ gratefully acknowledes financial assistance from Department of Science and Technology New Delhi, INDIA.

References

1. Ganz, T.: Defensins: antimicrobial peptides of vertebrates. Comptes Rendus Biologies 327(6), 539–549 (2004)
2. Giuliani, A., Benigni, R., Sirabella, P., Zbilut, J.P., Colosimo, A.: Nonlinear methods in the analysis of protein sequences: A case study in rubredoxins. Biophysics Journal 78(1), 136–149 (2000)

3. Zbilut, J.P., Giuliani, A., Webber, C.L.J., Colosimo, A.: Recurrence quantification analysis in structure-function relationships of proteins: an overview of a general methodology applied to the case of tem-1 beta-lactamase. Protein Eng. 11(2), 87–93 (1998)
4. Angadi, S., Kulkarni, A.: Nonlinear signal analysis to understand the dynamics of the protein sequences. The European Physical Journal - Special Topics 164(1), 141–155 (2008)
5. Mitra, J., Mundra, P.K., Kulkarni, B.D., Jayaraman, V.K.: Using recurrence quantification analysis descriptors for protein sequence classification with support vector machines. Journal of Biomolecular Structure and Dynamics 25(3), 141 (2007)
6. Eckmann, J.P., Kamphorst, S.O., Ruelle, D.: Recurrence plots of dynamical systems. EPL (Europhysics Letters) (9), 973 (1987)
7. Webber Jr., C.L., Zbilut, J.P.: Dynamical assessment of physiological systems and states using recurrence plot strategies. J. Appl. Physiol. 76(2), 965–973 (1994)
8. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
9. Diaz-Uriarte, R., Alvarez de Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1), 3 (2006)
10. Hamby, S., Hirst, J.: Prediction of glycosylation sites using random forests. BMC Bioinformatics 9, 500 (2008)
11. Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., Zhao, H.: Pathway analysis using random forests classification and regression. Bioinformatics (2006)
12. R Development Core Team: R: A Language and Environment for Statistical Computing. In: R. Foundation for Statistical Computing, Vienna, Austria (2009) ISBN 3-900051-07-0
13. Liaw, A., Wiener, M.: Classification and regression by randomforest. R. News 2(3), 18–22 (2002)

Data Mining by Navigation – An Experience with Systems Biology*

Amarnath Gupta[1], Michael Baitaluk[1], Animesh Ray[2], and Aditya Bagchi[3]

[1] San Diego Supercomputer Center, Univ. of California San Diego, La Jolla,
CA 92093, USA
{gupta,baitaluk}@sdsc.edu
[2] Keck Graduate Institute, 435 Watson Dr., Claremont, CA 91711, USA
Animesh.Ray@kgi.edu
[3] Indian Statistical Institute, Kolkata 700108, India
aditya@isical.ac.in

Abstract. This paper proposes a navigational method for mining by collecting evidences from diverse data sources. Since the representation method and even semantics of data elements differ widely from one data source to the other, consolidation of data under a single platform doesn't become cost effective. Instead, this paper has proposed a method of mining in steps where knowledge gathered in one step or from one data source is transferred to the next step or next data source exploiting a distributed environment. This incremental mining process ultimately helps in arriving at the desired result. The entire work has been done in the domain of systems biology. Indication has been given how this process can be followed in other application areas as well.

1 Introduction

In order to get insights of hypertension mechanisms, this paper ventures to discover the genes responsible for such hypertension. Traditionally the systems biologists depend on past experience, augmented by new laboratory experiments and statistical analysis. Results help in forming hypothesis substantiated by expert opinions. Though this approach enriches knowledge, it hardly utilizes past data obtained from different biological experiments. These past data are well documented in different databases and gene transcriptional/signaling networks, represented as graphs. This paper tries to make the knowledge discovery process sufficiently data driven avoiding dependence on hypotheses formed from experiences and expert opinions.

Since this type of knowledge discovery demands access to diverse data sources, it falls under distributed data mining paradigm. Moreover this study tries to associate related data items from different sources without deriving any rules. So the statistical processes applied here are different from the traditional *support* and *confidence* measures that use joint and conditional probabilities [7]. In addition, unlike considering only co-occurrence of items, combination of presence and absence of items may also

* This work was funded by the NIH/NIGMS grant R01 GM084881, which also paid for attending the PReMI 2009 Conference.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 158–164, 2009.

be studied to discover additional interesting patterns. A new measure to this effect has already been developed [5]. Some applications demand collection of data from different sources in order to find composite association of them. This notion of mining, popularly known as generalized association, tries to identify correlations among items considering both the absence and presence of them. A seminal work with this approach used chi-square test to measure significance of association. Authors demonstrated effectiveness of their method by testing it on census data and also by finding term dependence in a corpus of text documents [6]. Another recent work has proposed a method of compositional mining on multiple biological datasets. Authors have also proposed two new methods of mining; redescription and biclustering. However, here all the datasets are collected together to create a multirelational environment [9].

Though earlier research efforts explored the possibility of discovering association among items by extracting data from different sources, these efforts primarily handled same type of data representation and hence creation of a composite scenario was not a computationally challenging task. However, the application considered in this paper ventures to discover association among items with diverse representations distributed over different sites of a network. So this problem demands a new approach to mining different from the earlier efforts.

First, data representation, storage structure, access methods etc. are different in different data sources. So, porting data from different sites to one site and then bringing everything under a uniform representation may not be cost effective.

Secondly, semantic interpretation of even common items may be substantially different under different paradigms. For example, in a bioinformatics application, a gene may sometimes be viewed as a sequence and sometimes be viewed as a node in an interaction graph.

So, a mining effort is required which would choose to break the computational process to each data source and would transfer result of one to the next for further processing. In this navigational process a data source may be visited more than once depending on the application domain requirements. Thus the main motivation behind this paper is to develop a method of mining by navigation in a distributed environment.

While Section 2 of the paper covers the background knowledge for the data mining problem, Section 3 discusses about the diverse data sources utilized. Actual computational and navigational methodologies have been covered in Section 4. Section 5 discusses about the contribution made by the paper and concludes with possibility of other interesting applications.

2 Background Knowledge

Before going into the algorithmic detail, the following paragraph discusses about the problem domain for the appreciation of computer science community.

In general, Deoxyribonucleic Acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. DNA segments that carry this genetic information are called genes. DNA also contains the instructions or blueprints needed to construct other components of cells like, proteins and RNA molecules. Ribonucleic acid (RNA) is transcribed from DNA by enzymes RNA polymerases. RNA is central to the synthesis of proteins. Here, a type of

RNA called messenger RNA (mRNA) carries information from DNA and translates the information they carry into proteins. So in short, genes transcribe to mRNA in the process of protein production. Gene expression refers to the amount of mRNA produced. [8]. Two genes are considered to be co-expressed if, against a real life phenomenon, like a disease, both the genes are expressed simultaneously. Conversely, against a real life phenomenon, two genes are differentially expressed if while one gene is expressed other does not. Transcription of a gene is regulated (i.e. enabled and initiated) by a protein. While there can be other types of regulators, a regulator that helps in transcription is called a Transcription Factor (TF). The common expert opinion or hypothesis is - “if two genes are co-expressed, they may have a common TF”.

3 Data Sources

In this paper, the proposed mining effort uses three main data sources:

1. Gene Expression Omnibus (GEO), a well known repository that acts as curated, online resource for browsing, query and retrieval of gene expression data [4].
2. Promoter Analysis Pipeline (PAP) [2] is a Web-based workbench that provides the analysis of a set of co-expressed genes and the prediction of their transcriptional regulatory mechanisms.
   - PAP provides possible transcription factors that may regulate a set of co-expressed genes.
   - PAP can also predict other genes in the genome that are likely to be regulated by the same set of transcription factors.
   - It is also possible to get from PAP, the set of transcription factors that may regulate a particular gene.

   All these information are statistical in nature and offered with a possible likelihood. The results are accepted when they cross a certain threshold of likelihood.
3. PathSys is a graph-based system for creating a combined database of biological pathways, gene regulatory networks and protein interaction maps. PathSys integrates over 14 curated and publicly contributed data sources along with Gene Ontology and is structured as an acyclic graph [1].

4 Computation and Navigation

Starting with an experiment to find the genes responsible for hypertension mechanism in mammalian species, this paper ventures to navigate over the different data sources mentioned in Section 4. Ultimately, a set of interacting genes with their transcription factors are discovered from the graphs of PathSys. This graph is enriched by marking the connecting edges with their relative importance as obtained from computation with data from other sources. In the process, biologists acquire such knowledge about gene interactions that was not possible to get from laboratory experiments alone. The initial experiment was done on a hypertensive-hypotensive mice population to identify a set of differentially over-expressed genes apparently responsible for hypertension [3]. In the rest of the paper this would be referred as Friese experiment. Since the computational process has to navigate among diverse data sources it is necessary to justify the order of navigation so that the overall process of mining is well appreciated.

Genes identified by the Friese experiment, if placed on the PathSys graph directly as nodes, would give rise to a sub-graph as the molecular interaction graph that is responsible for hypertension in mammalian species. However, it does not reveal the importance of co-occurrence of all these genes. So it is necessary to identify which of these genes contribute significantly in the hypertension process when they co-express. This information is available in GEO database. So without pruning the gene set with co-expression data, directly accepting a sub-graph from PathSys may offer nodes and edges that may not have significant contribution to the study undertaken. Now since the co-expressed genes are likely to have common Transcription Factors and the genes that tend to co-express under hypertension have already been identified, the corresponding TFs can be obtained from the PAP workbench. Pruned set of genes as obtained from GEO database along with their TFs as obtained from PAP may now be placed on the PathSys graph. Experiment below would show that this process of identifying co-expressed genes and identification of sub-graph may be done more than once, ultimately to arrive at a final sub-graph.

Different steps of computation and navigation among data sources are detailed below:

Step 1. Correlation Phase:

Data Source: GEO Database.

Starting with the initial set of genes obtained from Friese experiment the GEO database is searched for experiments involving those genes, where in each such experiment at least one pair of co-expressed genes is identified from the initial set.

For each such pair of co-expressed genes Xi and Yi, the expression values for all experiments are taken where they co-occur and Pearson correlation coefficient is computed as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y}$$

where, S_X and S_Y are the standard deviation of X and Y respectively (with 25 genes in the initial set, GEO database offered data about 630 experiments, causing a maximum data volume = 630 * $({}^{25}C_2)$).

Out of the correlation coefficients computed for all pair of genes in the initial set only those crossing a predefined threshold are taken for further study (considering a threshold of 0.7 for Pearson correlation coefficient, similar to Minimum-Support, 11 genes, as listed in Table 1, are retained for further processing).

Table 1. Final set of co-expressed genes obtained from GEO

Gene Name
ACADVL (Acyl CoA dehydrogenase, very long chain)
CHGA (Chromogranin A)
CHGB (Chromogranin B)
CYP2E1 (Cytochrome P450 2E1)
EDN3 (Endothelin 3)
IGFBP4 (Insulin like growth factor binding protein-4)
IGFBP6 (Insulin like growth factor binding protein-6)
PNMT (Phenyl ethanolamine N-methyl transferase)
SCG2 (Chromogranin C)
SLC6A4 (Serotonin transporter)
TH (Tyrosine hydroxylase)

Step 2. Extraction of Transcription Factors (First Navigation):

Data Source: PAP Workbench.

Pairs of genes found to be highly co-expressed in Step 1 (correlation phase), are analyzed in the PAP workbench to find possible common Transcription Factors (TF). Considering human, rat and mouse data, only gene pairs present in all the three mammals are taken.

Since all TFs are identified with corresponding likelihood value, only TFs crossing a predefined threshold are listed in descending order and only a predefined number of members from the top of the list are taken for the subsequent step of analysis (the specific study considered 95% likelihood and only top 25 TFs are taken. These TFs are the regulator proteins for the genes obtained in Step 1).

Step 3. Identification of Protein-Protein Interaction (Second Navigation):

Data Source: PathSys Molecular Interaction Maps.

Since PathSys offers protein interaction maps, nodes corresponding to the original set of genes obtained from Step 1 and the nodes corresponding to the TFs obtained in Step 2 are identified along with their interconnecting edges. This study offers a sub-graph with the interconnected proteins and genes that is responsible for hypertension in mammalian species.

To extend the study further, one-neighborhood of these nodes is now considered on the PathSys graph thereby identifying even the regulator nodes connecting the TFs.

Step 4. Second Correlation phase (Third Navigation):

Data Source: GEO Database.

In order to ascertain the importance of the regulators in the original study of hypertension, the TF nodes as well as their regulators obtained from one-neighborhood are again taken for study of co-expression in GEO database. Repeating the process described in Step 1, the minimized set of highly co-expressed TFs and their regulators are identified.

Step 5. Graph Pruning (Fourth Navigation):

Data Source: PathSys Molecular Interaction Maps.

Sub-graph identified in Step 3 after considering one-neighborhood is now pruned with the result obtained from Step 4 and the extra nodes and edges obtained from one-neighborhood are deleted. Each co-expression edge (gene to gene and TF to TF) is now labeled with its co-expression value. This pruned graph with all its nodes as genes, proteins etc. and the interconnecting edges is functionally related to hypertension in mammalian species.

Step 6. Further Graph Pruning: **(An additional study)**

Hypothesis: *If the network, as obtained from Step 5 is functionally significant, then co-expressed and interacting genes and their regulator proteins should not only show functional relation to hypertension but also to other related diseases like cardiovascular disorders and diabetics.*

Accepting the hypothesis obtained from the domain experts, the graph pattern obtained from Step 5 is further pruned by using a fourth data source, the Online Mendelian Inheritance in Man or OMIM database. This database is a catalog of human genes and genetic disorders. Besides the original set of 11 genes, corresponding entries for

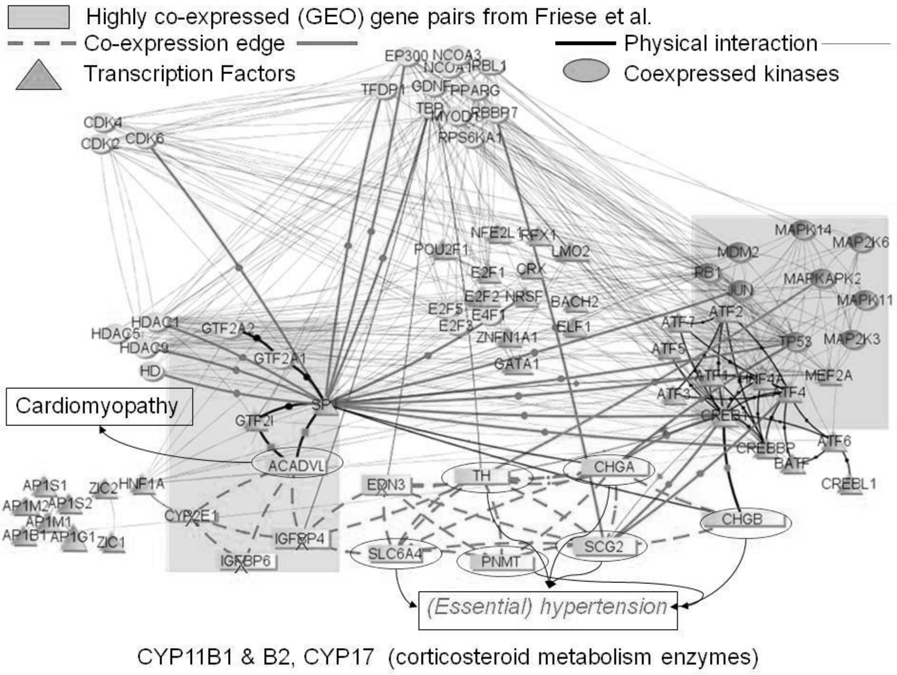


Fig. 1. PathSys Interaction Graph

other nodes are searched in OMIM database to ascertain whether they contribute to any or all of the diseases under scrutiny. For example, HNF1A a transcription factor for CYP2E1 gene is also a TF for other genes responsible for liver-specific disorders, and HDAC5 a regulator of SP1, a member of original set of TFs, is associated with cardiac problems.

Collecting these extra evidences from the fourth dataset used in this study, the set of selected genes and TFs is further pruned to get the final gene-protein interaction graph that contributes to the study of hypertension in mammalian species. Figure 1 shows the final pruned graph as obtained from PathSys.

5 Conclusions

Studying a problem in systems biology, this paper has shown how composite mining strategy can be devised when related information are distributed in different data sources at different sites with substantial variation in data representation, data structures, storage strategies etc. Besides this navigational process, this paper has also proposed a new graph mining technique. Deviating from well known graph mining strategies, this process wants to identify the most interesting sub-graph from a large graph and augments it by evidences obtained from other problem related data sources. Interestingly, it may be possible to apply this graph pruning strategy in many other application areas like, traffic management, tour planning, social network etc. Authors are now trying to develop a generalized methodology for this type of graph mining.

References

1. Baitaluk, M., Qian, X., Godbole, S., Raval, A., Ray, A., Gupta, A.: PathSys: Integrating Molecular Interaction Graphs for Systems Biology. BMC Bioinformatics 7, 55 (2006), `http://www.biomedcentral.com/1471-2105/7/55`
2. Chang, L.W., Fontaine, B.R., Stormo, G.D., Nagarajan, R.: PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis. Nucleic Acids Research, 238–244 (2007), `http://bioinformatics.wustl.edu/webTools/PromoterAnalysis.do`
3. Friese, R.S., Mahboubi, P., Mahapatra, N.R., Mahata, S.K., Schork, N.J., Schmid-Schönbein, G.W., O'Connor, D.T.: Common genetic mechanisms of blood pressure elevation in two independent rodent models of human essential hypertension. Am. J. Hypertension 18(5 Pt 1), 633–652 (2005)
4. Gene Expression Omnibus (GEO), `http://www.ncbi.nlm.nih.gov/geo/`
5. Pal, S., Bagchi, A.: Association against Dissociation: some pragmatic considerations for Frequent Itemset generation under Fixed and Variable Thresholds. ACM SIGKDD Explorations 7(2), 151–159 (2005)
6. Silverstein, C., Motwani, R., Brin, S.: Beyond Market Baskets: Generalizing Association Rules to Correlations. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 265–276 (1997)
7. Srikant, R.: Fast Algorithms for Mining Association Rules and Sequential Patterns, Ph.D. Thesis, University of Wisconsin, Madision, USA (1996)
8. Wikipedia, the free online encyclopedia, `http://www.wikipedia.org/`
9. Ying, J., Murali, T.M., Ramakrishnan, N.: Compositional Mining of Multirelational Biological Datasets. ACM TKDD 2(1), 1–35 (2008)

A Least Squares Fitting-Based Modeling of Gene Regulatory Sub-networks

Ranajit Das[1], Sushmita Mitra[1], C.A. Murthy[1], and Subhasis Mukhopadhyay[2]

[1]Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
{ranajit_r,sushmita,murthy}@isical.ac.in
[2]Department of Bio-Physics, Molecular Biology and Bioinformatics, Calcutta University, Kolkata 700 009, India
sm.bmbg@gmail.com

Abstract. This paper presents a simple and novel least squares fitting-based modeling approach for the extraction simple gene regulatory sub-networks from biclusters in microarray time series gene expression data. Preprocessing helps in retaining the strongly interacting gene regulatory pairs. The methodology was applied to public-domain data sets of Yeast and the experimental results were biologically validated based on standard databases and information from literature.

Keywords: Biclustering, transcriptional regulatory network, least squares, gene interaction network.

1 Introduction

During the recent years, rapid development in DNA microarray technology have resulted in the parallel generation of expression data of thousand of genes, of various organisms, under several experimental conditions. Genome expression profiling of many organisms have been completed in the past few years. The latest Affymetrix gene chips accommodate 750,000 unique 25-mer oligonucleotide features constituting more than 28,000 mouse gene-level probe sets. It is known that mRNA profiles are prone to different kinds of noise and ambiguity, and may be unequally sampled over time. Time series gene expression data is also essentially under-determined, involving high-dimensional genes with very few time-points. Clustering is one way of estimating such noisy expression data, by grouping co-expressed genes with the assumption that they are co-regulated. However, it is observed that a subset of genes is co-regulated and co-expressed over a subset of experimental conditions only. Biclustering (or co-clustering) aims at bringing out such local structure inherent in the gene expression data matrix. It refers to some sort of feature selection and clustering in the space of reduced dimension, at the same time [1].

To quantify the similarity among the co-expressed genes in a bicluster several distance measures have been employed. However, it is to be noted that any apparent similarity of expression profiles between a pair of genes need not always signify direct regulation. It may denote an indirect coregulation by other genes,

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 165–170, 2009.

or it may also be due to a mere coincidence involving no causal relationship. The indirect interaction between two genes may result from the regulation mediated by proteins, metabolites and non-coding RNAs (ncRNAs). Transcription factor (TF) is a protein that interacts directly with its target gene(s) (T) by up regulating (down regulating) its gene expression – resulting in activation (inhibition) of the target. There may also exist regulatory cascades (of first-order interactions between gene pairs), whereby the product of one gene influences the transcription rate of the second one, and so on [1].

In this paper we propose the method of least squares fitting using polynomials in the framework of continuous-column multiobjective evolutionary biclustering [2] to extract the interaction between gene pairs. Preprocessing, involving the discretization of the error matrix (using quantile partitioning) and the subsequent elimination of links with higher errors, is employed to retain strongly regulated gene pairs. An adjacency matrix is formed from the resulting error matrix, based on which the regulatory network is generated and biologically validated. The usefulness of the model is demonstrated, using time-series gene expression data from Yeast.

2 Reverse Engineering of Gene Regulatory Sub-networks

The various properties of the genome, along with the expression of a gene (which is the amount of mRNA it produces) are addressed in an important group of biological networks known as the genetic regulatory network (GRN). A GRN comprises of a complicated structure involving different gene products that activate or inhibit other gene products [1]. The Multi-objective evolutionary algorithm (MOEA), in association with the local search, were used for the generation of the set of biclusters. The algorithm followed is discussed in details in [2].

2.1 Algorithm for Extraction of Gene Gene Regulatory Sub-networks

The main steps of the proposed algorithm are outlined as follows.

1. Extraction of biclusters by MOEA.
2. Computation of pairwise error for least squares fitting between gene pairs.
3. Discretization of the error matrix for eliminating the higher errors.
4. Network generation from the connectivity matrix.
5. Biological validation.

2.2 Least Squares Fitting of Gene Pairs

The huge size of the gene expression data and the associated combinatorial problems in high-dimensional space, have opened up a challenge to traditional techniques. It emphasizes the need for dimensionality reduction in this context. The major problem with DNA microarray data analysis is that the data is essentially under-determined, with very few samples (or time-points) compared to the high-dimensional attributes/features (or genes) to be estimated; thus creating an additional constraint. Although standard statistical techniques for extracting relationships have proposed multiple models to fit the data [3], they often require additional data to resolve the ambiguities. These make the regression problem an ill-posed one; therefore a regularization technique is desired to be developed.

In this paper we propose a least squares fitting-based approach for the reconstruction of interactions in gene regulatory networks. Commonly used similarity measures like the Euclidean distance, Pearson correlation or the Spearman's rank correlation do not take into consideration the non-linear effects among genes. The above distance measures can serve as a satisfactory measure of relationship between two genes only when the two are linearly related. A low value of the correlation coefficient also does not rule out the possibility that the genes may be related in some other manner. Again, the fact that the coefficient of correlation between two genes is higher does not necessarily mean that they are causally related. Filkov *et al.* have designed an edge detection function for identifying regulation of genes, which demonstrates that less than 20% of the known regulatory gene pairs exhibit strong correlations [4].

The least-squares method, a very simple form of the regularization technique, can be helpful for model selection. Using the method of least squares we minimize the sum of squares of the error of estimation (S^2) of fitting one gene to another. If the computed error for fitting a particular gene $G1$ with another gene $G2$ (say), within the reduced localized domain of biclusters, be less than that for fitting $G2$ with $G1$ then we infer that $G1$ affects $G2$ and vice versa. The relationship is represented in terms of rules, linking the gene which regulates, to the regulated gene.

A gene expression profile e is represented over a series of n time points. Let S_1^2 denote the residual or the sum of squares of the error of estimation for fitting gene e_1 to e_2 (sampled at e_{1i} and e_{2i} over n time intervals). This is given by

$$S_1^2 \equiv \frac{1}{n}[\sum_i \{e_{2i} - (a_0 + a_1 e_{1i} + a_2 e_{1i}^2 + \ldots + a_k e_{1i}^k)\}^2]. \tag{1}$$

where a_k's denote the k coefficients of the kth order polynomial fitted. Analogously, the residual for fitting e_2 to e_1 can be represented as

$$S_2^2 \equiv \frac{1}{n}[\sum_i \{e_{1i} - (b_0 + b_1 e_{2i} + b_2 e_{2i}^2 + \ldots + b_k e_{2i}^k)\}^2]. \tag{2}$$

The coefficients, a_k's and b_k's, are obtained by solving the corresponding k normal equations

$$\frac{\partial S_1^2}{\partial a_k} = 0; \tag{3}$$

and

$$\frac{\partial S_2^2}{\partial b_k} = 0. \tag{4}$$

The corresponding errors $\xi(e_1, e_2)$ are represented as $\xi_1(e_1, e_2) = \sqrt{S_1^2}$ and $\xi_2(e_2, e_1) = \sqrt{S_2^2}$, which again are evaluated using eqns. (1), (2), (3) and (4), respectively.

The minimum error between the two fittings of the genes is selected, *i.e.*, $\xi(m,n)$ is chosen, if $\xi(m,n) < \xi(n,m)$ for the mn-th gene pair, identifying the gene that affects the other one more (for the pair). This results in the formation of the $ErrorMatrix$, $Err(i,j)$ (with $1 \leq i,j \leq N$, N being the total number of genes obtained in the bicluster) which is eventually discretized using quantile partitioning [5] for automatically eliminating the higher errors, which contribute less towards regulation. Only those error values are retained for which the absolute value is below a certain threshold, implying a larger interaction between the gene pairs. The entire range $[Err_{\max}(i,j), Err_{\min}(i,j)]$ is divided into three partitions each, using *quantiles* or *partition values*[1] so that the influence of noisy gene patterns are lessened. Error values smaller than Q_1^+ are indicative of less error and high interaction, while those with values in (Q_1^+, Q_2^+] indicate moderate interaction.

An adjacency matrix is calculated as follows:

$$A(i,j) = \begin{cases} +1 \text{ if} & Err(i,j) \leq Q_1^+ \\ 0 & \text{otherwise}, \end{cases} \tag{5}$$

Here we have avoided fitting a gene with itself, assuming self interaction to be absent. Thereafter, a network connecting the various genes is generated. Note that the threshold Q_1^+ is automatically determined from the data distribution, thereby eliminating the need for user-defined parameters.

3 Results

Yeast cell-cycle CDC28 data [6], a collection of 6220 genes for 17 time points, taken at intervals of 10-minutes, were chosen for applying our methodology. Genes that were not annotated and those with more than 30% missing expression values were first removed from the data set. Eventually, a total of 6029 genes were taken for imputation of missing values according to the methodology provided in [7]. The minimum of the errors for fitting one gene with another and *vice-versa* is chosen based on least squares fitting using eqns. (1) – (4). The orders of the polynomial, k are chosen as 1 (linear), 2 (quadratic) and 3 (cubic). It is noticed that better fitting (lower error) is provided by choosing a cubic polynomial to fit a pair of genes. The stronger interactions, using thresholding by quantile partitioning, were retained for forming a network connecting the various genes.

A sample extracted sub-network comprising 14 genes for $k = 3$ is depicted in Fig. 1. A transcription factor is connected to its target gene by an arrow if such a TF-T pair existed within the biclusters. The biclusters were biologically validated from gene ontology study based on the statistically significant GO annotation database[2].

From our computations we obtained a strong interaction between the TF-T pair $YHR084W$-$YNL192W$, indicated by the directed arrow in Fig. 1, which

[1] Quantiles or partition values denote the values of a variate which divide the entire frequency into a number of equal parts.

[2] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

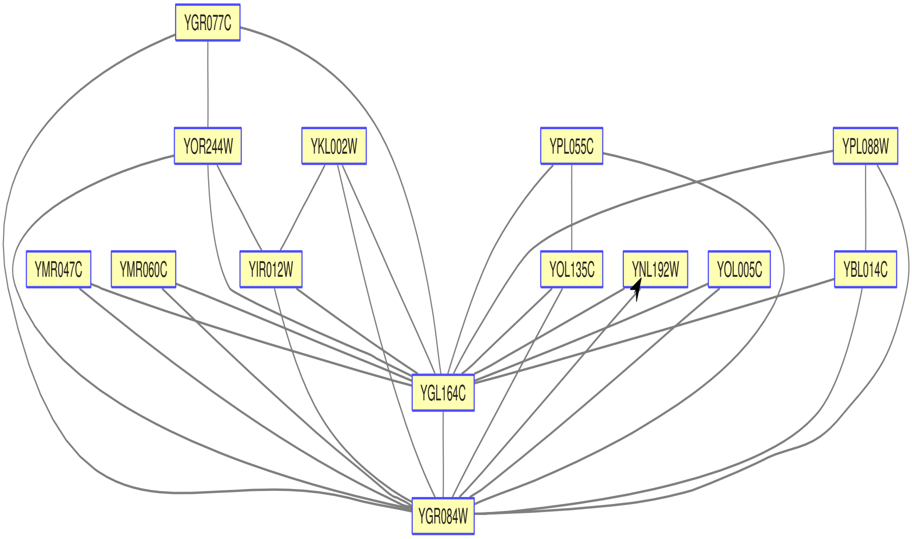


Fig. 1. A sample sub-network (bicluster) of 14 genes with transcription factor $YHR084W$ and target $YNL192W$

may be due to mating-specific binding behaviour or, they may belong to an additional mechanism for cell fusion. We also verified their summary from the *Saccharomyces Genome Database* (SGD)[3]. While identifying the hierarchical structure of regulatory networks [8] it was reported that $YHR084W$-$YNL192W$ forms a TF-T gene pair. One can also arrive at similar established conclusions for the other TF-T pairs (obtained in different biclusters). Our algorithm has not yet detected any false positive or false negative results.

4 Conclusions and Discussion

In this paper we have described an approach using the method of least squares fitting with polynomials, in the framework of continuous-column multiobjective evolutionary biclustering for the generation of gene interaction networks. Biologically relevant biclusters were obtained using multiobjective biclustering, from time-series gene expression data from Yeast. The pairwise time-lagged correlation coefficients among gene pairs were computed by eqn. (1) – (4), followed by the quantile partitioning. The orders of the polynomial, $k = 3$ (cubic) leads to a better fitting (lower error) for a pair of genes. A sample TF-T gene interaction sub-network is depicted in Fig. 1. We tried to analyze the non-linear relationship between two genes which was validated using the statistically significant GO annotation database[4]. The least squares fitting-based method takes care of most

[3] http://www.yeastgenome.org/

[4] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

higher-order dependencies between gene pairs, while automatically eliminating the need for user-defined parameters.

Acknowledgement

The authors gratefully acknowledge Dr. R. K. De and Ms. L. Nayak for their helpful discussion during the progress of this work. Dr. S. Mukhopadhyay gratefully acknowledges the financial assistance received in the form of a grant, BT/BI/04/001/93 from the Department of Biotechnology, Government of India.

References

1. Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics (to appear)
2. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition 39, 2464–2477 (2006)
3. Kohane, I.S., Kho, A.T., Butte, A.J.: Microarrays for an Integrative Genomics. MIT Press, Cambridge (2003)
4. Filkov, V., Skiena, S., Zhi, J.: Analysis techniques for microarray time-series data. Journal of Computational Biology 9, 317–330 (2002)
5. Mitra, S., Das, R., Banka, H., Mukhopadhyay, S.: Gene interaction - An evolutionary biclustering approach. Information Fusion 10, 242–249 (2009)
6. Cho, R.J., Campbell, M.J., Winzeler, L.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
7. Bo, T., Dysvik, B., Jonassen, I.: LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic Acids Research 32, 1–8 (2004)
8. Yu, H., Gerstein, M.: Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of National Academy of Sciences USA 103, 14724–14731 (2006)

Automated Identification of Protein Structural Features

Chandrasekhar Mamidipally[1], Santosh B. Noronha[1], and Sumantra Dutta Roy[2]

[1] Dept. of Chemical Engg., IIT Bombay, Powai, Mumbai - 400 076, India
chandra_m_sekhar@hotmail.com, noronha@che.iitb.ac.in
[2] Dept. of Electrical Engg., IIT Delhi, Hauz Khas, New Delhi - 110 016, India
sumantra@cse.iitd.ac.in

Abstract. This paper proposes an iterative clustering procedure for finding protein motifs which do not necessarily correspond to secondary structures, or to features along the protein backbone. We show the applicability of our method to two important applications namely, protein structure matching, and automatically identifying active sites and other biologically significant sub-structures in proteins.

Keywords: Protein Motifs, Robust Features, Protein Structure Comparison, Clustering, Active sites, Biologically Significant Substructures.

1 Introduction

Proteins are composed of different proportions of amino acids (residues) assembled using peptide bonds into unique 3-D structures. Existing protein structure comparison methods can be subdivided into three major categories: methods relying on subsequent overlapping polypeptide fragments: DALI [1] and CE [2]; methods based on Secondary Structure Elements (SSEs) such as VAST [3], TOP [4] and GRATH [5];and methods that search for generic 3-D patterns that are conserved: Clique detection approaches [6,7], selective groups of amino acids based on physio-chemical properties like Russell [8] and Vishweshwara [9]. Russell through his analysis infers that functional residues within a diameter of 12.0Å are likely to interact. Details of other classification methods can be found in Eidhammer et al. [10] and others [11,12].

In this work, we propose a iterative nearest-neighbour clustering method of generating meaningful cluster motifs (which may not necessarily include secondary structures or preserve sequential order along backbone). We observe correlation of our motifs with biologically significant protein fragment sizes, and suggest two broad implications: automatic detection of strongly conserved motifs including biologically significant sub-structures (active sites), and matching of protein structures using motif alignment. The subsequent sections of this paper are organized as follows: we describe our method for generating cluster motifs in Sec. 2, and analyse the efficacy of our method. Sec. 3 presents two important applications of our clustering method for finding motifs.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 171–176, 2009.

2 Clustering in Protein Feature Identification

We apply the Nearest-Neighbour clustering method for identifying groups of residues (nearest neighbours) as features, for the purpose of protein structure comparison in 3-D space. In this context, residues (C_α atoms) when represented as points in 3-D space can be thought to be interacting if they fall within a diameter of d_{thresh} (Å). Importantly, when *a priori* information of class labels is unavailable, unsupervised learning (clustering) methods are required to partition the collection.

We represent a protein as a collection of its C_α atoms, where each C_α atom is representative of the amino acid residue it lies in. We use a distance d_{thresh} to identify neighbouring C_α atoms. Our method uses three phases. (Fig. 1(a) outlines our method.) In the first phase, Nearest Neighbour clustering is used to create a joint participation matrix $V(q,r)$ where q and r represent random indices of C_α atoms. The joint participation of a pair of points a_q and a_r is expressed in terms of a binary matrix $V(q,r)_{N\times N}$. If a_q and a_r fall in same cluster then the value 1 is assigned to $V(q,r)$ (otherwise 0). A consensus of joint participation of pairs over M iterations $\bigcup_{q,r}^{N} s(q,r)$: (Fig. 1(a), second phase):

$$s(q,r) = \frac{1}{M}\sum_{m=1}^{M}\sum_{\forall q,r}^{N} V(q,r) \qquad (1)$$

M is an important parameter governing the time complexity of the entire procedure: our experiments with a very large number of PDB files show that it is reasonable to take M as a small constant. The final (third) phase involves a ranked search (in decreasing order) to identify high scoring $s(q,r)$ pairs. A final partition $P = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$ is constructed by reading points in the order of decreasing score using the same clustering algorithm, where K is the optimal number of clusters obtained in the consensus procedure. We refer to the centroids of these consensus clusters as 'consensus centroids' earlier work by Ghosh et al. on a artificial data set [13]. We observe that an empirically determined

```
(* --- First Phase: Basic NN --- *)
1. SELECT a random $C_\alpha$ atom, assign it to a new cluster
2. WHILE there are unexamined $C_\alpha$ atoms left REPEAT
    A. SELECT a new random $C_\alpha$
    B. FIND its distance $d_{\min}$ to centroid of nearest cluster
    C. IF $d_{\min} \le d_{thresh}$ THEN
            ASSIGN it to the corresponding cluster
            RECOMPUTE the cluster centroid
       ELSE assign it to a new cluster
(* --- Second Phase: --- *)
REPEAT steps 1, 2 $M$ times (* Sec 2 *)
COMPUTE $V(q,r)$ each time
(* --- Third Phase: Best Clusters --- *)
OUTPUT consensus clusters (* Sec 2 *)
```

(a)

Threshold (Å)	S(K)	GK
	Mean	Mean
4.0	0.066	0.985
4.5	0.121	0.980
5.0	0.177	0.977
5.5	0.246	0.974
6.0	0.317	0.971
6.5	0.413	0.966
7.0	0.559	0.960
7.5	0.727	**0.952**
8.0	**0.878**	0.944
8.5	1.204	0.932

(b)

Fig. 1. (a) Robust, repeatable and fast clustering: Sec. 2, and (b) Validity indices computed for different thresholds on 236 randomly selected proteins of varying sizes value of 100 iterations works well for almost all proteins in the PDB database. A threshold value of 6.0 Å has been used for individual attempts of clustering and further issues are considered in the next section.

We show the efficacy of our clustering method using standard cluster validation indices such as the Scattering Density Index and the Goodman-Kruskal Index [14,15,16]. Fig. 1(b) shows the variation in the above cluster validation indices with the clustering threshold. A low value of $S(K)$ indicates a low average variance i.e., an indication that the clusters are compact. This is desirable in any clustering procedure. Fig. 1(b) indicates that this behaviour occurs for clustering thresholds of up to ∼8.0 Å. The ratio then exceeds 1.0 indicating average variation across clusters is more than the variations as a whole (i.e., the variation of the residues from the protein centroid). A large value of GK indicates good clustering and the value will always be such that $GK \in [-1, 1]$. A value of -1 is indicative of all atoms being wrongly assigned; and a value of 1 is indicative of correct clustering. In Fig. 1(b), we observe that we get a large value of the Goodman-Kruskal Index for small clustering thresholds (<7.5Å, with smaller thresholds resulting in correctly classified points. We note from our analysis of clustering indices that suitable features can be obtained within a clustering threshold range of 6.0 to 7.5 Å. This threshold range ensures that the maximum separation between atoms in a cluster would be of the range 12.0 to 15.0 Å. *An interesting observation from our cluster threshold analysis is that for a threshold of 6 Å, the average number of C_α atoms per cluster is 5 for α and $\alpha + \beta$ and 4 for β class proteins.* Our feature size therefore compares well with methods like DALI [1] (which uses hexapeptide backbone based fragments), and the empirical value of 12.0 Å used to identify interacting residues in the detection of active sites [8]. We make a case for choosing a clustering threshold of about 6 Å, a value which turns out to be biologically significant [8], in addition to being statistically favourable as determined above using cluster validity indices.

3 Cluster Significance: Two Important Applications

This section examines the application of our fast and robust cluster-finding strategy for two important tasks - identifying biologically significant sub-structures in proteins, and comparing two protein structures.

3.1 Automatic Detection of Biologically Significant Substructures

We observe that biologically significant clusters superimpose well due to the retained sub-structure, with observed substitutions typically involving amino acids with similar properties. Fig. 2(a) shows results of experiments with clusters in the Phospholipase family (alpha class structures), with a clustering threshold of 6 Å. Each row in the table represents sample clusters in the protein family, with a cluster represented by single-letter codes corresponding to its constituent amino acid residues. Clusters comprising other class structures found are not shown due to space constraints. Phospholipase A2 is a lipolytic enzyme that is

Amino Acid Labels	Secondary structure	Proteins
CFVHD	$\alpha\alpha\alpha\alpha\alpha$	1JIA, 1PSJ, 1A2A, 1VIP, 1VPI
CFVHK		1PPA
CECDK	$\alpha\alpha\alpha\alpha\alpha$	1J1A, 1CL5
CECDR		1A2A, 1PSJ, 1VIP, 1VPI
DATDRC	$c\alpha\alpha\alpha\alpha\alpha$	1JIA, 1PPA, 1PSJ, 1VIP, 1VPI
DGTDRC		1A2A
1KKMTGK	$\alpha\alpha\alpha\alpha\alpha cc$	1JIA
1KEETGK		1A2A
1LEETGK		1CL5
1LQETGK		1PPA
1LQKTGK		1VPI
IVKMTGK		1VIP

(a)

1JIA	1VIP	SW_{score}	RMSD (Å)
LLQFRK	LFQFAE	15	0.20
PDILC	PPSQC	11	0.06
CECDK	CECDR	31	0.17
IKKMTGK	IVKMTGK	28	0.16
AICFRD	ATCFRD	29	0.22
CYEKV	CYEKV	30	0.14
PVVSYA	PLSSYS	18	0.56
TYS-VCG	SYS-VCG	31	0.31
DATDRC	DATDRC	35	0.08
WKNGTI	FQNGGI	16	0.30
CFVHD	CFVHD	33	0.06
NLKTY	NLNTY	22	0.09

(b)

Fig. 2. (a) Sample clusters of common substructure 'matches' occurring in some proteins of the Phospholipase family. Biologically significant motifs remain conserved, with substitutions typically involving amino acids with similar properties. The underlines letters H (His) and D (Asp) indicate residues in the active site. (b) Comparison of sub-structures in 1JIA and 1VIP using sequence alignment and the Kabsch superimposition method. Sequence alignment scores are generated using the Smith-Waterman matching algorithm, with the BLOSUM62 scoring matrix - details in Sec. 3.1.

involved in the biosynthesis of prostaglandin and other mediators of inflammation. The underlined amino acids in the table are part of the active site, and are implicated in the catalytic mechanism. The neighbourhood of these amino acids is also preserved with allowed substitutions of amino acids with similar properties. For example, the CECDK motif in 1J1A has lysine (K) substituted by arginine (R) in 1A2A; these are both positively charged polar amino acids. Superposition of CECDK motif in 1J1A and CECDR motif in 1A2A yields an rmsd of 0.15 Å indicating a good fit. In comparison, the overall sequence similarity between these two proteins is 59%. It is also evident from Fig. 2(a), that the secondary structure motifs for related clusters are conserved, and that the clusters in this case are likely parts of secondary structural elements (SSEs).

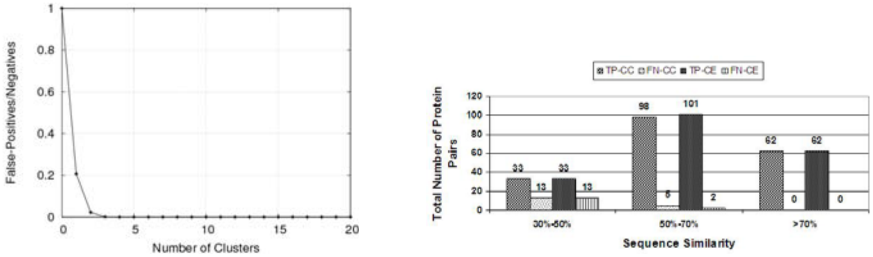
The quantitative extent to which the motifs are similar can be identified as follows. As an example, we compare clusters from 1JIA and 1VIP: these phospholipases share 63.6% sequence identity. We wish to compute an optimal cluster correspondence while accounting for possible amino acid substitutions (algorithm in section 3.2. Fig. 2(b) shows quantitative results for the cluster correspondence in proteins 1JIA and 1VIP. We analyze clusters having a minimum of 4 amino acid residues. Nonsequential residues in a cluster are separated by a '-' symbol. Matched residues are then superposed using the Kabsch method and corresponding rmsd values are shown.

3.2 Protein Structure Matching

In this section, we show the utility of clusters as features for fast and efficient protein structure comparison, using dynamic programming and the Kabsch rotation matrix. The method involves three steps: Representation of clusters, matching, and refinement of matches. We search for inter-cluster chemical similarity across a pair of proteins using the Smith-Waterman (SW) local alignment approach. Only those clusters constituted of at least 4 residues are chosen for alignment. A suitable threshold for the SW score ($SW_{score} >= 10$), derived empirically, is used to detect similarity. Simultaneously, the matched residues are superimposed to within a threshold rmsd value ($R < 0.8$). The resulting pair of matching subsequences are retained as if there are at least 4 residue equivalences.

Towards evaluating the efficacy of our structure comparison approach we have considered 213 similar protein pairs and 868 dissimilar pairs covering five major families: Hemoglobin, Immunoglobulin, Lysozyme, Phosphilipase and Plastocyanins. Similar proteins are observed to have similar substructures even when the extent of sequence similarity is low (data not shown due to space constraints). The minimum number of consensus clusters required to correctly classify protein pairs as similar or dissimilar is seen to be 3 from Figure 3(a). This therefore corresponds approximately to at least a 12 residue match given a minimum cluster size of 4 residues. It should be reiterated that our comparison procedure discounts clusters with less than 4 residues, and also cluster pairs with less than 4 residue equivalences. *It is therefore inappropriate to compare rmsd values from a full protein alignment with the value derived from the structure comparison approach described above.* In Fig. 3(b) we show a comparison of the structure comparison capabilities of our method versus that of CE. We require that homologous pairs have at least three matching cluster pairs in our method. For CE, similarity is claimed given a Z-score > 5 and an rmsd < 4.0 Å[2].

We further analyzed some difficult protein pairs: 1SCE and 1PUC possess high sequence identity but were treated as unique structures by DALI as observed by Shindyalov et al. [2]. On the same proteins we obtained 9 aligned cluster pairs (45 residues) as compared to CE (93/1.26Å). Similarly a comparison of 2SEC and 1EGP yielded 11 residue matches while CE detected 34 (1.2Å). Alignment of the sequentially dissimilar protein pair 1A2Y and 1A4J revealed a potentially strong motif WVRQPPGK–EWL (36-43,46-48) and WYLQKPGQ–KLL (40-47,50-52) respectively. 1HDA and 1PBX, proteins that share 48% sequence



(a) (b)

Fig. 3. (a) Cut-off threshold for similarity across pair of proteins at 6Å across 1081 pairs of protein structures. (b) Number of homologous protein pairs classified as true positives and false negatives. TP-CC, FN-CC represent true positives and false negatives predicted by our method. Similarly TP-CE, FN-CE represents predictions of CE.

identity resulted in a 72 residue alignment as compared to 141 residues aligned by CE. In Immunoglobulins two homologous pairs 1BBD - 1DFB and 1BBD - 4FAB gave 29 and 25 residue matches and were correctly classified as similar; while CE aligned them with rmsd 4.27Å and 4.42Å respectively.

References

1. Holm, L., Sander, C.: Protein Structure Comparison by Alignment of Distance Matrices. Journal of Molecular Biology 233, 123–138 (1993)
2. Shindyalov, I.N., Bourne, P.E.: Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path. Protein Engineering 11, 739–747 (1998)
3. Gibrat, J.F., Madej, T., Spouge, J.L., Bryant, S.H.: The VAST Protein Structure Comparison Method. Biophysics Journal 72, MP298 (1997)
4. Lu, G.: TOP: a new method for protein structure comparisons and similarity searches. Journal of Applied Crystallography 33, 176–183 (2000)
5. Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J., Orengo, C.: Recognizing the fold of a protein structure. Bioinformatics 19, 1748–1759 (2003)
6. Grindley, H.M., Artymiuk, P.J., Rice, D.W., Willett, P.: Identification of Tertiary Structure Resemblance in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm. Journal of Molecular Biology 229, 707–721 (1993)
7. Koch, I., Lengauer, T., Wanke, E.: An algorithm for finding maximal common subtopologies in a set of protein structures. Journal of Computational Biology 3, 289–306 (1996)
8. Russell, R.B.: Detection of Protein Three-Dimensional Side-chain Patterns: New Examples of Convergent Evolution. Journal of Molecular Biology 279, 1211–1227 (1998)
9. Kannan, N., Vishveshwara, S.: Identification of side-chain clusters in protein structures by a graph spectral method. Journal of Molecular Biology 292, 441–464 (1999)
10. Eidhammer, I., Jonassen, I., Taylor, W.R.: Structure Comparison and Structure Pattern. Journal of Computational Biology 7, 685–716 (2000)
11. Kolodny, R., Petrey, D., Honig, B.: Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Current Opinion Structural Biology 16, 393–398 (2006)
12. Dundas, J., Binkowski, T., Dasgupta, B., Liang, J.: Topology independent protein structural alignment. BMC Bioinformatics 8, 388 (2007)
13. Strehl, A., Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research 3, 583–617 (2002)
14. Goodman, L., Kruskal, W.: Measures of associations for cross-validations. Journal of American Statistical Association 49, 732–764 (1954)
15. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: part II. SIGMOD Record 31, 19–27 (2002)
16. Bolshakova, N., Azuaje, F.: Cluster Validation Techniques for Genome Expression Data. Signal Processing 83, 825–833 (2003)

Using Supervised Learning and Comparing General and ANTI-HIV Drug Databases Using Chemoinformatics

Taneja Shweta, Raheja Shipra, and Kaur Savneet

Guru Teg Bahadur Institute of Technology
Rajouri Garden, New Delhi- 110064
shweta_taneja08@yahoo.co.in, shipraraheja@gmail.com,
savneetkaur@yahoo.com

Abstract. Earlier People used to discover new drugs either by chance that is serendipity or by screening the natural products. This process was time consuming, costly as well as required a lot of investment in terms of man-hours. The process of discovering a new drug was very complex and had no rational. Prior to Data Mining , researchers were trying computer methods such as potential drug molecules interactions with the targets and that was also time consuming, costly and required high expertise. Data mining is often described as a discipline to find hidden information in a database. It involves different techniques and algorithms to discover useful knowledge lying hidden in the data. Data mining and the term Knowledge Discovery in Databases (KDD) are often used interchangeably. In this paper, we are implementing the classification technique using WEKA tool for the analysis of similarity between GENERAL DRUGS and ANTI-HIV DRUGS.

Keywords: Classification, Chemoinformatics, Data mining, HIV drugs.

1 Introduction

1.1 Motivation

Earlier all over the world, when a new drug was sought pharmacological researchers used to conduct a blind study of tens or hundreds or thousands of chemical compounds, applying them to an assay for a disease. Drug discovery process had no rationale and it tended to be a bit hit and miss.[1].

Then the field of Data Mining emerged. It combines artificial intelligence with innovative knowledge discovery techniques to analyze the results of pharmacological experiments it conducts itself. By relating the chemical structure of different compounds to their pharmacological activity, Now We are able to learn which chemical compounds should be tested next, bringing a degree of predictability to drug screening procedures. This will help scientists and pharmaceutical companies identify more effective compounds to treat different diseases, allowing them to find drug leads in a fraction of the time and at a fraction of the cost of current methods and could minimize the need for random testing of chemical compounds.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 177–183, 2009.

In our work, we are trying to find the general drugs which can be used instead of ANTI-HIV drugs .All available ANTI-HIV drugs in the market have some side effects. The solution of this problem was to either search for a new molecule which could be very expensive or use the already existing drug .The general drugs and their properties are already known and it can be used. Using the WEKA tool, we have implemented the Classification Technique and produced decision Trees[10] of both the databases .There are various other classifiers also available but we selected decision trees for our similarity search as they can be directly converted into the Classification rules. To construct a rule, a path is traced from root to a leaf node.

1.2 ANTI-HIV Drugs

A virus from the Latin, meaning toxin or poison is a sub-microscopic infectious agent that is unable to grow or reproduce outside a host cell. Viruses infect all cellular life. Viruses consist of two or three parts: all viruses have genes made from either DNA or RNA, long molecules that carry genetic information; all have a protein coat that protects these genes; and some have an envelope of fat that surrounds them when they are outside a cell. A retrovirus is a virus with an RNA genome that replicates by using a viral reverse transcriptase enzyme to transcribe its RNA into DNA in the host cell. The DNA is then incorporated into the host's genome by an integrase enzyme. Antiretroviral drugs are medications for the treatment of infection by retroviruses primarily HIV. There are different classes of antiretroviral drugs that act at different stages of the HIV life cycle. Antiretroviral drugs are broadly classified by the phase of the retrovirus life-cycle that the drug inhibits.

1.3 Data Mining

The term ‘data mining’ is the extraction of interesting (non-trivial, implicit, previously Unknown and potentially useful) information or patterns from data in large databases. Its alternative names are: Knowledge discovery (mining) in databases (K.D.D.), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

2 Classification Technique

Data Classification is a supervised learning technique.

Classification is a Two-Step Process:

1. Model construction: It is used for describing a set of predetermined classes
   - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
   - The set of tuples used for model construction: training set

- The model is represented as classification rules, decision trees, or mathematical formula

2. Model usage: It is used for classifying future or unknown objects

- Estimate accuracy of the model
- The known label of test sample is compared with the classified result from the model
- Accuracy rate is the percentage of test set samples that are correctly classified by the model
- Test set is independent of training set, otherwise over-fitting will occur.

3 Decision Tree

A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. The topmost node in the tree is the root node. Decision trees are used for classification. Given a tuple X,for which the class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple.Decision trees can be easily converted to classification rules. The decision tree classifiers are very popular as it does not require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans .In general decision tree classifiers have good accuracy.

4 About the Tool WEKA

“WEKA” stands for the Waikato Environment for Knowledge Analysis. (Also, the weka pronounced to rhyme with Mecca, is a flightless bird with an inquisitive nature found only on the islands of New Zealand.). Weka is developed at the University of Waikato in New Zealand, the system is written in JAVA an object oriented language that is widely available on all computer platforms, and weka has been tested under various operating systems like Linux, Windows, and Macintosh. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre- and post processing and for evaluating the result of learning schemes on any given dataset. There are several different levels at which Weka can be used. First of all it provides implementations of state-of-the-art learning algorithms that you can apply to your dataset from the command line. It also includes a variety of tools for transforming datasets, like the algorithms for discrimination .We can preprocess a dataset, feed it into a learning scheme, and analyze the resulting classifier and its performance—all without writing any program code at all.

5 Dataset Construction

5.1 Drug Bank- A Knowledge Base for Drugs, Drug Actions and Drug Targets

Drug Bank is a richly annotated resource freely available on internet that combines detailed drug data with comprehensive drug target and drug action information. Since its first release in 2006, Drug Bank has been widely used to facilitate in silico drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education.

We have created two databases with descriptors (Formula weight, Predicted water solubility, Predicted Log P, Experimental Log P, and Predicted Log S) as follows-

- A section of 350 drugs has been made from DRUG BANK
- Then a database of 25 ANTI HIV drugs is made.

A sample of records containing 10 instances is as shown in following tables-Table 5.1 and Table 5.2.

Table 5.1. Database of General drugs

Name	Type	Formula Weight	Predicted Water Solubility mg/mL	P LogP	E LogP	P LogS
Astrozole	Approved	293.3663	6.61E-02	2.32	2.4	-3.65
Acamprosate	Approved	181.21	1.88E+01	-1.78	-1.1	-0.98
Acetazolamide	Approved	222.245	2.79E+00	-0.39	?	-1.9
Adenosine	Approved	267.2413	1.40E+01	-1.2	-1.6	-1.28
Alendronate	Approved	249.096	1.69E+01	-1.34	-4.3	-1.17

Table 5.2. Database of Anti HIV Drugs

Name	Type	Formula Weight	Predicted Water Solubility mg/mL	P LogP	E LogP	P LogS
Abacavir	Approved	286.3323	1.21E+00	0.61	1.1	-2.37
Amprenavir	Approved	505.627	4.91E-02	1.85	?	-4.01
Atazanavir	Approved	704.8555	3.27E-03	4.08	4.5	-5.33
Cidofovir	Approved	279.187	1.15E+01	-2.11	-3.9	-1.38
Darunavir	Approved	547.664	6.68E-02	1.89	1.8	-3.91

6 Experimental Results

There are various tools which are used for Data Mining .Because the WEKA tool is a freely available tool, that's why we implemented classification technique with WEKA tool. WEKA supports input files in 2 formats- ARFF format and CSV format. The figures 6.1 and 6.2 given below show the decision trees of GENERAL drugs and ANTI-HIV drugs.

6.1 Decision Trees

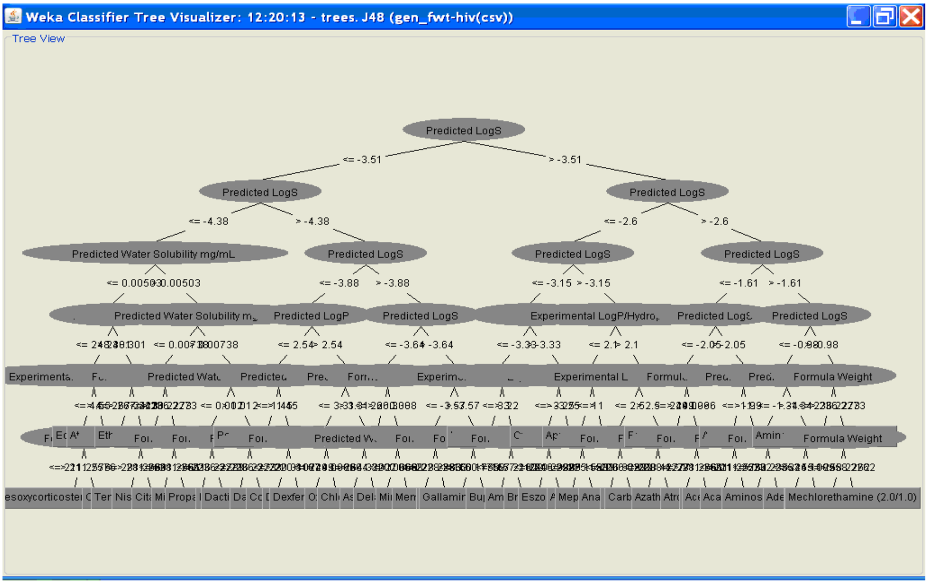


Fig. 6.1. Decision tree of general drugs

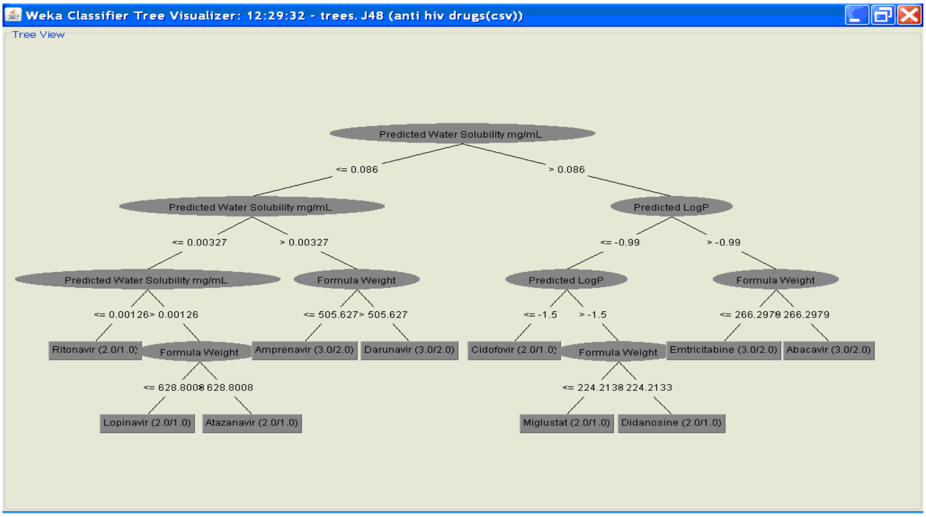


Fig. 6.2. Decision tree of Anti HIV drugs

6.2 Classification Rules

GENERAL DRUGS

1. if P.W.S<=0.00503 and F.W.T<=248.301 and F.W.T<=211.2576
   =>Drug=**Celecoxib**
2. if PWS<=0.00503 and FWT<=248.301 and FWT>211.2576
   =>Drug=**Desoxycorticortesene Pivalate**
3. if PWS<=0.00503 and FWT>.248.301 and FWT<267.2413
   =>Drug=**Atonoxetine**
4. if PWS>0.00503 and PWS>0.00738 and PWS<=0.012 and
   FWT>236.2273=>Drug=**Propafenone**
5. if PWS>0.00503 and PWS>0.00738 and PWS<=0.012 and FWT>236.227
   =>Drug=**Dactinomycin**

ANTI HIV DRUGS

1. if PWS<=0.086 and PWS<=0.00327 and PWS<=0.00126=>Drug=**Ritonavir**
2. if PWS<=0.086 and PWS<=0.00327 and PWS>=0.00126 and
   FWT<=628.800=>Drug=**Lopinavir**
3. if PWS<=0.086 and PWS<=0.00327 and PWS>=0.00126 and
   FWT>628.800=>Drug=**Atazanavir**
4. if PWS<=0.086 and PWS>0.00327 and
   FWT<505.627=>Drug=**Amprenavir**
5. if PWS<=0.086 and PWS>0.00327 and
   FWT>=505.627=>Drug=**Darunavir**

7 Conclusion

The table 7.1 given below shows the predicted results of similarity between GENERAL DRUGS and ANTI-HIV DRUGS on the basis of range of descriptors.

Table 7.1. Predicted results of similarity between GENERAL DRUGS and ANTI-HIV DRUGS on the basis of range of descriptors

Name	Type	Formula Weight	Predicted Water Solubility mg/mL	P LogP	E LogP	P LogS
Abacavir	Approved	286.3323	1.21E+00	0.61	1.1	-2.37
Amprenavir	Approved	505.627	4.91E-02	1.85	?	-4.01
Atazanavir	Approved	704.8555	3.27E-03	4.08	4.5	-5.33
Cidofovir	Approved	279.187	1.15E+01	-2.11	-3.9	-1.38
Darunavir	Approved	547.664	6.68E-02	1.89	1.8	-3.91

FWT-Formula Weight.
PWS-PredictedWater Solubility.

As shown in the above table, some ANTI-HIV drugs are similar to GENERAL drugs on the basis of range of descriptors specified. These results are only forecasted results and may be clinically tested.

References

1. Data Mining Promises To Dig Up New Drugs, Science Daily (February 3, 2009)
2. Zimmermann, A., Bringmann, B.: CTC - correlating tree patterns for classification. In: Fifth IEEE International Conference on Data Mining, Novemmber 27-30, p. 4 (2005), doi:10.1109/ICDM.2005.49
3. Bjorn, B., Albrecht, Z.: Tree Decision Trees for Tree Structured Data. Institute of Computer Science, Machine Learning Lab, Albert-Ludwig-University Freiburg, Georges-Kohler-Allee 79, 79110 Freiburg, Germany
4. Ósk, J.S., Steen, J.F., Brunak, S.: Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates 21(10), 2145–2160 (2005), doi:10.1093/bioinformatics/bti314
5. Lumini, A., Nanni, L.: Machine Learning for HIV-1 Protease Cleavage Site Prediction. Elsevier Science Inc., New York (2005)
6. Niko, B., et al.: Diversity and complexity of HIV-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype. IEEE Intelligent Systems 16(6), 35–41 (2001)
7. Hiroto, S.: Mining complex genotypic features for predicting HIV-1 drug resistance. National Institute of Informatics (2007)
8. Lin Ray, S.: A combined data mining approach for infrequent events: analyzing HIV mutation changes based on treatment history. Stanford University Stanford, CA 94305, United States
9. Wasim, A.M., Sidhu, M.: Chemoinformatics: Principles and Applications. Pesticide Residue Laboratory, Department of Agricultural Chemicals, Department of Agricultural Chemistry and Soil Science, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur-741252, Nadia, West Bengal, India
10. Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., Selbig, J.: Identifying drug resistance-associated patterns in HIV genotypes

Multiple Sequence Alignment Based Upon Statistical Approach of Curve Fitting

Vineet Jha, Mohit Mazumder*, Hrishikesh Bhuyan, Ashwani Jha,
and Abhinav Nagar

InSilico Biosolution, 103B, North Guwahati College Road, Abhoypur, Near IIT-Guwahati,
P.O – College Nagar, North Guwahati – 781031, Assam, India
mazumder.mohit@gmail.com

Abstract. The main objective of our work is to align multiple sequences together on the basis of statistical approach in lieu of heuristics approach. Here we are proposing a novel idea for aligning multiple sequences in which we will be considering the DNA sequences as lines not as strings where each character represents a point in the line. DNA sequences are aligned in such a way that maximum overlap can occur between them, so that we get maximum matching of characters which will be treated as our seeds of the alignment. The proposed algorithm will first find the seeds in the aligning sequences and then it will grow the alignment on the basis of statistical approach of curve fitting using standard deviation.

Keywords: Multiple Sequence Alignment, Sequence Alignment, Word Method, Statistically Optimized Algorithm, Comparative Genome Analysis, Cross Referencing, Evolutionary Relationship.

1 Introduction

Multiple sequence alignment is a crucial prerequisite for biological sequence data analysis.

It is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. A large number of multi-alignment programs have been developed during last twenty years. There are three main considerations in choosing a program: biological accuracy, execution time and memory usage. Biological accuracy is generally the most important concern amongst all. Some of the prominent and accurate programs according to most benchmarks are *CLUSTAL W* [1], *DIALIGN* [2], *T-COFFEE* [3], MAFFT, MUSCLE, PROBCONS . An overview about these tools and other established methods are given [4].

T-COFFEE is a prototypical consistency- based method which is still considered as one of the most accurate program available. MAFFT and MUSCLE have a similar design, building on work done by Gotoh in the 1990s that culminated in the PRRN

* Corresponding author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 184–192, 2009.

and PRRP programs [10,11], which achieved the best accuracy of their time but were relatively slow and were not widely adopted.

The recently published methods ALIGN-M [12], DIALIGN [2, 13], POA [14, 15] and SATCHMO [16] have relaxed the requirement for global alignment by allowing both alignable and non-alignable regions. Although these methods are sometimes described as 'local'; alignable regions must still be co-linear (i.e. appear in the same order in each sequence). This model is appropriate for protein families with well-conserved core blocks surrounded by variable regions, but not when input sequences have different domain organizations.

2 Objective

The main purpose of undertaking this study is to develop a new algorithm which can align multiple sequences together in lesser time.

The alignment of sequences will be done on the basis of statistical approach in lieu of heuristics approach. Statistical method of curve fitting by means of standard deviation and stochastic approaches will be used.

3 Algorithm for Multiple Sequence Alignments

1. Take 2 sequences.
2. Find the seed for alignment.
3. Build the distance matrix between the sequences.
4. Plot a graph.
5. Plot the scores of sequences in distance matrix on the graph about the line x=y.
6. Find the minimum deviation of that curve or line of minimum deviation.
7. Draw the minimum deviation onto the graph and name it as master alignment.
8. Arrange the alignment according to the master sequence alignment on distance matrix.

3.1 Detail Explanation of above Algorithm

Here we consider a DNA sequence as line not as string where each character represents a point in the line. The sequences are lined up in such a way that maximum overlap occurs between them, thus giving maximum matching of characters.

3.1.1 Hypothesis

For example consider these two sequences

DNA1 ATCGGGGCTATC
DNA2 ATCCCCCTATTG

A matrix was built according to the following conditions. If the character matches in the two sequences the distance between the two point is 0 and if mismatch occurs then the distance is given by the PAM [17] or BLOSSUM [18] matrix. The distance contributed from the deviation to the line where x=y. Then the matrix is constructed and plotted on a graph(fig 1). The graph and matrix are as following-

Table 1. This is matrix formed between the two DNA sequences in this match score=0 and mismatch score is given by PAM and BLOSSUM matrix 1st and 2nd row consist of DNA1 & DNA2 unaligned row is taken as zero at this stage ,3rd &5th rows are the score of matrix

A	T	C	G	G	G	G	C	T	A	T	C	DNA1
A	T	C	C	C	C	C	T	A	T	T	G	DNA2
0	0	0	1	1	1	1	2	1	1	0	1	DNA1
0	0	0	0	0	0	0	0	0	0	0	0	UNALIGNED SEQ
0	0	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2

The graph will look like fig 1 for the above sequences. The matching string will be called as seed (where our final alignment will grow) and the mismatch string will expand like a bulb. Now the main problem whice arises at this stage is how to deflate the bulb. This is where statistics comes into action. The bulb is like a scattered plot. We have to find a line or curve of minimum deviation in the bulb which will be our master alignment line.

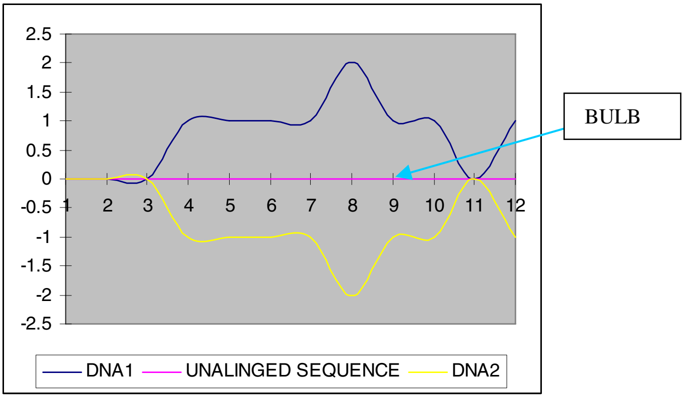


Fig. 1. This graph shows the score of matches & mismatches of both DNA1 & DNA2, In this matches are called as “seed” & mismatches will appear as “bulb”& the middle line represents the Master Alignment

A matrix was built which contains the master alignment. A score was assigned to it. The master alignment line will guide the sequences to align properly as following:

Table 2. This matrix consist Aligned sequence score along with previous score of DNA1 &DNA2 which is obtained from above graph

A	T	C	G	G	G	G	C	T	A	T	C	DNA1
A	T	C	C	C	C	C	T	A	T	T	G	DNA2
0	0	0	1	1	1	1	2	1	1	0	1	DNA1
0	0	0	0	0	0	-1	2	-2	1	-1	0	ALING SEQ
0	0	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2

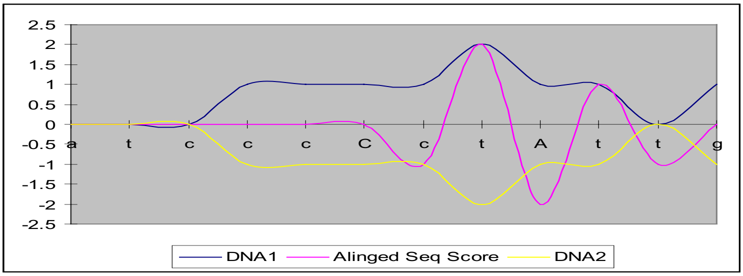


Fig. 2. This graph shows the curve of scores of DNA1 &DNA2 along with the curve obtained by the minimum deviation of these named as ”MASTER ALIGNMENT”

After the matrix is prepared the final alignment is done on the basis of scoring matrix, adjusting gaps and mismatch, checking the maximum aligning score. Introduction of gaps are done to maximize the alignment so that better alignment and homology is achieved.

Table 3. This matrix shows the alignment done by the help of “MASTER ALIGNMEENT” obtained on the previous graph by the help of minimum deviation method

a	t	c	g	g	g	g	c	t	a	t	#	c	DNA1
a	t	c	c	c	c	#	c	t	a	t	t	g	DNA2
a	t	c	g/c	g/c	g/c	g	c	t	a	t	t	c/g	MASTER ALINGMENT (MA)
0	0	0	1	1	1	0	0	0	0	0		1	DNA1 SCORE
0	0	0	0	0	0		0	0	0	0	0	0	MA SCORE
0	0	0	-1	-1	-1	0	0	0	0	0	0	-1	DNA2 SCORE

Final graph based on the matrix score was obtained. Size of the bulb is reduced. The bulb that is visible in the graph is showing the mismatch between the bases present in the alignment.

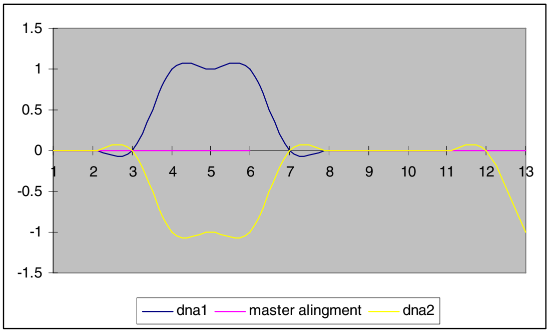


Fig. 3. This figure shows the alignment of DNA1 & DNA2 with the final score along with the Master Alignment curve

Final alignment of the sequences will be

ATCGGGGCTAT- C DNA1

ATCCCC - CTATTG DNA2

Colour Notaions

ATCG Shows the mismatch in the sequences

ATCG shows the gaps in the sequences

Consider the alignment of tree or more sequences

DNA1 ATCGGGGCTATC
DNA2 ATCCCCCTATTG
DNA3 CGCCCGGCTATG

A matrix was built according to the following conditions. If the character matches in the two sequences the distance between the two point is 0 and if mismatch occurs then the distance is given by the PAM [17] or BLOSSUM [18] matrix. The distance contributed from the deviation to the line where x=y. Then the matrix is constructed and plotted on a graph(fig 4). The graph and matrix are as following-

Table 4. This matrix is same as that in previous example but this is for aligning 3 DNA sequences

a	t	c	g	g	g	g	c	t	a	t	c	DNA1
a	t	c	c	c	c	c	t	a	t	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	t	g	DNA3
2	2	0	1	1	1	1	2	1	1	0	1	DNA1SCORE
2	2	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2 SCORE
-2	-2	0	-1	-1	1	1	2	1	1	0	-1	DNA3 SCORE
0	0	0	0	0	0	0	0	0	0	0	0	UNALING SCORE

The graph will look like fig 4 for the above sequences. The matching string will be called as seed (where our final alignment will grow) and the mismatch string will expand like a bulb. Now the main problem whice arises at this stage is how to deflate the bulb. This is where statistics comes into action. The bulb is like a scattered plot. We have to find a line or curve of minimum deviation in the bulb which will be our master alignment line.

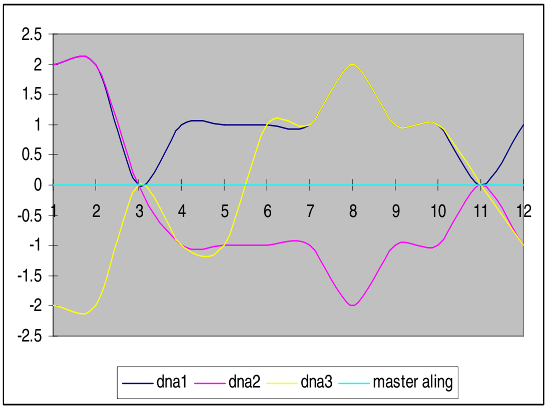


Fig. 4. This graph shows the score of matches & mismatches of both DNA1, DNA2, DNA3similarly as in the early case & the middle line represents the Master Alignment

A matrix was built which contains the master alignment. A score was assigned to it. The master alignment line will guide the sequences to align properly as following:

Table 5. This matrix consist Unaligned sequence score along with score of DNA1, DNA2& DNA3 obtained from above graph

a	t	c	g	g	g	g	c	t	a	t	c	DNA1
a	t	c	c	c	c	c	t	a	t	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	t	g	DNA3
2	2	0	1	1	1	1	2	1	1	0	1	DNA1SCORE
2	2	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2 SCORE
-2	-2	0	-1	-1	1	1	2	1	1	0	-1	DNA3 SCORE
2	2	0	-1	-1	1	1	-2	1	1	0	-1	UNALING SCORE

After the matrix is prepared the final alignment is done on the basis of scoring matrix, adjusting gaps and mismatch, checking the maximum aligning score. Introduction of gaps are done to maximize the alignment so that better alignment and homology is achieved.

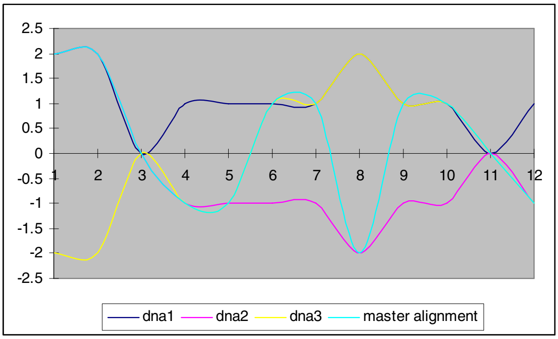


Fig. 5. This graph shows the curve of scores of DNA1, DNA2 &DNA3 along with the curve obtained by the minimum deviation of these curve for the alignment purpose named as "MASTER ALIGNMENT"

Table 6. It consist of the final alignment score of DNA1, DNA2& DNA3 which is used to draw the final alignment curve , these scores are obtained by aligning the DNA's with the help of Master Alignment by shifting of bases

a	t	c	g	g	g	g	c	t	a	T	#	c	DNA1
a	t	c	c	c	c	#	c	t	a	T	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	T	#	g	DNA3
2	2	0	1	1	1	0	0	0	0	0		1	DNA1SCORE
2	2	0	-1	-1	-1		0	0	0	0	0	-1	DNA2 SCORE
-2	-2	0	-1	-1	1	0	0	0	0	0		-1	DNA3 SCORE
2	2	0	-1	-1	1	0	0	0	0	0	0	-1	ALIGNMNT SCORE

Final graph based on the matrix score was obtained. Size of the bulb is reduced. The bulb that is visible in the graph is showing the mismatch between the bases present in the alignment.

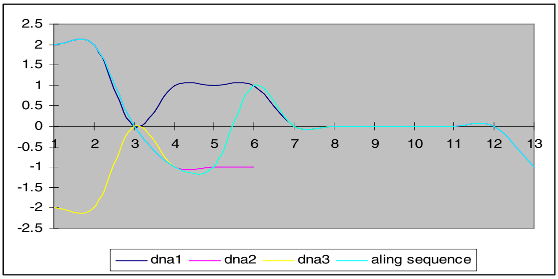


Fig. 6. This figure shows the alignment curve of DNA1, DNA2& DNA3 with the final score in the Matrix along with the Master Alignment curve

Table 7. This matrix shows the alignment of all 3 DNA sequences done by the help of "MASTER ALIGNMENT" obtained on the previous graph by the help of minimum deviation method

a	t	c	g	g	g	g	c	t	a	t	#	c	DNA1
a	t	c	c	c	c	#	c	t	a	t	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	t	#	g	DNA3 SCORE
a/c	t/g	c	g/c	g/c	g/c	g	c	t	a	t	t	c/g	MASTER ALIGNMENT

Final alignment of the sequences will be: **COLOUR NOTAION**

DNA1 ATCGGGGCTAT –C

DNA2 ATCCCC – CTATAG

DNA3 CGCCCGGCTAT -G

ATCG Shows the mismatch in the sequences

ATCG shows the gaps in the sequences

4 Result and Discussion

The main feature of the algorithm is sequential execution through seed finding and growing those seeds on the basis of statistical approach of curve fitting using standard deviation.

The seed of alignment can be found either by using heuristics approach of BLAST [19] or FASTA algorithm or using DOTPLOT approach. Here we have used DOTPLOT approach to find out the short or long inter sequences match (seed). Seeds are then filtered out and un-aligned sequences will be aligned using this algorithm.

This algorithm is unique as it is not based on progressive method and also doesn't divide the sequences in bits and pieces for alignment so it gives better alignment. As it uses seeds to align the sequences it gives better homology.

Results have shown that this algorithm worked well even if seed is absent or the sequences are far apart from each other.

It can align cDNA with gene which most of MSA algorithms fails to do. Advantage of cDNA not containing the introns can be used here. Exons on the cDNA are aligned with coding genes in the DNA sequence which can than be used as seeds.

This leads to a phenomenon of DNA looping. These looping DNA are aligned by creating gap in the cDNA.Here these gaps indicates introns in the genome.

Here we have aligned the EST with gene to find out point mutation. The data was downloaded from UCSC genome browser[20,21]. UCSC refgene contains the co-ordinates for exon and intron in gene. While aligning EST with gene the EST must match with exon (as a biological phenomenon). Thus we have calibrated our MSA verses (Stretcher[22], Matcher[22,23], Needle[24] and Waunch[25]) to find match and mismatch in EST verses gene alignment.

While performing the alignment we noticed that exon can be subdivided if it contains gaps with respect to EST and vice versa. Thus we divided the gene in such a way that no gaps occur in EST and exon. Now two sequences were taken containing only mismatches. Any gap here is considered as false positive(FP) as no sequence contains gap. Any mismatch which is missed will be reported as true negative(TN) . Any wrong mismatch reported will be false negative(FN). All true matches is true positive(TP).

Sensitivity
TP (true positive)
FN (False negative)
Sensitivity= TP/TP+FN

Specificity
TN(true negative)
FP(false positive)
Specificity= TN/TN+FP

Table 8. Showing sensitivity and specificity of the algorithm in comparison to other popular algorithms

Program	Sensitivity	Specificity
Our Algorithm	99.91	98.73
Stretcher	99.51	107.37
Matcher	99.44	106.25
Needelman-Waunch	98.14	61.17
Smith-Waterman	95.9	60.96

Above comparison shows that this algorithm gives better sensitivity among all other algorithms taken into account and gives specificity comparable to all other algorithms.

References

1. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673–4680 (1994)
2. Morgenstern, B.: DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiServ. Nucleic Acids Research 32, W33–W36 (2004)
3. Notredame, C., Higgins, D., Heringa, J.: T-Coffee: a novel algorithm for multiple sequence alignment. J. Mol. Biol. 302, 205–217 (2000)

4. Notredame, C.: Recent progress in multiple sequence alignment: a survey. Pharmacogenomics 3, 131–144 (2002)
5. Lee, C., Grasso, C., Sharlow, M.F.: Multiple sequence alignment using partial order graphs. Bioinformatics 18(3), 452–464 (2002)
6. Edgar, R.: MUSCLE: Multiple sequence alignment with high score accuracy and high throughput. Nuc. Acids Res. 32, 1792–1797 (2004)
7. Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S.: ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Research 15, 330–340 (2005)
8. Katoh, K., Misawa, K., Kuma, K., Miyata, T.: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066 (2002)
9. Edgar, R.C.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004)
10. Gotoh, O.: Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J. Mol. Biol. 264, 823–838 (1996)
11. Gotoh, O.: A weighting system and algorithm for aligning many phylogenetically related sequences. Comput. Appl. Biosci. 11, 543–551 (1995)
12. Van Walle, I., Lasters, I., Wyns, L.: Align-m-a new algorithm for multiple alignment of highly divergent sequences. Bioinformatics 20, 1428–1435 (2004)
13. Morgenstern, B.: DIALIGN: 2 improvement of the segment-tosegment approach to multiple sequence alignment. Bioinformatics 15, 211–218 (1999)
14. Grasso, C., Lee, C.: Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics 20, 1546–1556 (2004)
15. Lee, C., Grasso, C., Sharlow, M.F.: Multiple sequence alignment using partial order graphs. Bioinformatics 18, 452–464 (2002)
16. Edgar, R.C., Sjölander, K.: SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics 19, 1404–1411 (2003)
17. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. In: Dayhoff, M.O. (ed.) Atlas of Protein Sequence and Structure, vol. 5(3), pp. 345–352 (1978)
18. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89(biochemistry), 10915–10919 (1992)
19. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)
20. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: Genome Res. 12(6), 996–1006 (June 2002)
21. University of California santa Cruz, `http://genome.ucsc.edu/`
22. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000)
23. MacLaughlin, D.S.: MATCHER: a program to create and analyze matched sets. Comput. Programs Biomed. 14(2), 191–195 (1982)
24. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48(3), 443–453 (1970)
25. Smith, T.F., Waterman, M.S., Fitch, W.M.: Comparative biosequence metrics. J. Mol. Evol. 18(1), 38–46 (1981)

A Constraint Based Method for Optimization in Metabolic Pathways

Mouli Das[1], Subhasis Mukhopadhyay[2], and Rajat K. De[1]

[1] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
{mouli_r,rajat}@isical.ac.in

[2] Department of Bio-Physics, Molecular Biology and Bioinformatics, Calcutta University, Kolkata 700 009, India
smbmbg@caluniv.ac.in

Abstract. The analysis of pathways in metabolic networks has become an important approach for understanding metabolic systems. Constraint-based approaches recently brought new insight into our understanding of metabolism. Here we introduce a method for identifying optimal metabolic pathways. The method, generates data on reaction fluxes based on biomass conservation constraint and formulates constraints incorporating weighting coefficients corresponding to concentration of enzymes catalyzing the reactions. The effectiveness of the present method is demonstrated on the pentose phosphate pathway (PPP) in *T. cruzi* and on a large network of the plant polyamine pathway existing in the literature. A comparative study with the existing extreme pathway analysis (EPA) also forms a part of this investigation.

Keywords: FBA, stoichiometric matrix, polyamine pathway.

1 Introduction

Cellular metabolism, the integrated interconversion of thousands of metabolic substrates into products through enzyme-catalysed biochemical reactions, is the most investigated complex intracellular web of molecular interactions. Metabolic pathways are defined by a series of chemical reactions that are interlinked with each other [1]. Metabolic pathway modeling is needed as metabolism is the "chemical engine" that drives the living process, and is a fundamental determinant of cell physiology. The flux balance analysis (FBA), a constraint-based approach applied to genome-scale metabolic models can be used to make predictions of flux distributions based on linear optimization [2]. Here we describe a method based on the framework of FBA for identifying optimal metabolic pathways that yields the maximum amount of the target metabolite from a given substrate. Results, comparison with the existing EPA [3] and their biological relevance have been provided for the PPP in *T. cruzi* and on one large polyamine biosynthesis pathway observed in plants.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 193–198, 2009.

2 A Constraint Based Method

A metabolic network consists of internal fluxes v within the network and exchange fluxes b that can either enter or exit the system. The rate of growth of the target metabolite B on the starting metabolite A, which needs to be maximized is the algebraic sum of the weighted fluxes of $R_1, R_2, \ldots, R_s$ reactions by s different paths and is given by

$$z = \sum_{k=1}^{s} c_k v_k \tag{1}$$

c_k indicates the enzyme concentration level. The flux vectors, $\mathbf{v}$ that satisfy approximately the quasi-steady state condition $\mathbf{S}.\mathbf{v} \approx \mathbf{0}$ and the inequalities described later are generated. $\mathbf{S}$ is the $m \times n$ stoichiometric matrix with m as the number of metabolites and n as the number of reactions. We generate l number of basis vectors $\mathbf{v}_b$ that form the null space of $\mathbf{S}$ and also generate l number of non-negative random numbers a_p, $p = 1, 2, \ldots, l$ to generate a vector $\mathbf{v} = \sum_{p=1}^{l} a_p \mathbf{v}_{bp}$ satisfying the following inequality constraints. All the internal fluxes are positive yielding: $v_i \geq 0, \forall i$. The constraints on the exchange fluxes depending on their direction can be expressed as $\alpha_j \leq b_j \leq \beta_j$ where $\alpha_j \epsilon \{-\infty, 0\}$ and $\beta_j \epsilon \{0, \infty\}$. In real systems the genes may not be expressed at the required level. So we define a new set of constraints as

$$\mathbf{S}(\mathbf{C}\mathbf{v}) = \mathbf{0} \tag{2}$$

where $\mathbf{C}$ is an $n \times n$ diagonal matrix whose i-th diagonal element is c_i for each i. That is, if $\mathbf{C} = [\gamma_{ij}]_{n \times n}$, then $\gamma_{ij} = \delta_{ij} c_i$, where δ_{ij} is the Kronecker delta. So the problem boils down to a maximization problem, where z is maximized with respect to $\mathbf{c}$, subject to satisfying the constraint given in equation(2) along with the inequality constraints. Combining equations (1) and (2), we can reformulate the objective function as

$$y = 1/z + \mathbf{\Lambda}^T.(\mathbf{S}.(\mathbf{C}.\mathbf{v})) \tag{3}$$

that has to be minimized with respect to the weighting factors c_i for all i. The term $\mathbf{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_m]^T$ is the regularizing parameter. Initially, a set of random values in $[0, 1]$ corresponding to c_i's are generated which are then modified iteratively by gradient descent optimization algorithm, where $\Delta c_i = -\eta \partial y / \partial c_i$. The term η is a small positive quantity indicating the rate of modification. The modified value of c_i is given by $c_i(t+1) = c_i(t) + \Delta c_i, \ \forall i, \ t = 0, 1, 2, \ldots$ λ is chosen empirically from 0.1 to 1.0 in steps of 0.1. For each value of λ as we are increasing the number of iterations, the value of y gradually decreases and the corresponding c_i-values indicate an optimal pathway.

3 Results and Comparison

When applied to various organisms, the constraint based method yields more significant and biologically relevant results as compared to that of EPA. Here,

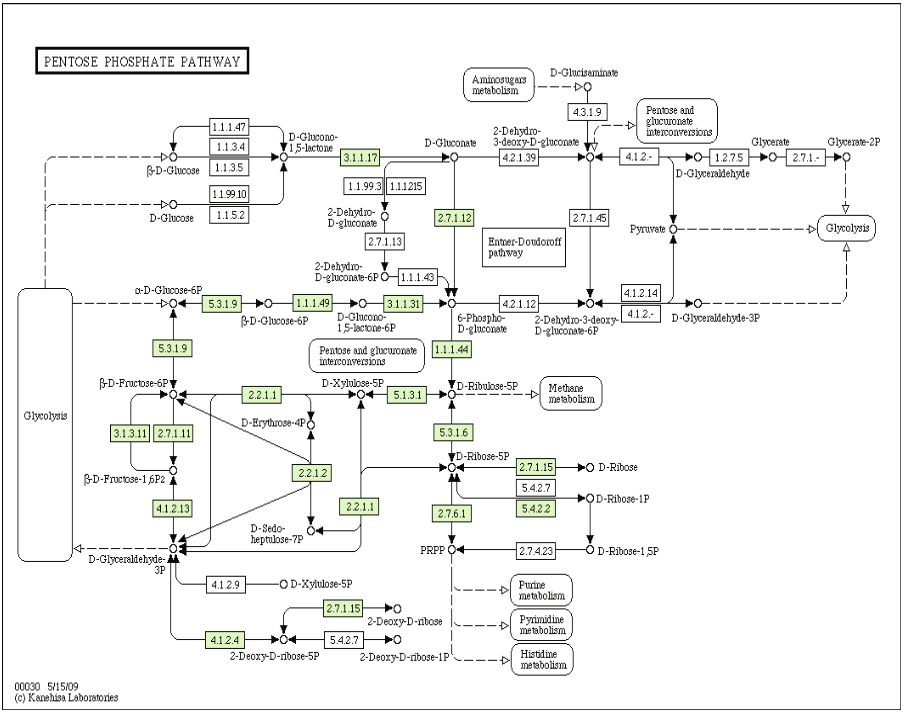


Fig. 1. Pentose Phosphate Pathway in *T. cruzi*

we apply our method to the PPP in the parasite *T. cruzi* and on one large plant polyamine pathway obtained from the KEGG database [1].

PPP is an anabolic pathway that utilizes 6 carbons of glucose to generate NADPH (reduced form of NADP+) and pentose(5-carbon) sugars. The PPP is a key metabolic process in the cell because it provides reductive power, ribose-5-phosphate and glycolytic intermediates. The pathway has an oxidative phase that generates NADPH, and the second non-oxidative phase synthesizes 5-carbon sugars [4]. The biochemical evidence obtained so far suggest that the oxidative branch is essential for protection of the parasite against oxidative stress. The PPP has been shown to be functional in T. cruzi and the seven enzymes of the pathway have been expressed and fully characterized. The balance between the 2 branches is necessary to maintain the metabolic efficiency of the cancer cell for growth and proliferation [5]. The pathway consists of 18 metabolites and 33 fluxes (Fig. 1) where the starting and target metabolites are α-D-Glucose-6P, D-Glyceraldehyde-3P and β-D-Fructose-6P respectively. The present method generates the optimal pathway as $\alpha-D-Glucose-6P \rightarrow \beta-D-Glucose-6P \rightarrow D-Glucono-1,5lactone-6P \rightarrow 6-Phospho-D-Gluconate \rightarrow D-Ribulose-5P \rightarrow D-Xylulose-5P + D-Erythrose-4P \rightarrow D-Glyceraldehyde-3P + \beta-D-Fructose-6P$. The EPA method generates a different pathway

[1] http://www.genome.jp/kegg/pathway.html

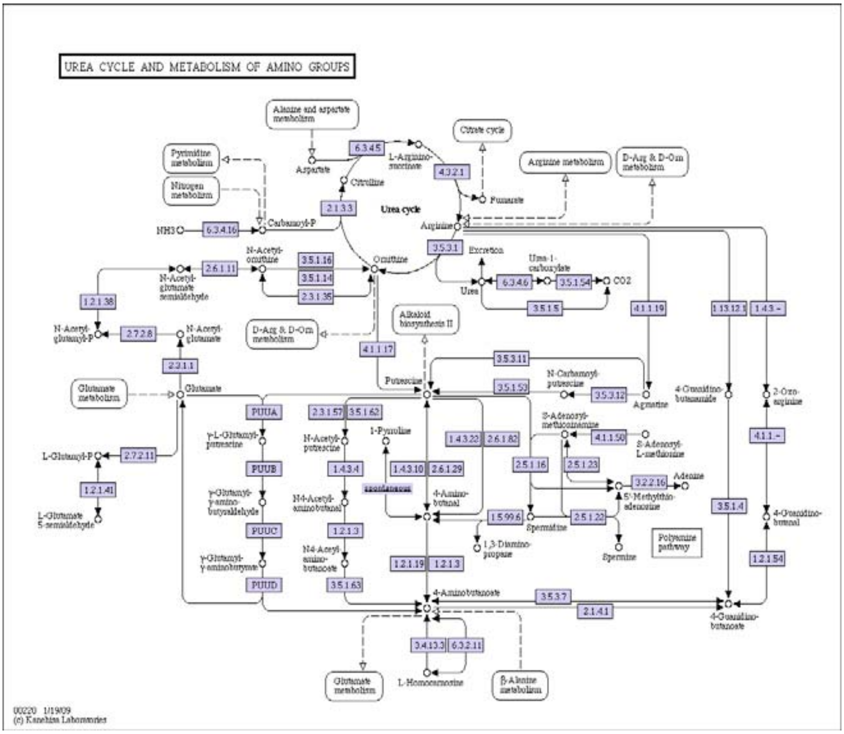


Fig. 2. Polyamine pathway

as $\alpha - D - Glucose - 6P \rightarrow \beta - D - Fructose - 6P \rightarrow D - Xylulose - 5P + D - Erythrose - 4P \rightarrow D - Glyceraldehyde - 3P + \beta - D - Fructose - 6P$.

Polyamines (PAs) are naturally occurring low molecular weight, polycationic, aliphatic nitrogenous compounds found in all organisms and are essential for cellular proliferation and normal cellular function. The most common PAs in plants are spermidine, spermine and their precursor putrescine [6]. PAs take part in many important cellular processes such as cell division, protein synthesis, DNA replication and response to abiotic stress. The polycationic nature of PAs is one of the main properties believed to mediate their biological activity. They are, indeed, able to bind several negatively charged molecules, such as DNA, proteins, membrane phospholipids and proteins. Polyamine biosynthetic pathways in plants have been well elucidated [7]. The intermediate metabolite putrescine is formed either by direct decarboxylation of L-ornithine by the enzyme ornithine decarboxylase (ODC, EC 4.1.1.17), or by decarboxylation of arginine by arginine decarboxylase (ADC, EC 4.1.1.19) via agmatine [8]. Spermidine and the target metabolite spermine are synthesized by the sequential addition of an aminopropyl group to putrescine by spet-midine synthase and spermine synthase, respectively. The aminopropyl group is donated by decarboxylated S-adenosylmethionine (SAM), which is produced by S-adenosylmethionine decarboxylase (SAMDC, EC 4.1.1.50). There are 66 reactions and 42 metabolites for the pathway (Fig. 2) where we are maximizing the yield of Spermine from the starting metabolite Arginine. The present method generates the optimal paths as follows: $Arginine \rightarrow Ornithine \rightarrow Putrescine \rightarrow Spermidine \rightarrow Spermine$ and $Arginine \rightarrow Agmatine \rightarrow Putrescine \rightarrow Spermidine \rightarrow Spermine$. The EPA method results in $Arginine \rightarrow Agmatine \rightarrow Putrescine \rightarrow Spermidine \rightarrow Spermine$ as an optimal pathway. It has been demonstrated in various plant tissues that the pathway obtained by our method (ornithine decarboxylase pathway) is active in cell proliferation and the the pathway obtained by EPA (arginine decarboxylase pathway) is involved in embryo and organ differentiation [9].

4 Biological Relevance and Validation

There are 2 paths starting from $\alpha - D - Glucose - 6P$ in the PPP in *T. cruzi* (Fig. 1) leading to formation of $\beta - D - Glucose - 6P$ and $\beta - D - Fructose - 6P$ in the 2 branches. The path through $\beta - D - Glucose - 6P$ is followed till the intermediate metabolite $D - Ribulose - 5P$ which gets divided into 2 paths of which the one leading to formation of $D - Xylulose - 5P + D - Erythrose - 4P$ is followed as it terminates in the desired target and the other path leading to $D - Ribose - 5P$ is not followed as it ends up in some other target. Of the 2 paths emerging from $D - Xylulose - 5P + D - Erythrose - 4P$, the path leading to $\beta - D - Fructose - 6P$ is not followed as it terminates in $2 - Deoxy - D - ribose$ which is not the desired target. The other path leads to formation of $D - Glyceraldehyde - 3P + \beta - D - Fructose - 6P$ which are the desired targets. The first 3 steps of the PPP which has been produced by the constraint based method, glucose 6-phosphate converted to ribulose 5-phosphate by the actions of the enzymes glucose 6-phosphate dehydrogenase (Glc6PD, EC 1.1.1.49), 6-phosphogluconolactonase (6PGL, EC 3.1.1.31) and 6-phosphogluconate dehydrogenase (6PGDH, EC 1.1.1.44) are crucial paths. These reactions are the only source of NADPH, which is needed to reduce peroxides and other oxidizing agents that may otherwise damage the cell. The sequence of steps that leads to formation of the optimal path has been cited in [10].

Starting from arginine there are five paths leading to formation of ornithine, urea, agmatine, 4-Guanidino-butanamide and 2-Oxo-arginine (Fig. 2). The path that leads to urea, 4-Guanidino-butanamide and 2-Oxo-arginine are not followed as they donot lead to the desired target. The paths leading to agmatine and ornithine are followed as they terminate in the desired target. Of the 3 paths emerging from ornithine the paths leading to formation of citrulline and N-acetyl-ornithine are not followed as they donot terminate in the desired target. The path that leads to putrescine is followed. Of the 4 paths emerging from Putrescine, the paths leading to γ-L-glutamyl-putrescine, N-Acetyl-putrescine and 4-Amino-butanal are not followed as they terminate in some other end products. The path leading to spermidine is followed as it leads to the desired target. From spermidine there are 3 paths forming 1,3-Diamino-propane, 4-Aminobutanal and spermine as the intermediate metabolites. The path leading to spermine is followed which is the desired end product. The occurrence of the optimal path following the above mentioned methodology has been observed in [6,9].

5 Conclusions and Discussions

We have developed a simple constraint based method for identifying an optimal metabolic pathway that involves weighting coefficients indicating the concentration levels of enzymes catalyzing biochemical reactions in the pathway. The method can suitably be used using reaction databases without going into complex mathematical calculations, and without using various kinetic parameters that are hard to estimate. It has been found that the method is able to produce biologically more relevant results than EPA and can be applied in metabolic engineering. This method could be of a great interest for the scientific community, as current pathway identification methods, e.g. elementary flux modes and extreme pathways cannot be applied to many real life models due to their numerical complexity.

References

1. Papin, J.A., Price, N.D., Wiback, S.J., Fell, D.A., Palsson, B.O.: Metabolic pathways in the post-genome era. Trends in Biochemical Sciences 28, 250–258 (2003)
2. Lee, J.M., Gianchandani, E.P., Papin, J.A.: Flux balance analysis in the era of metabolomics. Briefings in Bioinformatics 7, 1–11 (2006)
3. Schilling, C.H., Letscher, D., Palsson, B.O.: Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. J. Theor. Biol. 203, 229–248 (2000)
4. Stryer, L., Berg, J.M., Tymoczko, J.L., Clarke, N.D. (eds.): Biochemistry. W H Freeman, New York (1998)
5. Montoya, A.R., Lee, W.P., Bassilian, S., Lim, S., Trebukhina, R.V., Kazhyna, M.V., Ciudad, C.J., Noe, V., Centelles, J.J., Cascante, M.: Pentose phosphate cycle oxidative and nonoxidative balance: A new vulnerable target for overcoming drug resistance in cancer. Int. J. Cancer 119, 2733–2741 (2006)
6. Mapelli, S., Brambilla, I.M., Radyukina, N.L., Ivanov, Y.V., Kartashov, A.V.: Free and bound polyamines changes in different plants as a consequence of uv-b light irradiation. Gen. Appl. Plant Physiology (special issue 34), 55–66 (2008)
7. Kumar, A., Altabella, T., Taylor, M.A., Tiburcio, A.F.: Recent advances in polyamine research. Trends Plant Sci. 2, 124–130 (1997)
8. Malmberg, R.L., Watson, M.B., Galloway, G.L., Yu, W.: Molecular genetic analyses of plant polyamines. Critical Reviews in Plant Sciences 17, 199–224 (1998)
9. Ponce, M., Martinez, L., Galmarini, C.: Influence of ccc, putrescine and gellam gum concentration on gynogenic embryo induction in allium cepa. Biologia Plantarum 50(3), 425–428 (2006)
10. Igoillo-Esteve, M., Maugeri, D., Stern, L., Beluardi, P., Cazzulo, M.J.: The pentose phosphate pathway in trypanosoma cruzi: a potential target for the chemotherapy of chagas disease. Annals of the Brazilian Academy of Sciences 79, 649–663 (2007)

Cross-Correlation and Evolutionary Biclustering: Extracting Gene Interaction Sub-networks

Ranajit Das[1], Sushmita Mitra[1], and Subhasis Mukhopadhyay[2]

[1] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
{ranajit_r,sushmita}@isical.ac.in

[2] Department of Bio-Physics, Molecular Biology and Bioinformatics, Calcutta University, Kolkata 700 009, India
sm.bmbg@gmail.com

Abstract. In this paper we present a simple and novel time-dependent cross-correlation-based approach for the extraction of simple gene interaction sub-networks from biclusters in temporal gene expression microarray data. Preprocessing has been employed to retain those gene interaction pairs that are strongly correlated. The methodology was applied to public-domain data sets of Yeast and the experimental results were biologically validated based on standard databases and information available in the literature.

Keywords: Biclustering, transcriptional regulatory network, time-delay, time-lagged correlation, gene interaction network.

1 Introduction

With the current development in microarray technology (gene chips), today researchers in Bioinformatics have, at their disposal, expression data of thousand of genes of different organisms under various experimental conditions. DNA microarray technology, thus, forms an indispensable tool for exploring transcriptional regulatory networks from the system level and is very helpful when one dwells into the cellular environment to investigate various complex interactions. Biological pathways can be conveniently characterized as networks and broadly classified as *metabolic pathways*, *gene regulatory networks or gene interaction networks* and *signal transduction pathways.* Gene regulatory networks connect genes, gene products (in the form of protein complexes) or their groups to one another. A network of coregulated genes may form gene clusters that can encode proteins, which interact amongst themselves and take part in common biological processes. Clustering has been applied to locate co-expressed groups of genes and extract gene interaction/gene regulatory networks from gene expression data [1].

Genes with similar expression profiles may regulate one another or be regulated by some other common parent gene. However, one need to observe that a subset of genes is co-regulated and co-expressed only over few conditions. The genes also share local rather than global similar patterns in their expression profiles [2]. Such sets of genes may be identified in the form of biclusters [3] using continuous columns, to represent a continuous interval of time [4].

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 199–204, 2009.

Evolutionary biclustering has been used for extracting gene interaction networks from time series gene expression data [5]. The networks were validated incorporating domain knowledge from transcription factor (TF)-target (T) databases like *TRANSFAC*[1] and literature [6]. Co-expressed gene pairs were seldom found to exhibit simple simultaneous relationships. On closer inspection of their profiles it could rather be noted that there exists a time shifted response of the target gene to its TF [6]. Time-lagged correlation or cross-correlation helps in analyzing the positive or negative correlation among time-lagged profiles of gene pairs. This motivated us to explore the use of cross-correlation with time-delay for the appropriate modeling of temporal interactions.

In this paper continuous-column multiobjective evolutionary biclustering [4] has been used for extracting time-correlated gene pairs. A gene pair having correlation magnitude above a detection threshold was considered to be interacting or regulating each other. Preprocessing was done to eliminate the weakly correlated (positive or negative) gene interaction pairs. An adjacency matrix was constituted from the resulting cross-correlation matrix, which was eventually used for reverse engineering transcriptional gene interaction sub-networks using regulatory information among genes. The usefulness of the model is demonstrated, using time-series gene expression data from Yeast and biologically validated.

2 Gene Interaction Sub-network Extraction: A Multi-objective Evolutionary Approach

A gene produces a protein by *transcription* (formation of many copies of mRNA molecules) followed by *translation* (resulting into production of protein), the reaction taking place in the presence of an enzyme. In turn, a protein is responsible for a particular biological function. A TF is a gene product that binds to the promoter region of a target gene, up-regulating or down-regulating its expression. Every gene has one or more such activators/repressors. Their identification, and the subsequent elucidation of the biological networks demonstrating TF-T relationship is quite a challenging task. Analysis of their profiles brings out several complex relationships between the co-regulated gene pairs, including co-expression, time shifted, and inverted relationships [6].

2.1 Evolutionary Biclustering

Biclustering refers to the simultaneous clustering and redundant feature reduction involving both attributes and samples. This results in the extraction of biologically more meaningful, less sparse partitions from high-dimensional data, that exhibit similar characteristics. A bicluster may be defined as a pair (g, c), where $g \subseteq \{1, \ldots, m\}$ denotes a subset of genes and $c \subseteq \{1, \ldots, n\}$ denotes a subset of conditions (or time points). The optimization task [3] involves finding the maximum-sized bicluster subject to a certain homogeneity constraint. The size (or volume) $f(g, c)$ of a bicluster is defined as the number of cells in the gene

[1] http://www.gene-regulation.com/pub/databases.html

expression matrix E (with values e_{ij}) that are covered by it. The homogeneity $\mathcal{G}(g, c)$ is expressed as a mean squared residue score (or error). Since these two characteristics of biclusters are conflicting to each other, multi-objective evolutionary algorithms, in association with local search, was applied to provide an alternative, more efficient approach [4].

2.2 Time-Lagged Cross-Correlation between Gene Pairs

Often genes are found to share similar sub-profiles (over a few time points) instead of the complete gene expression profiles. Considering the global correlation among gene pairs, *i.e.*, computation of correlation amongst genes employing the complete gene expression data matrix, may not reveal the proper relationship between them. Since the transcriptional response of a gene can occur from tens of minutes to several hours, time delay correlation may help determine the underlying causal relationship. The concept of cross-correlation has been introduced to take into account the time-shifted behaviour between TF-T pairs. This allows a more realistic modeling of gene interactions within the reduced localized domain of biclusters. In this work we extend our earlier network extraction algorithm [5] to include such temporal influence.

The expression profile e of a gene may be represented over a series of n time points. The cross-correlation $CC_d(e_1, e_2)$ between gene pair e_1 and e_2, with delay d,is expressed as

$$CC_d(e_1, e_2) = \frac{\sum e_{1i}e_{2i-d} - \sum e_{1i} \sum \frac{e_{2i-d}}{n}}{\sqrt{(\sum e_{1i}^2 - \frac{(\sum e_{1i})^2}{n})(\sum e_{2i-d}^2 - \frac{(\sum e_{2i-d})^2}{n})}}. \tag{1}$$

Here we select that delayed time $d = \Delta$t which maximizes the correlation in absolute value by eqn. (2) as

$$CC(e_1, e_2) = \max |CC_d(e_1, e_2)| \qquad where - 2 \leq d \leq 2 \tag{2}$$

The maximum delay of two time point are allowed as longer shifts are hard to explain from a biological point of view [2].

Filtering out the weaker correlation coefficients, which presumably contribute less towards regulation, serves as the first preprocessing step. This allows us to avoid an exhaustive search of all possible interactions among genes. The remaining coefficients, having absolute values above a detection threshold, imply a larger correlation among the gene pairs. The correlation range $[CC_{\max}, CC_{\min}]$ is divided into three partitions each, using *quantiles* or *partition values* [5] to reduce the influence of extreme values or noisy patterns. Negative correlation for a gene pair is not zero correlation. Time lagged-correlation coefficients with values greater than Q_2^+ (less than Q_2^-) indicate high positive (negative) correlation, while those with values in $[Q_1^+, Q_2^+)$ $([Q_2^-, Q_1^-))$ indicate moderate positive (negative) correlation.

An adjacency matrix is calculated as follows:

$$A(i, j) = \begin{cases} -1 \text{ if} & CC \leq Q_2^- \\ +1 \text{ if} & CC \geq Q_2^+ \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where self correlations among the genes are assumed to be absent. Next the extraction of gene interaction sub-networks is attempted. Biological validation is made in terms of ontology study.

3 Experimental Results

We applied our analysis to the Yeast cell-cycle CDC28 data gathered by Cho *et al.* [7]. It is a measure of the expression levels of 6220 gene transcripts (features/attributes) for 17 conditions (time points/samples), taken at 10-minute time intervals covering nearly two cycles. The synchronization of the yeast cell cultures was done using the so-called CDC28 arrest and the experiment was performed using Affymetrix oligonucleotide arrays. The missing values in the data set were imputed similar to that of our earlier network extraction technique [5] and the biclusters were extracted to detect sets of co-regulated genes. Pairwise time-lagged cross-correlation coefficients were calculated between gene pairs in the extracted biclusters using eqn. (1). The weaker interactions, as thresholded by quantile partitioning, were excluded.

A sample extracted sub-network comprising of four biclusters (modules or sub-networks) is depicted in Fig. 1. A transcription factor is connected to its target gene by an arrow if such a TF-Target pair existed within any of the biclusters. Gene pairs connected by solid lines depict positive correlation, while those connected by dashed lines are negatively correlated. As an example, the TF named $YHR084W$ (encircled with solid lines) belonging to the sub-network of 10 genes has targets in all the four sub-networks. These biclusters were biologically validated from gene ontology study, based on the statistically significant GO annotation database[2], and discussed below.

Fig. 2 demonstrates various complex relationships *viz.* and their interplay in the pairwise profile relationship of the TF and its target in Fig. 1. We observe that the target $YGR233C$ poses a mixed simultaneous (during 0-40 minutes) and time-shifted relationship (during 40-60 and 100-140 minutes) with the TF $YHR084W$ and that the relationship is not a simple direct one.

It has been reported during the prediction of regulatory network structure [8] that the gene pair $YHR084W$-$YGR233C$ (where both the TF and the target belongs to the 10-node network) form a TF-Target pair. We verified their GO summary in terms of *Molecular Function*, *Biological Process* and *Cellular Component* from the Saccharomyces Genome Database (SGD). Our computations also indicated an analogous interaction between the TF-Target pair and were supported by literature [9]. Based on the SGD it could be gathered that the $YHR084W$ is activated by a MAP kinase signaling cascade – that controls many important functions of living organisms like cell-cycle, apoptosis[3], differentiation, etc. while the protein $YGR233C$ has cyclin-dependent protein kinase inhibitor activity (TAS). One can think of several models considering the

[2] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

[3] Programmed cell death.

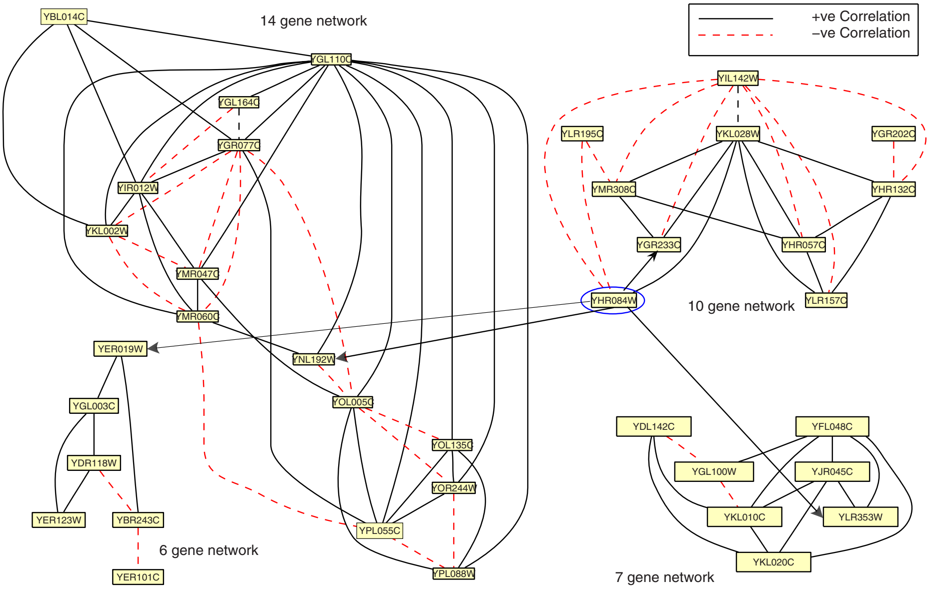


Fig. 1. Sub-network (bicluster) of 10 genes connected by transcription factor $YHR084W$ to sub-networks (biclusters) of 6, 7 and 14 genes

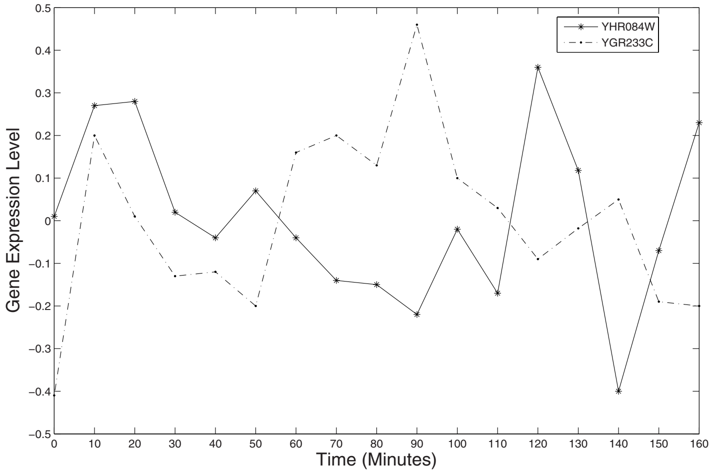


Fig. 2. Expression profile of transcription factor $YHR084W$ and its target $YGR233C$ (10-node network)

transcription of $YGR233C$ by $YHR084W$ to occur inside the nucleus, followed by the regular translation mechanism, based on their cellular component.

4 Conclusions and Discussion

In this paper we have described an approach for the extraction of cross correlated gene pairs for the generation of gene interaction networks. Biologically relevant biclusters were obtained using multiobjective biclustering, from time-series gene expression data from Yeast. The pairwise time-lagged correlation coefficients among gene pairs were computed by eqn. (1), followed by the quantile partitioning. Strongly correlated genes were considered for extracting sample TF-Target gene interaction sub-networks, as in Fig. 1. We tried to analyze the expression profiles of the regulator and the regulated genes to reveal several complex (time shifted, inverted, simultaneous, etc.) biological relationships from information available in the literature/databases. The sparsity and time-shifted behaviour between TF-T pairs in gene regulatory networks was reflected well on choosing cross-correlation as the similarity measure.

References

1. Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Transactions on Computational Biology and Bioinformatics (to appear)
2. Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kamper, J.: Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 21, 1069–1077 (2005)
3. Cheng, Y., Church, G.M.: Biclustering of gene expression data. In: Proceedings of ISMB 2000, pp. 93–103 (2000)
4. Mitra, S., Banka, H.: Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition 39, 2464–2477 (2006)
5. Mitra, S., Das, R., Banka, H., Mukhopadhyay, S.: Gene interaction - An evolutionary biclustering approach. Information Fusion 10, 242–249 (2009)
6. Qian, J., Lin, J., Luscombe, N.M., Yu, H., Gerstein, M.: Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19, 1917–1926 (2003)
7. Cho, R.J., Campbell, M.J., Winzeler, L.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
8. Yu, H., Gerstein, M.: Genomic analysis of the hierarchical structure of regulatory networks. Proceedings of National Academy of Sciences USA 103, 14724–14731 (2006)
9. Zeitlinger, J., Simon, I., Harbison, C., Hannett, N., Volkert, T., Fink, G., Young, R.: Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell 113, 395–404 (2003)

Learning Age and Gender of Blogger from Stylistic Variation

Mayur Rustagi, R. Rajendra Prasath*, Sumit Goswami, and Sudeshna Sarkar

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur,
West Bengal 721302, India
mrustagi@iitkgp.ac.in, rajendra@cse.iitkgp.ernet.in,
sgoswami@iitkgp.ac.in, sudeshna@cse.iitkgp.ernet.in

Abstract. We report results of stylistic differences in blogging for gender and age group variation. The results are based on two mutually independent features. The first feature is the use of slang words which is a new concept proposed by us for Stylistic study of bloggers. For the second feature, we have analyzed the variation in average length of sentences across various age groups and gender. These features are augmented with previous study results reported in literature for stylistic analysis. The combined feature list enhances the accuracy by a remarkable extent in predicting age and gender. These machine learning experiments were done on two separate demographically tagged blog corpus. Gender determination is more accurate than age group detection over the data spread across all ages but the accuracy of age prediction increases if we sample data with remarkable age difference.

1 Introduction

Stylistic classification can improve the results achieved through Information Retrieval (IR) techniques by identifying documents that matches a certain demographic profile. Gender and age are the common demographic features used for experimentation using stylistics as the blogs generally contain these information provided by the author. Style in writing is a result of the subconscious habit of the writer of using one form over a number of available options to present the same thing. The variation also evolves with the usage of the language in certain period, genre, situation or individuals. Variations are of two types - variation within a norm which is grammatically correct and deviation from the norm which is ungrammatical. The variations can be described in linguistic as well as statistical terms[15]. Concept and themes[20] can be determined from variations within the norm while the usage of non-dictionary words or *slang* is an example of deviation from a norm.

2 Related Work

The research in last few decades on usage of language pattern by different social groups was constrained due to unavailability of sufficient annotated data.

* Currently Rajendra is an ERCIM Fellow at IDI, NTNU, Norway.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 205–212, 2009.

Analysis of effects of bloggers age and gender from weblogs, based on usage of keywords, parts of speech and other grammatical constructs, has been presented in [2,6,19,22]. Age linked variations had been reported by Pennebaker, et al. [11], Pennebaker and Stone[14] and Burger and Henderson, 2006 [6]. J. Holmes distinguished characteristics of male and female linguistic styles [9]. Expert used spoken language [15], Palander worked on electronic communications[13], and S. Herring analyzed correspondence[18]. Simkins analysed that there are no difference between male and female writing style in formal contexts [10]. Koppel et al. estimated author's gender using the British National Corpus text [12]. By using function words and part-of-speech, Koppel et al. reported 80% accuracy for classifying author's gender. Koppel et al. also stated that female authors tend to use pronoun with high frequency, and male authors tend to use numeral and representation related numbers with high frequency. Corney et al. estimated author's gender from e-mail content [5]. In addition to function words and part-of-speech and n-grams [19,12], they used HTML tags, the number of empty lines, average length of sentences for features for SVM [4].

3 Dataset

A blog corpus[1] is made available by ICWSM 2009[1] and the blogs in this corpus did not have any tag for demographic information. However, it had the resource link which had the URL of the blogger's home page. In the above corpus, blogs from blog.myspace.com had the maximum occurrence and had the demographic details of the blogger in its home page. The home page of these URLs were crawled and processed to retrieve gender, status (married, unmarried), age, zodiac sign, city, state and country corresponding to each URL. With the available valid URL list, the downloaded data from these URLs gives 92,381 files. The blogs in which the blogger's age has been reported as below 20

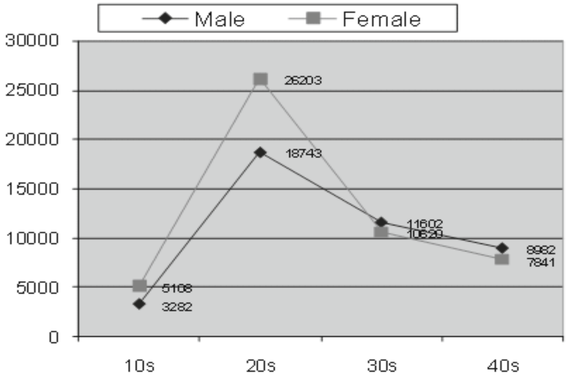


Fig. 1. Number of files in age groups and gender

[1] Provided by Spinn3r.com [3].

has been grouped in 10s age group, those in the age group of 20 to 29 as 20s, those in 30 to 39 as 30s and 40 and above has been put in 40s age group. The distribution of these files over age and gender is given in Figure 1.

4 Feature Selection and Classification Algorithm

The highest frequency words collected from a corpora may not be a good distinguishing feature. However, an analysis of the words that are highest occurring in a subcorpora can be the marker [16]. Reference to ‘attending school’ results in an instant ‘teenage’ classification. A feature may be represented by its relative frequency or by its mere presence or absence. Features for stylistics are generally based on character or morphological features or lexical features. In our experiments we used the sentence length and non-dictionary words as the features. As per our literature survey, the usage of slang word has not yet been explored for the study of stylistic variation.

Koppel[12] used a list of 30 words each as a distinguishing feature for gender and age respectively. These words were detected to be having an extreme variation in usage across gender and age groups. Similarly out-of-dictionary words were augmented to increase the accuracy of results[17]. For the purpose of learning age and gender classifier, each document is represented as a numerical vector in which each entry represent the normalized frequency of a corresponding word in the feature set. Table 1 and Table 2 lists a few content words used for learning gender and age groups respectively.

Many Stylistic results had been reported using average sentence length as a feature. Still we selected to work on this feature because, most of the reported work was on formal writing and generally on classical works of literature. Analysis of blogs based on average sentence length is challenging as blogs lack editorial and grammatical checks. Figure 2 shows the variation of average sentence length on age and gender basis.

As blogs are informal writing without any editorial bounds, blogosphere has slowly filled up with many non-dictionary words that are understandable and commonly used by online community. We refer to some of them as slangs, smiley, out of vocabulary words, chat abbreviations etc. The named entities are also non-dictionary words. There are words that are intentionally misspelled, repeated, extended or shortened to have a different effect on the reader, express emotion or save the time of blogging. All these words and even the frequency of use of such words are contributable features in stylistics. A taboo word for a particular age group can be a commonly used word for another. For our experiments with non-dictionary words, Ispell [8] was run and the frequency of all the non-dictionary words used by males and females for detecting gender variation was obtained. From these, only those words were selected as feature which had an occurrence of >50 and for which the usage among male and female was atleast double. This generated a list of 52 words. Figure 3 shows the usage of out of vocabulary words among age and gender variation and Figure 4 shows the usage frequency of a few selected Non-Dictionary words among different gender.

Table 1. List of Content word frequency per 10000 words in gender

	$\frac{\text{Male Occ}}{10000}$	$\frac{\text{Female Occ}}{10000}$
mom	4.543	7.844
microsoft	0.921	0.594
gaming	0.131	0.045
server	0.152	0.108
software	0.131	0.051
gb	0.436	0.519
programming	0.069	0.045
google	0.318	0.228
data	0.249	0.114
graphics	0.076	0.108
india	0.069	0.085
nations	0.464	0.142
democracy	0.048	0.011
users	0.159	0.102
economic	0.159	0.079
shopping	0.304	0.845
cried	0.159	0.759
freaked	0.048	0.119
pink	0.256	0.497
cute	0.671	1.662
gosh	0.083	0.182
kisses	0.096	0.217
yummy	0.069	0.091
mommy	0.027	0.314
boyfriend	0.297	1.411
skirt	0.062	0.154
adorable	0.027	0.285
husband	0.297	1.765
hubby	0.034	0.359

Table 2. List of Content word frequency per 10000 words in age groups

	$\frac{WC}{(\sum \text{WC in that age grp})\text{x}10000}$		
Word	10s age	20s age	30s age
college	4.433	1.173	0.829
maths	0	0.006	0
homework	0.299	0.126	0.078
bored	2.399	1.892	0.789
sis	3.433	4.750	4.844
boring	0.966	0.687	0.618
awesome	2.533	2.971	2.264
mum	0.499	0.277	0.329
crappy	0.266	0.283	0.289
mad	9.832	9.236	8.384
dumb	1.266	0.870	0.447
semester	1.333	0.813	0.263
apartment	0.599	1.205	0.487
drunk	0.799	1.318	0.974
beer	0.466	0.826	0.908
student	0.766	0.504	0.855
album	0.966	1.463	1.684
someday	0.199	0.302	0.184
dating	0.699	0.889	0.710
bar	3.733	3.470	3.922
marriage	0.133	0.403	0.394
development	0.099	0.176	0.171
campaign	0.033	0.258	0.605
tax	0.066	0.391	0.539
local	0.499	0.706	1.803
democratic	0.033	0.044	80.10
son	30.26	28.80	28.55
systems	0	0.050	0.105
provide	0.433	0.378	0.552
workers	0.099	0.233	0.394

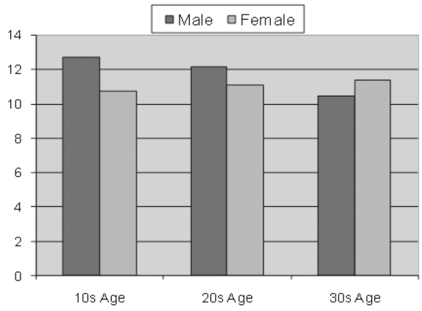


Fig. 2. Average Sentence length on Gender Age Basis

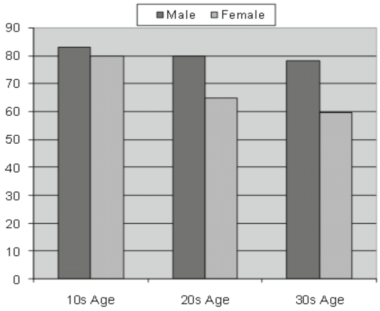


Fig. 3. Out-of-dictionary words used per 1000 words across various age groups

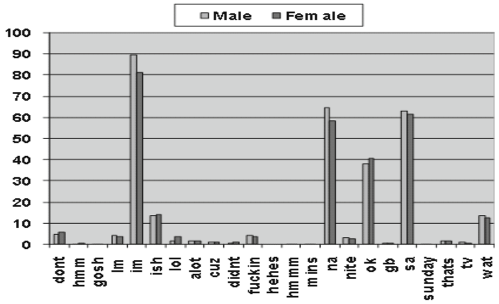


Fig. 4. Measure of usage of a few selected Non-dictionary words used by Males and Females

Naive Bayes classifier for predicting the blogger's age group or gender from the stylistic features were trained using the WEKA toolkit [21]. During training, classifiers are created by the selection of a set of variables for each feature and classifier parameters are tuned through cross-validation. To evaluate the classifier, the given data is split into training and test data and the trained classifier is used to predict the blogger's demographic profile on the test data [7].

5 Results and Discussion

Analysis of Figure 3 tells that teenagers generally use more out-of-dictionary words than the adults. Here, we call those words as non-dictionary which are not available in the Ispell ver. 3.1.20. These words can be the slang, exclamatory word with a small or large repetition of last character like 'soooo sweet' or typing errors due to less concern towards spelling and grammar or idiosyncrasies. Though, the number of slang words used in text can be a remarkable feature but a single feature can not make a good classifier. To build a classifier for age variation, we initially took only those bloggers who are in their 10s and those who are in their 30s so that there is a remarkable difference between their usage of non-dictionary word pattern and thus simpler to classify. For our experiments with non-dictionary words, we selected the list of 52 non-dictionary words.

Table 3. Confusion matrix for the gender classification using 52 non-dictionary words as features

a	b	← classified as
42609	0	a = male
15147	34608	b = female

Naive Bayes Classifier yielded an accuracy of 83.6% for gender based classification and 95% accuracy for the age group classification between 10s and 30s age. We did not measure the percent accuracy for age group classification between 10s, 20s and 30s due to similarity of style in overlapping age groups.

5.1 Average Sentence Length

Since the average sentence length is a remarkable feature, we used this feature in combination with slang words reported above and the interesting content words reported on this corpus in [19]. The classification results and Figure 2 are not sufficient to interpret that the average sentence length in a persons writing increases with age. The collected blogposts had been written across a span of 5 years and is not sufficient to predict this trend. The trend of increase in the average sentence length with age can be tested only if we have sufficient blog data in which the person had been blogging for a few decades so as to look into the trend of change in average sentence length with his age. It may happen that the average sentence length in English writing is decreasing with time. The bloggers of today may continue blogging at the same average sentence length but those who start blogging after ten years may use smaller sentence lengths.

Table 4. Detailed Accuracy By Class gender detection using out-of-dictionary words only

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1.000	0.304	0.738	1.000	0.849	0.925	male
0.696	0.000	1.000	0.696	0.820	0.925	female

Table 5. Confusion matrix for 10s and 30s age group classification using 52 non-dictionary words

a	b	← classified as
6890	1498	a = 10s age
3	22219	b = 30s age

5.2 Augmented Features

The gender and age experiments were conducted initially only on 35 content words and it gave an accuracy of 95.2% for gender classification, and 92.51% for age classification (refer to Table 3).

The age experiments were run on four categories of age group considered above: 10s, 20s, 30s and higher. The feature list comprised of 35 content words reported in [19] combined with 52 slang words mined by us from blog data based on our acceptance index. [19] has reported an accuracy of 92.51% with the content words. The augmented feature list yielded an accuracy of 94.13%.

Table 6. Confusion matrix for the gender classification using 35 Content words as features

a	b	← classified as
42571	0	a = male
4451	45262	b = female

Table 7. Detailed Accuracy By Class gender detection using content words only

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.09	0.905	1	0.95	1	male
0.91	0	1	0.91	0.953	1	female

Table 8. Confusion matrix for the gender classifier using 52 slang word and 35 content words

a	b	← classified as
30931	0	a = male
1731	34191	b = female

Addition of average sentence length to this set of features did not contribute to a significant amount.

Similarly, experimentation was done for gender variation after augmenting the 35 content word feature reported in [19] with our 52 slang words. Schler *et* al.[19] have reported an accuracy of 80.1% in gender determination on their dataset and we received an accuracy of 83.6% on ICWSM dataset. However, our augmented feature list gave an accuracy of 97.41%, the confusion matrix of which is given in Table 8. After further enhancement of this augmented feature list with average sentence length, there was not much increase in the accuracy and so is not reported here.

6 Conclusion and Future Work

Teenage bloggers use more out-of-dictionary words than the adult bloggers. Furthermore, for bloggers of each gender, there is a clear distinction between usage of a few slangs. Generally in their present age, teenager use smaller sentences compared to the adult bloggers but we found a variation to this in this dataset. With the available data and the existing experiments, it cannot be confirmed that the average sentence length increases or decreases with age. The stylistic difference in usage of slang predicts the age and gender variation with certain accuracy. Average sentence length itself is not a good feature to predict the variation as there is a wide variation in sentence length in informal writing. However, the feature of average sentence length can be augmented with slang to slightly increase its prediction efficiency. Both these features when augmented with other features like content words reported earlier, increases the prediction accuracy by a good amount.

The usage of slang can also be a good feature to predict the geographical location or the ethnic group of the user due to the heavy usage of a particular out-of-dictionary word or named entities at certain regions. A sufficiently huge corpus collected over a decade will be useful to study the variation of sentence length of users with age and variations in individuals language use over the course of their lives. This corpus can also be used to study the evolution and death of the slang words with time.

References

1. ICWSM 2009, Spinn3r Dataset (May 2009)
2. Argamon, S., Koppel, M., Avneri, G.: Routing documents according to style. In: Proc. of First Int. Workshop on Innovative Inform. Syst. (1998)
3. Spinn3r Indexing Blogosphere, www.spinn3r.com (last accessed on March 01, 2009)
4. Brank, J., Grobelnik, M., Milic-Frayling, N., Mladenic, D.: Feature selection using support vector machines. In: Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Eng., Finance, and Other Fields, pp. 84–89 (2002)
5. Corney, M., de Vel, O., Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: 18th Annual Computer Security Appln. Conference (2002)
6. Burger, J.D., Henderson, J.C.: An exploration of observable features related to blogger age. In: Proc. of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (2006)
7. Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Tat: an author profiling tool with application to arabic emails. In: Proc. of the Australasian Language Technology Workshop, pp. 21–30 (2007)
8. Ispell (2009), http://www.gnu.org/software/ispell/ (last accessed on April 02, 2009)
9. Holmes, J.: Women's talk: The question of sociolinguistic universals. Australian Journal of Communications 20(3) (1993)
10. Simkins-Bullock, J., Wildman, B.: An investigation into relationship between gender and language Sex Roles 24. Springer, Netherlands (1991)
11. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count. In: LIWC 2001 (2001)
12. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
13. Palander-Collin, M.: Male and female styles in 17th century correspondence: I think. Language Variation and Change 11, 123–141 (1999)
14. Pennebaker, J.W., Stone, L.D.: Words of wisdom: Language use over the lifespan. Journal of Personality and Social Psychology 85, 291–301 (2003)
15. McMenamin, G.R.: Forensic Linguistics: Advances in Forensic Stylistic. CRC Press, Boca Raton (2002)
16. Datta, S., Sarkar, S.: A comparative study of statistical features of language in blogs-vs-splogs. In: AND 2008: Proc. of the second workshop on Analytics for noisy unstructured text data, pp. 63–66. ACM, New York (2008)
17. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. To appear in: Proc. of ICWSM (2009)
18. Herring, S.: Two variants of an electronic message schema. In: Herring, S. (ed.) Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives, vol. 11, pp. 81–106 (1996)
19. Argamon, S., Schler, J., Koppel, M., Pennebaker, J.: Effects of age and gender on blogging. In: Proc. of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs (April 2006)
20. Leximancer Manual V.3, www.leximancer.com (last accessed on January 22, 2009)
21. Witten, I.H., Frank, E.: DataMining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
22. Yan, R.: Gender classification of weblog authors with bayesian analysis. In: Proc. of the AAAI Spring Symp. on Computational Approaches to Analyzing Weblogs (2006)

Hypertext Classification Using Tensor Space Model and Rough Set Based Ensemble Classifier

Suman Saha, C.A. Murthy, and Sankar K. Pal

Center for Soft Computing Research, Indian Statistical Institute
{ssaha_r,murthy,sankar}@isical.ac.in

Abstract. As WWW grows at an increasing speed, a classifier targeted at hypertext has become in high demand. While document categorization is quite a mature, the issue of utilizing hypertext structure and hyperlinks has been relatively unexplored. In this paper, we introduce tensor space model for representing hypertext documents. We exploit the local-structure and neighborhood recommendation encapsulated in the proposed representation model. Instead of using the text on a page for representing features in a vector space model, we have used features on the page and neighborhood features to represent a hypertext document in a tensor space model. Tensor similarity measure is defined. We have demonstrated the use of rough set based ensemble classifier on proposed tensor space model. Experimental results of classification obtained by using our method outperform existing hypertext classification techniques.

Keywords: Hypertext classification, tensor space model, rough ensemble classifier.

1 Introduction

As the Web is expanding, where most Web pages are connected with hyperlinks, the role of automatic categorization of hypertext is becoming more and more important [1]. This is the case especially with the limitation of the retrieval engines; too much information to be searched and too much information retrieved. By categorizing documents a priori, the search space can be reduced dramatically and the quality of search result improved. Besides, Web users often prefer navigating through search directories as in portal sites.

In this article we have proposed a novel tensor space model for hypertext representation. Our model relies on different types of features, which are extracted from a hypertext document and its neighbors. The proposed model consists of a second order tensor for each hypertext document and a tensor component for each of the different types of features. In each tensor components a base level has been performed. The different types of classifications have been combined using rough set based ensemble classifier [2].

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 213–218, 2009.

2 Hypertext Features

A hypertext document consists of different types of features which are found to be useful for representing a web page [3]. Written in HTML, web pages contain additional informations other than text content, such as HTML tags, hyperlinks and anchor text (Fig 1). These features can be divided into two broad classes: on-page features, which are directly located on the page to be represented, and features of neighbors, which are found on the pages related in some way with the page to be represented.

Most commonly used on-page features are URL of the web page, outgoing links of web page, HTML tags, title-headers and text body content of the web page.

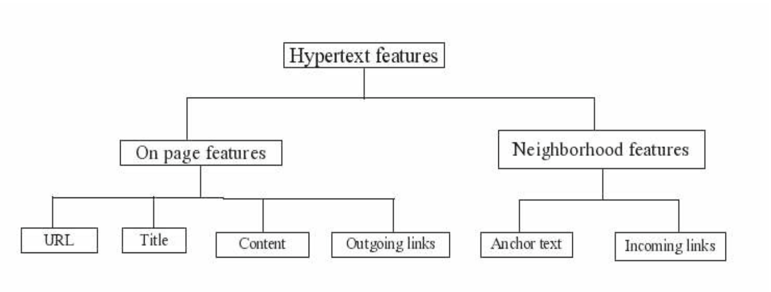


Fig. 1. Different type of features of hypertext document

3 Proposed Method

In this article we propose tag based tensor space model for the representation of hypertext documents and Rough Set based approach for its classification [2]. Splits of the features has been performed based on the different types of features existing in the hypertext documents. Tensor space model has been used to represent the hypertexts using the information of text content, internal mark-up structure and link structure. Classification of hypertext documents, represented as tensor, can be obtained in two ways by integrating classifier's parameters of different tensor components and by integrating classifiers output obtained on different tensor components. In this article two classification methods have been discussed using the two ways mentioned above. In the first method, a tensor similarity measure has been proposed. K-NN classification has been performed using tensor similarity measure. In the second method ensemble classification has been performed. For ensemble classification, base level classification has been performed on individual tensor components and combined classification has been obtained using rough set based ensemble classifier.

3.1 Preprocessing and TSM Representation

Hypertext documents are tokenized with syntactic rules and canonical forms. First we select a set of relevant features from a HTML document. For each type of feature an individual tensor component is constructed. A tensor component is a vector, which represents the terms of particular type corresponding to the component. The tensor space model captures the structural representation of hypertext document [4].

1) Preprocessing text content:

- The text is stemmed using Porter's stemming algorithm and stop words are removed.
- Unique words present in the text are represented as a tensor component. This tensor component corresponds to the text contents of the hypertext documents.

2) Preprocessing URL:

- A URL is first divided to yield a baseline segmentation into its components as given by the URI protocol (e.g., scheme :// host / path elements / document . extension), and further segmented wherever one or more non-alphanumeric characters appear.
- These segmented substrings are treated as words. All these words found in a URL will be represented as a tensor component corresponding to features of URLs.

3) Preprocessing anchor text:

- Anchor text is a small text content. The text is stemmed using Porter's stemming algorithm and stop words are removed.
- It is computed the same way as text content, except substituting each document by a virtual document consisting of all the anchor text inside that document.
- Unique words present in this virtual document are represented as a tensor component corresponding to features of anchor text.

4) Preprocessing title and headers:

- Title and headers are text contents. The text is stemmed using Porter's stemming algorithm and stop words are removed.
- Unique words present in these text are represented as a tensor component corresponding to features of title and headers.

5) Preprocessing in-links:

- All in-links are first divided to yield a baseline segmentation into its components as given by the URI protocol (e.g., scheme :// host / path elements / document . extension), and further segmented wherever one or more non-alphanumeric characters appear.
- The tokens obtained by segmentations of the in-links are stored in a tensor component corresponding to features of in-links.

6) Preprocessing out-links:

- All out-links are first divided to yield a baseline segmentation into its components as given by the URI protocol and further segmented wherever one or more non-alphanumeric characters appear.
- The tokens obtained by segmentations of the out-links are stored in a tensor component corresponding to features of out-links.

3.2 Ensemble Classifications on Tensor Space

We now describe how we generate partitions for each one of the components of the tensor using classifiers. To generate the initial partitions for RSM we assumes a base level classifier and train it on each different tensor component. This trained classifiers provide partitions on the tensor space. Outputs of the base classifiers and the actual class information are used to construct meta level decision table. This meta data represented in the form of decision table is the input of rough set based ensemble classifier. Unlike word vector representation of web documents meta data has a simple brief format, where classifier in the ensemble contribute the existence of an attribute, values of this attribute can be any class level that is determined by the base classifier corresponding to the tensor component. So the number of attributes is the same as number of tensor components. Rough set based attribute reduction techniques eliminate superfluous attributes and create a minimal sufficient subset of attributes for a decision table. Such minimal sufficient subset of attributes, called a reduct. Once the reduct is computed we remove redundant classifiers from the ensemble and construct new reduced decision table. Rough set based decision rules extracted from this reduced decision table are applied to obtain final classification.

4 Experimental Results

We performed a large number of experiments to test the output of RSM. We now describe the corpora, methodology, and results.

4.1 Data Sets

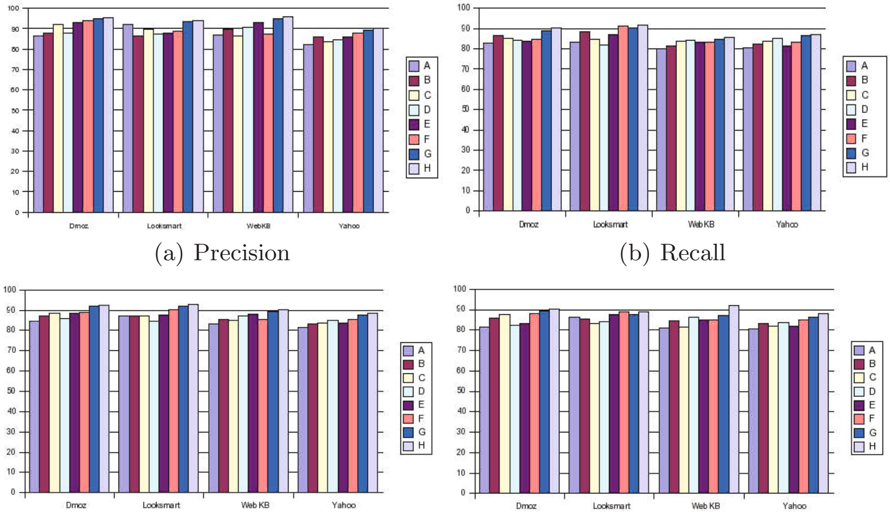
We used four data set, Looksmart, Dmoz, webkb and Yahoo for our experiments. We crawled the Looksmart and Dmoz web directories. These directories are well known for maintaining a categorized hypertext documents. The web directories are multi-level tree-structured hierarchy. The top level of the tree, which is the first level below the root of the tree, contains 13 categories in Looksmart and 16 categories for Dmoz. Each of these categories contains sub-categories that are placed in the second level below the root. We use the top-level categories to label the web pages in our experiments. The webkb data set was collected from the WebKB project. The pages in the WebKB dataset are classified into one of the categories Student, Course, Department, Faculty, Project, Stuff and

Other. Here there are 8077 documents in 7 categories. The largest category (Other) consists of 3025 pages; while the smallest category (Stuff) consists of only 135 pages. Another data set consists of 40000 Web pages crawled from the Yahoo topic directory (http://dir.yahoo.com). This is a big hypertext corpora, manually classified by the human experts. The extracted subset includes 33253 pages, which are distributed among 14 top level categories. The largest category (Science) consists of 4627 pages; while the smallest category (Regional) consists of only 782 pages. We processed the data sets to remove image and scripts followed by stop-words removal and stemming. Link graph has been constructed for each of the datasets for extracting neighborhood features. URLs has been segmented for extracting URL features. Finally features extracted from all the components of hypertext has been represented using tensor space model and vector space model for our experiments.

4.2 Classification Results on TSM

We have compared the performance of the proposed methods with existing classification techniques. The following methods are considered for comparisons.

A) Enhanced hypertext categorization using hyperlinks[5], *B*) Improving A Page Classifier with Anchor Extraction and Link Analysis [6], *C*) Fast webpage classification using URL features [7], *D*) Link-Local Features for Hypertext Classification [8], *E*) Graph based Text Classification: Learn from Your Neighbors [9], *F*) Web Page Classification with Heterogeneous Data Fusion [10], *G*) Tensor space model for hypertext representation [4] and *H*) The proposed technique.



(a) Precision (b) Recall

(c) Micro F_1 (d) Macro F_1

Fig. 2. Comparison of RSM with other hypertext classification methods

We have compared our method with other hypertext classification algorithms. Results in terms of precision, recall, micro-F_1 and macro-F_1 of A, B, C, D, E, F, G and H have been shown in figure 2. It can be observed that performance of the proposed methods are better than others in terms of precision, recall, micro-F_1 and macro-F_1.

5 Conclusion

We proposed a methodology for representing hypertext documents in tensor space model. The proposed model consists of a second order tensor for each hypertext document and a tensor component for each of the different types of feature. In this representation the features extracted from URL or Title is assigned in different tensor components. Base level classification has been performed on individual tensor components. Finally combined classification has been obtained by using rough set based ensemble classifier. Two step improvement on the existing classification results of web services has been shown. In the first step we achieve better classification results by using proposed tensor space model. In the second step further improvement of the results has been obtained by using Rough set based ensemble classifier.

References

1. Yang, Y., Slattery, S., Ghani, R.: A study of approaches to hypertext categorization. Journal of Intelligent Information Systems 18(2-3), 219–241 (2002)
2. Saha, S., Murthy, C.A., Pal, S.K.: Rough set based ensemble classifier for web page classification. Fundamentae Informetica 76(1-2), 171–187 (2007)
3. Furnkranz, J.: Web mining. The Data Mining and Knowledge Discovery Handbook, pp. 899–920. Springer, Heidelberg (2005)
4. Saha, S., Murthy, C.A., Pal, S.K.: Tensor space model for hypertext representation. In: ICIT 2008: Proceedings of the 2008 International Conference on Information Technology, pp. 261–266. IEEE Computer Society, Los Alamitos (2008)
5. Chakrabarti, S., Dom, B., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In: SIGMOD 1998, pp. 307–318. ACM, New York (1998)
6. Cohen, W.: Improving a page classifier with anchor extraction and link analysis (2002)
7. Kan, M.Y., Thi, H.O.N.: Fast webpage classification using url features. In: CIKM 2005: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 325–326. ACM, New York (2005)
8. Utard, H., Fürnkranz, J.: Link-local features for hypertext classification. In: Ackermann, M., Berendt, B., Grobelnik, M., Hotho, A., Mladenič, D., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., van Someren, M. (eds.) EWMF 2005 and KDO 2005. LNCS (LNAI), vol. 4289, pp. 51–64. Springer, Heidelberg (2006)
9. Angelova, R., Weikum, G.: Graph-based text classification: learn from your neighbors. In: SIGIR 2006, pp. 485–492. ACM, New York (2006)
10. Xu, Z., King, I., Lyu, M.R.: Web page classification with heterogeneous data fusion. In: WWW 2007, pp. 1171–1172. ACM, New York (2007)

Feature and Opinion Mining for Customer Review Summarization

Muhammad Abulaish[1,*], Jahiruddin[1], Mohammad Najmud Doja[2], and Tanvir Ahmad[2]

[1] Department of Computer Science, Jamia Millia Islamia, New Delhi, India
abulaish@ieee.org, jahir.jmi@gmail.com
[2] Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India
ndoja@yahoo.com, tanvir.ce@jmi.ac.in

Abstract. In this paper, we present an opinion mining system to identify product features and opinions from review documents. The features and opinions are extracted using semantic and linguistic analysis of text documents. The polarity of opinion sentences is established using polarity scores of the opinion words through Senti-WordNet to generate a feature-based summary of review documents. The system is also integrated with a visualization module to present feature-based summary of review documents in a comprehendible way.

Keywords: Opinion mining, Opinion analysis, Sentiment analysis, Text mining, Review summarization, Natural language processing.

1 Introduction

In recent past, due to existence of numerous forums, discussion groups, and blogs, individual users are participating more actively and are generating vast amount of new data – termed as *user-generated contents*. These new Web contents include customer reviews and blogs that express opinions on products and services – which are collectively referred to as customer feedback data on the Web. As customer feedback on the Web influences other customer's decisions, these feedbacks have become an important source of information for businesses to take into account when developing marketing and product development plans.

Recent works have shown that the distribution of an overwhelming majority of reviews posted in online markets is bimodal. Reviews are either allotted an extremely high rating or an extremely low rating. In such situations, the average numerical star rating assigned to a product may not convey a lot of information to a prospective buyer. Instead, the reader has to read the actual reviews to examine which of the positive and which of the negative aspect of the product are of interest. Several sentiment analysis approaches have proposed to tackle this challenge up to some extent. However, most of the classical sentiment analysis mapping the customer reviews into binary classes – *positive* or *negative*, fails to identify the product features liked or disliked by the customers.

[*] To whom correspondence should be addressed.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 219–224, 2009.

In this paper, we present an opinion mining system which uses linguistic and semantic analysis of text to identify key information components from text documents. The information components are centered on both product features, and associated opinions, which are extracted using natural language processing techniques and co-occurrence-based analysis. The novelty of the system lies in mining associated modifiers with opinions to represent the degree of expressiveness of opinions. For each extracted feature, the list of opinions and associated modifiers are compiled and their polarity is established using numerical scores obtained through Senti-WordNet [8]. We also present a visualization technique that provides a feature-based summary of review documents in a graphical way. The feature-based summary can help the customers as well as manufacturers to know about the positive and negative aspects of the products without going through pile of documents.

The remaining paper is structured as follows: Section 2 presents related works on opinion mining. Section 3 presents the architectural details of proposed opinion mining system. The evaluation of the feature and opinion extraction process is presented in section 4. Finally, section 5 concludes the paper with possible enhancements to the proposed system.

2 Related Work

Research on opinion mining started with identifying opinion bearing words, e.g., *great*, *amazing*, *wonderful*, *bad*, *poor* etc. Many researchers have worked on mining such words and identifying their semantic orientations. In [3], a bootstrapping approach is proposed, which uses a small set of given seed opinion words to find their synonyms and antonyms in WordNet. The history of the phrase *sentiment analysis* parallels that of *opinion mining* in certain respects. A sizeable number of papers mentioning *sentiment analysis* focus on the specific application of classifying customer reviews as to their polarity – *positive* or *negative* [4,6]. Although, classical sentiment classification attempts to assign the review documents either positive or negative class, it fails to find what the reviewer or opinion holder likes or dislikes. To obtain detailed aspects, feature-based opinion mining is proposed in literature [1,3,5]. In [1], a supervised pattern mining method is proposed. In [3,5], an unsupervised method is used. A lexicon-based approach has been shown to perform quite well in [2,3]. The lexicon-based approach basically uses opinion words and phrases in a sentence to determine the orientation of an opinion on a feature.

Although, some opinion mining methods extract features and opinions from document corpora, most of them do not explicitly exploit the semantic relationships between them. The proposed method differs from all these approaches predominantly in its use of pure linguistic techniques to identify only those features for which customers have commented using opinionated words. Moreover, extraction of associated modifiers used in review documents to represent the degree of expressiveness of opinions is unique in our work.

3 Proposed Opinion Mining System

Fig. 1 presents the architectural details of the proposed opinion mining system, which consists of five major modules – *Document Processor*, *Subjectivity/ Objectivity*

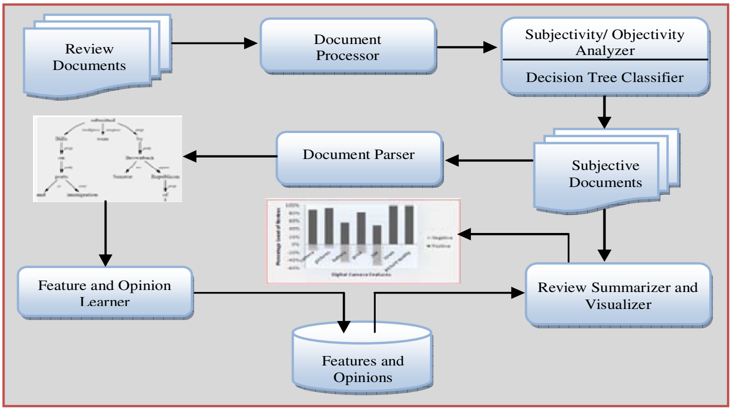


Fig. 1. Architecture of the proposed opinion mining system

Analyzer, *Document Parser*, *Feature and Opinion Learner*, and *Review Summarizer and Visualizer*. The working principles of these components are explained in the following sub-sections.

3.1 Document Processor and Subjectivity/Objectivity Analyzer

Subjective sentences are expressive of the reviewer's sentiment about the product, and objective sentences do not have any direct or obvious bearing on or support of that sentiment [7]. Therefore, the idea of subjectivity analysis is used to retain segments (sentences) of a review that are more subjective in nature and filter out those that are more objective. This increases the system performance both in terms of *efficiency* and *accuracy*. We employ the *Document Processor* which consists of a Markup Language (ML) tag filter, divides an unstructured web document into individual record-size chunks, cleans them by removing ML tags, and presents them as individual unstructured record documents for further processing.

The cleaned documents are converted into numeric-vectors using unigram model for the purpose of subjectivity/objectivity analysis. In document vectors a value represents the likelihood of each word being in a subjective or objective sentence. We have used a corpus of subjective and objective sentences described in [7] for training purpose. The training set is used to get the probability for each word to be subjective or objective. The Decision Tree classifier of Weka[1] is trained to classify the unseen review sentences into subjective and objective classes.

3.2 Document Parser, and Feature and Opinion Learner

The *Document Parser* module uses Stanford parser, which assigns Parts-Of-Speech (POS) tags to every words based on the context in which they appear. The POS information is used to locate different types of information of interest inside text documents. For example, generally noun phrases correspond to product features, adjectives

[1] http://www.cs.waikato.ac.nz/~ml/weka/

represent opinions, and adverbs are used as modifiers to represent the degree of expressiveness of opinions. Since, it is observed that opinion words and product features are not independent of each other rather, each sentence is also converted into dependency tree using the parser. The dependency tree, also known as word-word relationship, encodes the grammatical relations between every pair of words.

The *Feature and Opinion Learner* module is responsible to extract feasible information components from review documents which is analyzed further to identify product features and opinions. It takes the *dependency tree* input and output feasible information components after analyzing noun phrases and the associated adjectives possibly preceded with adverbs. On observation, we found that product features are generally noun phrases and opinions are either only adjectives or adjectives preceded by adverbs. Therefore, we have defined information component as a triplet $<\mathcal{F}, \mathcal{M}, \mathcal{O}>$ where, $\mathcal{F}$ is a noun phrase, $\mathcal{O}$ is adjective possibly representing product feature and $\mathcal{M}$ is adverb that acts as modifier to represent the degree of expressiveness of $\mathcal{O}$. $\mathcal{M}$ is also used to capture negative opinions explicitly expressed in reviews. The information component extraction mechanism is implemented as a rule-based system which analyzes dependency tree to extract information components.

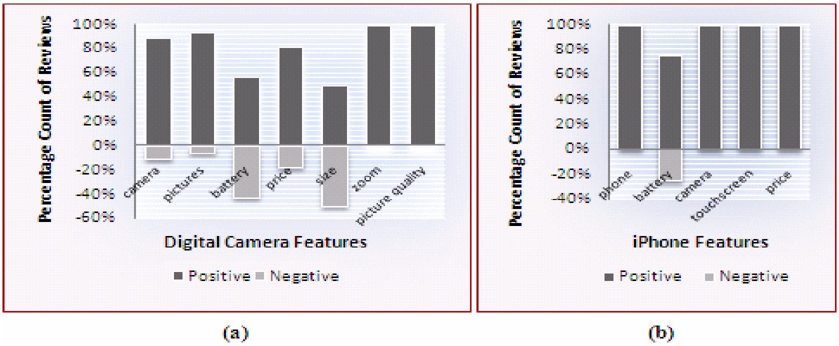
Though a large number of commonly occurring noun and adjective phrases are eliminated due to the design of the information component itself, it is found that further processing is necessary to consolidate the final list of information components and thereby the product features and opinions. During the consolidation process, we take care of two things. In the first stage, since product features are the key noun phrases on which opinions are applied, so a feasible collection of product features is identified using term frequency (*tf*) and inverse document frequency (*idf*). In the second stage of analysis, however, for each product feature the list of all opinions and modifiers is compiled that are used later for polarity determination of the opinion sentences. A partial list of product features, opinions, and modifiers extracted from a corpus of 286 customer reviews on *digital camera* is shown in table 1.

Table 1. A partial list of extracted features, opinions and modifiers for digital camera

Product	Feature	Modifier	Opinion
Digital Camera	picture	not, really, very	beautiful, clear, fantastic, good, great, professional, sharp
	battery	very	decent, excellent, rechargeable, short, long
	price	---	cheap, excellent, good, great, high

3.3 Review Summarizer and Visualizer

In order to generate feature-based summary of review documents, firstly, the polarity of extracted opinions for each feature are classified using Senti-WordNet [8], a lexical resource in which each WordNet synset s is associated to three numerical scores $\mathcal{O}bj(s)$, $\mathcal{P}os(s)$ and $\mathcal{N}eg(s)$, describing how objective, positive, and negative the terms contained in the synset are. For each feature, the opinion sentences are examined and mapped into one of the *positive* or *negative* class based on the maximum score value of the opinions present in them. In case of presence of multiple features in an opinion



(a) (b)

Fig. 2. A feature-based summary generated by the proposed opinion mining system for (a) Digital camera, and (b) iPhone

sentence, the one having highest score value is used to decide its class. Finally, the total number of positive, and negative opinion sentences for each feature is calculated to generate a feature-based review summary which is presented to user in a graphical way as shown in Fig. 2.

4 Evaluation

Since terminology and complex proper names are not found in Dictionaries, an obvious problem of any automatic method for concept extraction is to provide objective performance evaluation. Therefore manual evaluation has been performed to judge the overall performance of the proposed system. From the extraction results, the value of performance measures frequently used for information retrieval tasks - *precision*, *recall*, *F1-measure* and *accuracy* is calculated for each category of experimental data. Table 2 summarizes the performance measure values for our system. The recall value is lower than precision indicating that certain correct feature-opinion pairs could not be recognized by the system correctly. This is justified since most of the reviewers do not follow grammatical rules strictly while writing reviews due to which the parser fails to assign correct POS tag and thereby correct dependency relations between words. However, almost all identified feature-concept pairs are correct, which leaves scope for enhancing our grammar to accommodate more dependency relations.

Table 2. Performance evaluation of feature-opinion extraction process

Product Name		**TP**	**FP**	**FN**	**TN**	**Precision (%)**	**Recall (%)**	**F1-measure (%)**	**Accuracy (%)**
Digital Camera	**Canon**	37	03	27	320	92.50	57.81	71.15	93.46
	Kodak	55	03	73	365	94.83	42.97	59.14	85.02
	Nikon	44	04	63	390	91.67	41.12	56.77	87.85
	Panasonic	32	03	18	155	91.43	64.00	75.29	89.90
iPhone		23	04	14	185	85.19	48.94	62.16	88.14
					Macro-Average	**91.12**	**50.97**	**64.90**	**88.87**

5 Conclusion and Future Work

In this paper, an opinion mining system is proposed to identify product features and opinions from review documents. The proposed method also finds the sentiment polarity of opinion sentences using Senti-WordNet and provides feature-based review summarization and visualization. Presently, we are refining the rule-set to consider more relations to improve the *accuracy* of the system. We are developing a query-answering system to handle opinion-based queries over review documents.

References

1. Liu, B., Hu, M., Cheng, J.: Opinion Observer - Analyzing and Comparing Opinions on the Web. In: Proceedings of the 14th International Conference on World Wide Web (WWW 2005), Japan, pp. 342–351 (2005)
2. Ding, X., Liu, B., Philip, S.Y.: A Holistic Lexicon-Based Approach to Opinion Mining. In: Proceedings of the 1st ACM International Conference on Web Search and Data Mining (WSDM 2008), California, USA, pp. 231–240 (2008)
3. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004), USA, pp. 168–177 (2004)
4. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), USA, pp. 79–86 (2002)
5. Popescu, A.M., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP 2005), Canada, pp. 339–346 (2005)
6. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), Philadelphia, Pennsylvania, pp. 417–424 (2002)
7. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: Proceedings of ACL 2004, pp. 271–278 (2004)
8. Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of 5th Conference on Language Resources and Evaluation, Genova, Italy, pp. 417–422 (2006)

A Semi-supervised Approach for Maximum Entropy Based Hindi Named Entity Recognition

Sujan Kumar Saha, Pabitra Mitra, and Sudeshna Sarkar

Indian Institute of Technology, Kharagpur, India - 721302
{sujan.kr.saha,shudeshna,pabitra}@gmail.com

Abstract. Scarcity of annotated data is a challenge in building high performance named entity recognition (NER) systems in resource poor languages. We use a semi-supervised approach which uses a small annotated corpus and a large raw corpus for the Hindi NER task using maximum entropy classifier. A novel statistical annotation confidence measure is proposed for the purpose. The confidence measure is used in selective sampling based semi-supervised NER. Also a prior modulation of maximum entropy classifier is used where the annotation confidence values are used as 'prior weight'. The superiority of the proposed technique over baseline classifier is demonstrated extensively through experiments.

1 Introduction

Machine learning based approaches are commonly used for the development of named entity recognition (NER) systems (Borthwick 1999, Li and McCallum 2004, Saha et al. 2008). In this approach a classifier is trained using the annotated data with a suitable set of features. The performance of such statistical classifier largely depends on the amount of annotated data. But in many languages and domains sufficient annotated data do not exist. Manual creation of a sufficiently large named entity (NE) annotated corpus is costly and time consuming. Semi-supervised learning (SSL) may be adopted in such cases (Collins and Singer 1999, Mohit and Hwa 2005). SSL techniques use a limited annotated corpus along with a large raw corpus.

In this paper we propose a semi-supervised approach to named entity recognition and applied it on the Hindi NER task. A maximum entropy (MaxEnt) based classifier (baseline) is trained using a training corpus and a set of suitable features. The amount of training data is not sufficient and the baseline classifier suffers from poor recall. In order to improve the recall of the classifier we have used a large raw corpus for SSL.

A novel statistical confidence measure specific to the NER task is proposed for the purpose. A large raw corpus is annotated using the baseline classifier and the confidence weight (between 0 to 1) of the baseline annotation (for each word) is computed. The high confident part of the corpus is then selected and added to the original training data. A new classifier is built using the 'extended corpus'. In addition, we have modified the classifier in such a way that it accepts a 'prior

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 225–230, 2009.

confidence weight' corresponding to each training sample. Each annotated word in the manually created training corpus is assigned a prior confidence of *one* and the samples in the baseline system annotated corpus is assigned a confidence weight in [0-1] using the proposed algorithm. The merged corpus is given to the modified classifier along with the prior confidence weights of the samples. Experimental results show that the classifiers built using the SSL techniques achieve a considerable amount of performance improvement over the baseline.

2 Computing Confidence Weights for NER Task

We present here the procedure for computing the confidence weight of the output obtained from a NE classifier. This method assigns a weight, ranging between 0 to 1, to each word and its annotation in the corpus. The procedure uses word based statistics derived from the training data. The details are presented below.

The confidence weight is computed using a word window of length p+q+1, containing p previous words and q next words of a target word. For each of the p+q+1 positions, we compute a class specific weight corresponding to the words in the lexicon. For each word (w) in the corpus the class specific word weight, $Wt_{\{C,\ pos\}}(w)$, is defined as,

$$Wt_{\{C,\ pos\}}(w) = \frac{\#\ occurrence\ of\ \text{`w'}\ in\ position\ \text{`pos'}\ of\ a\ NE\ of\ class\ \text{`C'}}{\#\ total\ occurrence\ of\ \text{`w'}\ in\ the\ corpus} \tag{1}$$

where *pos* denotes a particular position of the p+q+1 window. Using this equation we calculate the word weights for each position in the window for the NE-class predicted by the classifier. The weighted sum of these word-wights can be used as confidence weight.

During the computation of the weighted sum, the current position (i.e. the target word) is given the highest weight. We assign 50% of the total weight of the word window to the current word. The rest 50% weight is equally divided to the p previous positions and q next positions. That is the p previous words share a total of 25% weight. Now this weight is distributed to the p positions in such a way that the weight becomes inversely proportional to the distance; i.e., (-1) position shares more weight and (-p) shares the minimum. Similar distribution is followed for the next positions also.

Let us explain the word unigram based confidence computation method with an example. Assume in the Hindi sentence, "*Aja pradhAnama.ntrI manamohana ne kahA ...*[1]" (Today prime-minister Manmohan has said), '*manamohana*' is identified as person name by a classifier. To get the confidence of '*manamohana*' as a person name, here we consider its previous two words and next two words. So the confidence of the word as person is, $\lambda_{-2} \times Wt_{\{per,-2\}}(Aja)$ + $\lambda_{-1} \times Wt_{\{per,-1\}}(pradhAnama.ntrI)$ + $\lambda_0 \times Wt_{\{per,0\}}(manamohana)$ + $\lambda_{+1} \times Wt_{\{per,+1\}}(ne)$ + $\lambda_{+2} \times Wt_{\{per,+2\}}(kahA)$. The λ_i denotes the weight factor of position i. Using unigram word weights computed using Eqn. 1 and

[1] All Hindi words are written in italics using the 'Itrans' transliteration.

corresponding λ_i values, the confidence weight of *manamohana* as person becomes 0.874.

During the confidence computation, few issues arise. To handle these issues the λ_i values are modified in such cases. Examples of such issues are, whether the current word is unknown (not present in the training corpus), whether any of the context words is unknown, whether any of the context words is postposition (these have very high occurrence in the corpus and as unigram these are not informative in identifying NEs) etc. Also in the NER task ambiguity between two classes is quite common. A NE belonging to a particular NE class, might occur as a NE or part of a NE of another class. Presence of some *clue words* often helps to resolve the ambiguous NEs. For example, presence of clue words like road, *palli* (locality), *setu* (bridge) etc. after a person name modifies it to a location NE. Such lists of clue words are collected for each pair of ambiguous NE classes. These words are given higher importance if they occur in the word window.

3 Confidence Prior Based MaxEnt

Maximum entropy (MaxEnt) principle is a commonly used learning technique which provides the probability of belongingness of a token to a class. MaxEnt computes the probability $p(o|h)$ for any o from the space of all possible outcomes O, and for every h from the space of all possible histories H. In NER task, history may be viewed as all information derivable from the training corpus relative to the current token. The computation of probability ($p(o|h)$) of an outcome for a token in MaxEnt depends on a set of features that are helpful in making the predictions about the outcome. Given a set of features and a training corpus, the MaxEnt estimation process produces a model in which every feature f_i (i - feature count index) has a weight α_i. We can compute the conditional probability as (Berger et al. 1996, Borthwick 1999):

$$p(o|h) = \frac{1}{Z(h)} \prod_i {\alpha_i}^{f_i(h,o)} \tag{2}$$

$$Z(h) = \sum_o \prod_i {\alpha_i}^{f_i(h,o)} \tag{3}$$

The conditional probability of the outcome is the product of the weights of all active features, normalized over the products of all the features. In the MaxEnt model the features are binary valued which are defined as,

$$f(h,o) = \begin{cases} 1, if\ (h,o) \in R; \\ 0, otherwise. \end{cases} \tag{4}$$

where R is an equivalent class over *(H,O)*. An example of MaxEnt feature is, if *previous word is professor* (h) = true and *NE_tag* (o) = person-begin then $f(h,o) = 1$. During the computation of $p(o|h)$ all such features which are active on the (h,o) pair, are considered.

Table 1. Features used in the MaxEnt based Hindi NER system

Type	**Features**
Word	current, previous two and next two words
NE tag	NE tag of previous two words
Digit information	Contains digit, digit & spl. char, numerical word
Affix information	Fixed length suffix, suffix list, fixed length prefix
POS information	POS of words, POS based binary features
Lists	Common locations, designations, person honorary terms, organization end terms

In simple supervised learning all the annotations in the training corpus are assumed to have same confidence values (say, 1 for each annotation). Now we introduce a confidence prior ($\rho_k \in [0,1]$) for each annotation (a_k) in the training corpus. These values are multiplied with $f_i(h,o)$ in Equation 2 i.e. $\acute{f}_i(h,o)$ becomes $\rho_k \times f_i(h,o)$.

In MaxEnt training algorithm (Borthwick 1999) for a feature $f(h,o)$ the entire training corpus is scanned to determine the number of times ($\#f$) the feature is active. As the feature is binary, simple occurrence count represents the corresponding feature value over the corpus. The prior based MaxEnt does not simply counts the number of times the $f(h,o)$ fires. Here for each fire, the corresponding confidence prior is multiplied with the feature value (which is 1 in case of binary features), so the $\#f$ is now replaced by $\sum_k \ \rho_k \times f_i(h,o)$.

4 Experimental Result and Discussion

The experimental results are discussed in this section for the Hindi NER task.

4.1 Training and Test Data

The training data for the Hindi NER task is composed of about $200K$ words which is collected from the popular daily Hindi newspaper "Dainik Jagaran". Here three types of NEs are considered; namely *Person*, *Location* and *Organization*. The corpus has been manually annotated and contains $5,429$ person, $4,390$ location and $2,678$ organization entities. The Hindi test corpus contains $25K$ words, which is distinct from the training corpus. The test corpus contains 678 person, 480 location and 322 organization entities.

4.2 Features

A list of candidate features are selected for the Hindi NER task. Several experiments are conducted with the features, individually and in combination to find the best feature set for the task. The features are mentioned in Table 1.

Table 2. Semi-supervised Hindi NER accuracies

Confidence Wt.	**Pre**	**Rec**	**F-val**
>1 (baseline)	85.92	68.37	76.15
0.9	85.07	71.86	77.91
0.8	84.82	72.64	77.8
0.7	84.93	70.47	77.03
0.5	84.3	69.6	76.25
0 (all data)	83.35	69.4	75.74

4.3 Baseline System

Here we present the baseline system accuracy. In NER task the accuracies are measured in terms of *f-measure* or *f-value*, which is the weighted harmonic mean of precision and recall. *Precision* is the percentage of correct annotations and *recall* is the percentage of the total NEs that are successfully annotated.

The baseline NE classifier is trained only using the available annotated data and the above mentioned features. The highest baseline accuracy, which is a f-value of 76.15, is obtained using words, NE tag of the previous word, suffix, prefix, digit information, POS information and list based features. The precision of the system is 85.92% and the recall is 68.37%. For this the class specific f-values are, person - 73.13, location - 81.05 and organization - 75.55. The accuracy is achieved when the word window of length three (previous one word to next one word) is used. Use of wider window reduces the accuracy. From the baseline accuracy we observe that the baseline system suffers from poor recall. Now we investigate whether we can improve the result by making use of additional raw corpus which is often easily available.

4.4 Semi-supervised NER: Selecting High-Confidence Samples

To perform semi-supervised experiments we have used a raw corpus containing $2000K$ words. This corpus is annotated by the baseline classifier. The confidence of the annotation is measured using the proposed approach (Section 2) and the the sentences with high confidence are selected and added to the training data. We have experimented with different confidence threshold values. The details of the experiments are given in Table 2.

It can be observed from Table 2 that the semi-supervised classifier performs better than the baseline if a suitable confidence threshold is chosen. The highest accuracy is achieved when the threshold is 0.9, which is a f-value of 77.91. Here amount of corpus selected and added with the training corpus is 179K words. Comparing with the baseline accuracy here we have achieved higher recall with a small decrease in precision. The recall is increased to 71.86 from 68.37. Use of lower threshold i.e. selecting more *additional samples* reduces the accuracy.

Table 3. Performance of semi-supervised approaches in Hindi NER

System	Pre	Rec	F-val
Baseline classifier	85.92	68.37	76.15
Semi-supervised high confident	85.07	71.86	77.91
Semi-supervised prior based	85.73	72.63	78.64

4.5 Semi-supervised NER: Confidence Prior in Classifier

Now we present the performance of the classifier which uses the annotation confidence value as a prior weight during training as discussed in Section 3. Here we have used all the baseline classifier annotated data along with their confidence weights. The MaxEnt classifier with prior based semi-supervised approach yields a f-value of 78.64. The prior based method performs better than baseline classifier as well as the high-confident portion selection based SSL approach (see Table 3.

5 Conclusion

We propose a semi-supervised learning framework for named entity recognition. The approach is based on a novel statistical confidence measure computed using NER specific cues. We use a sample selection approach as well as a prior modulation approach to enhance the performance of the NE classifier using the confidence measure. The approach is particularly useful for resource poor languages like Hindi where annotated data is scarce. Experimental results on a Hindi corpus consisting of a small annotated set and a much larger raw data set demonstrate that semi-supervised learning improves performance significantly over baseline classifier.

References

Berger, A., Pietra, S., Pietra, V.: A maximum entropy approach to natural language processing. Computational Linguistic 22(1), 39–71 (1996)

Borthwick, A.: A maximum entropy approach to named entity recognition. Ph.D. thesis, Computer Science Department, New York University (1999)

Collins, M., Singer, Y.: Unsupervised models for named entity classification. In: Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (1999)

Li, W., McCallum, A.: Rapid development of Hindi named entity recognition using conditional random fields and feature induction. ACM Transactions on Asian Language Information Processing (TALIP) 2(3), 290–294 (2004)

Mohit, B., Hwa, R.: Syntax-based semi-supervised named entity tagging. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 57–60. Association for Computational Linguistics, Ann Arbor (2005)

Saha, S., Sarkar, S., Mitra, P.: A hybrid feature set based maximum entropy Hindi named entity recognition. In: Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP), pp. 343–349 (2008)

A News Analysis and Tracking System

Sk. Mirajul Haque, Lipika Dey, and Anuj Mahajan

TCS Innovation Labs, Delhi
{skm.haque,lipika.dey,anuj.mahajan}@tcs.com

Abstract. Continuous monitoring of web-based news sources has emerged as a key intelligence task particularly for Homeland Security. We propose a system for web-based news tracking and alerting. Unlike subscription-based alerts, alerting is implemented as a personalized service where the system is trained to recognize potentially important news based on user preferences. Preferences are expressed as combinations of topics and can change dynamically. The system employs Latent Dirichlet Allocation (LDA) for topic discovery and Latent Semantic Indexing (LSI) for alerting.

1 Introduction

The amount of news content available online is increasing at a steady rate. News belongs to a very specific genre of textual data. News contents from multiple sources are usually similar in that they largely look at the same set of events, yet they are at the same time highly unstructured and open ended. Timely assimilation and interpretation of information available from these sources is extremely important for political and economic analysts. It is imperative to analyze the whole collection to eliminate the possibility of missing anything.

Automated acquisition, aggregation and analysis of news content is a challenging and exciting line of research, lying at the crossroads of information extraction, machine learning, machine translation, pattern discovery, etc. The complexity of the problem increases due to the unpredictability of incoming content that needs to be analyzed. User preferences cannot be expressed as pre-defined categories. A news item of any category - be it political, sports, or entertainment can become interesting to an analyst, depending on its content. An analyst may wish to track it for some time. Further the interest changes dynamically. The key challenge in recognizing relevant news lies in recognizing the concepts of interest to a user.

In this paper, we have presented a system that is distinct in functionality from other news analysis systems. It is designed as a watch-guard which continuously scans online sources for new stories and alerts a registered user whenever there is any new development that could be of potential interest to the user. The news repository is organized topically. Users can register their interests with the system in terms of topics. The system generates an alert if a new story is judged to be conceptually relevant to a user. Alerts are also sent to the user's mobile. The novelty of the proposed system lies in its capability to track conceptually relevant news items as opposed to news in pre-specified categories.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 231–236, 2009.

2 News Analysis Systems – A Review

News categorization and topic tracking through news collections has been an active area of research for a few years now. In [1] five categorization techniques were compared for news categorization. It was found that Support Vector Machines (SVM), k-Nearest Neighbor (kNN) and neural networks significantly outperform Linear Least squares Fit (LLSF) and Naive Bayes (NB) when the number of training instances per category is small. [4] presents a system for automatic categorization of news stories written in Croatian language into a pre-defined set of categories using kNN based classifier. [10] described a news analysis system to classify and index news stories in real time. The system extracted stories from a newswire, parsed the sentences of the story, and then mapped the syntactic structures into a concept base.

Since news contents change in unpredictable ways supervised techniques that require large training sets are not ideally suited for news repository organization. Topic tracking and detection [2] is another area of research where the emphasis is to monitor online feed of news articles, and determine the underlying topics and their distribution in the news items in an unsupervised way. In [7] a topic tracking system based on unigram models was presented. Given a sparse unigram model built from topic training data and a large number of background models built from background material, the system finds the best mixture of background models that matches the unigram model. [8] proposed a clustering based approach to summarize and track events from a collection of news Web Pages. In [9] the problem of *skewed data* in topic tracking was tackled using semi-supervised techniques over a bi-lingual corpus. The Lydia project [3] describes a method for constructing a relational model of people, places and other entities through a combination of natural language processing and statistical analysis techniques.

3 Overview of the Proposed News Analysis System

Fig.1 presents the software architecture of the proposed news analysis system. The key components of the system are:-

- *News acquisition module* - The News acquisition module collects news from a host of pre-defined websites from their RSS[1] feeds. The RSS feeds contain news headlines, date and time of publishing, a brief description, and the URL of the full news story. These are periodically updates. The URLs extracted from the RSS feeds are fed to customized web crawlers, to extract the news stories.
- *Topic Discovery* – This module builds a topic map for the entire news collection using Latent Dirichlet Allocation (LDA). It also indexes the collection topically. Each time a news repository is updated, the existing topic maps are backed up and a new topic map is built.
- *Topic Tracker and Trend Analyzer* – Topic maps are time-stamped. These can be compared to analyze how the strength of a topic has changed over time. Correlation analysis of topic maps identifies entry and exit of new topics into the system. Topic tracking and trend analysis can provide early warnings and actionable intelligence to avoid disasters.

[1] http://en.wikipedia.org/wiki/RSS_(file_format).

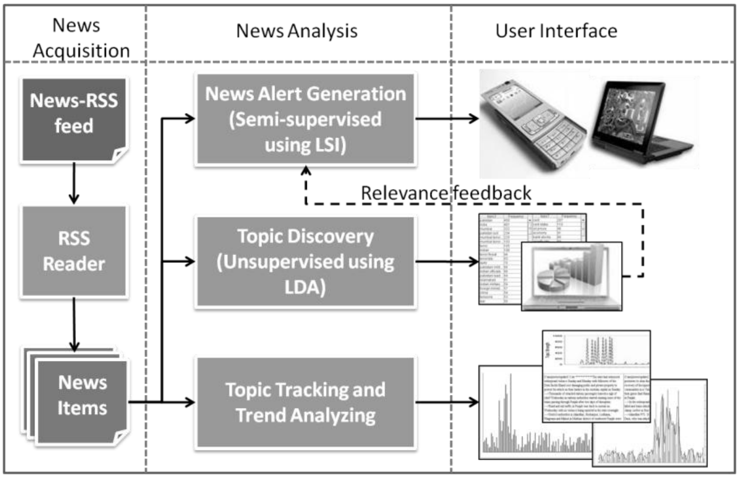

News Acquisition
News Analysis
User Interface
News-RSS feed
RSS Reader
News Items
News Alert Generation (Semi-supervised using LSI)
Relevance feedback
Topic Discovery (Unsupervised using LDA)
Topic Tracking and Trend Analyzing

Fig. 1. Architecture of proposed News Analysis System

- *News Alert Generator* – This is a trainable module that implements Latent Semantic Indexing (LSI) to categorize an incoming news story as relevant or irrelevant to any of the multiple interest areas expressed by the user. The alerting mechanism is designed as a trainable multi-classifier system. User interests can be further classified as *long-term* and *short-term.* Concepts in long term interest category change at a slower rate than those in the short-term interest category. While long-term interests encode generic needs of a news monitoring agency, short-term interests help in tracking evolving stories in any category.
- *Topic Explorer* – This is the interface through which the user interacts with the system. A topic snap-shot shows all topics and concepts present in a collection. Users can drill-down to detailed reports on any event along any topic. This allows a single story to be read from multiple perspectives based on its contents. Topic-based aggregation allows the user to get a bird's eye view of the key events. Users can also view topic strengths and localizations.

3.1 Topic Extraction Using LDA

An important step to comprehend the content of any text repository is to identify the key topics that exist in the collection, around which the text can be organized. Latent Dirichlet Allocation (LDA) [5] is a statistical model, specifically a topic model, in which a document is considered as a mixture of a limited number of topics and each meaningful word in the document can be associated with one or more of these topics. A single word may be associated to multiple topics with differing weights depending on the context in which it occurs. Given a collection of documents containing a set of words, a good estimation of unique topics present in the collection is obtained with the assumption that each document d can be viewed as a multinomial distribution over k topics. Each topic z_j , j = 1 · · · k, in turn, is assumed to be a multinomial distribution $\Phi(j)$ over the set of words W. The problem of estimating the various distributions is in general intractable. Hence a wide variety of approximate algorithms, that attempt to maximize likelihood of a corpus given the model have been proposed for LDA. We have used the Gibb's sampling based technique proposed in [5] which is computationally efficient. The topic extractor creates a *topic dictionary* where each topic is represented as a collection of words along with their probabilities of belonging to the topic. Each document *d* is represented as a *topic distribution vector* < ($p(t_1,d),\ldots,p(t_k, d)$>, where $p(t_j,d)$ denotes the strength of topic t_j in document *d* is computed as follows:

$$\mathrm{p}(\mathrm{t_j}, \mathrm{d}) = \sum_{w=1}^{w=n} p(w, t_j)\ . \tag{1}$$

where *n* is the total number of unique words in the collection.

3.2 News Alert Generation Using Latent Semantic Indexing

All incoming news stories are scored for relevance based on its topic distribution. The system employs Latent Semantic Indexing (LSI) to assign relevance scores. LSI uses Singular Value Decomposition (SVD) to identify semantic patterns between words and concepts present in a text repository. It can correlate semantically related terms within a collection by establishing associations between terms that occur in similar context.

User interests are conveyed in terms of topics. Given that the underlying repository has stories containing different topics, of which some are relevant and some not, the aim is to find new stories which are conceptually relevant to the user, even though they may not be content-wise similar to any old story as such. The training documents for the system are created as follows. For each relevant topic, all stories containing the topic with strength greater than a threshold are merged to create a single training document d_R. All remaining stories are merged to create another training document d_{IR}. Let T_{m2} denote the term-document matrix for this collection, where m represents the number of unique words. The weight of a word in T is a function of its probability of belonging to the topics in each category. Applying SVD to T yields,

$$T = UWV^T\ . \tag{2}$$

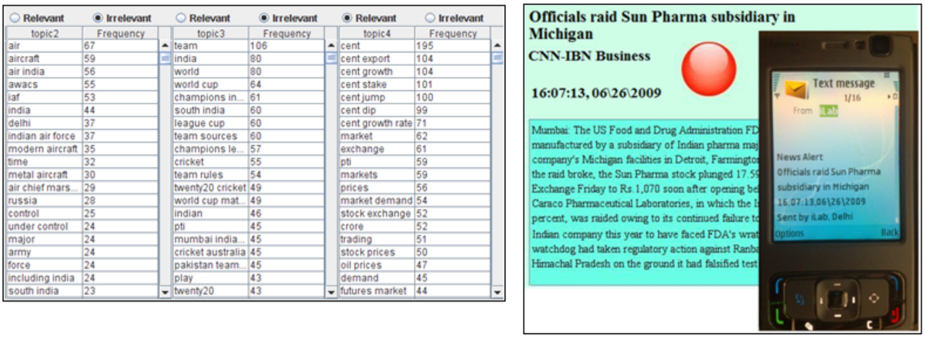
where U and V are orthonormal matrices and W is a diagonal matrix. A modified term-document matrix $\hat{T}$ is now constructed using $\hat{U}$ which consists of the first two columns of U such that $\hat{T}$ represents a transformed 2 dimensional concept-space, where each dimension is expressed as a linear combination of the original terms and the relevant and irrelevant documents are well-separated. For each new news story, its cosine similarity with the relevant and irrelevant documents is computed in the new concept space. Let Q represent the term vector for the new story. Q is then projected into the transformed concept space to get a new vector $\hat{Q}$. Let S_R and S_{IR} denote the cosine similarity of $\hat{Q}$ with the relevant and irrelevant documents respectively. The new story is judged as relevant if

$$S_R > \alpha * S_{IR}\ . \tag{3}$$

where α is a scaling parameter in the range (0, 1). Since number of news stories in irrelevant category is usually much higher than number of relevant news, there is an automatic bias towards the irrelevant category. The *scaling* parameter α adjusts this bias by reducing the similarity to irrelevant category.

4 Experimental Results

Fig. 2 presents some snapshots. On the left is a portion of the topic dictionary created for a 2 week time-period ending on 9th June, 2009. Users select topics to indicate preferences for receiving alert. On the right an alert generated for the user is shown.



○ Relevant ● Irrelevant ○ Relevant ● Irrelevant ● Relevant ○ Irrelevant

topic2	Frequency	topic3	Frequency	topic4	Frequency
air	67	team	106	cent	195
aircraft	59	india	80	cent export	104
air india	56	world	80	cent growth	104
awacs	55	world cup	64	cent stake	101
iaf	53	champions in...	61	cent jump	100
india	44	south india	60	cent dip	99
delhi	37	league cup	60	cent growth rate	71
indian air force	37	team sources	60	market	62
modern aircraft	35	champions le...	57	exchange	61
time	32	cricket	55	pti	59
metal aircraft	30	team rules	54	markets	59
air chief mars...	29	twenty20 cricket	49	prices	56
russia	28	world cup mat...	49	market demand	54
control	25	indian	46	stock exchange	52
under control	24	pti	45	crore	52
major	24	mumbai india...	45	trading	51
army	24	cricket australia	45	stock prices	50
force	24	pakistan team...	45	oil prices	47
including india	24	play	43	demand	45
south india	23	twenty20	43	futures market	44

Officials raid Sun Pharma subsidiary in Michigan

CNN-IBN Business

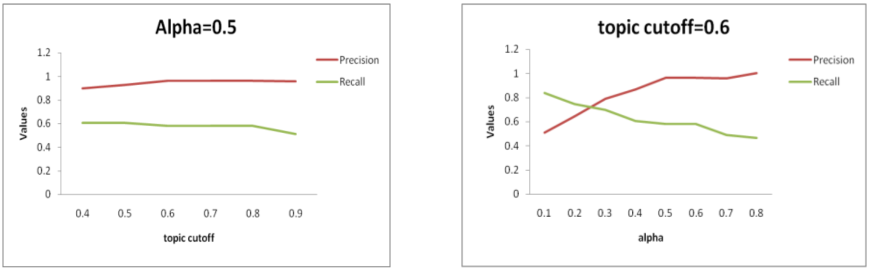
16:07:13, 06\26\2009

Mumbai: The US Food and Drug Administration FD
manufactured by a subsidiary of Indian pharma maj
company's Michigan facilities in Detroit, Farmingto
the raid broke, the Sun Pharma stock plunged 17.59
Exchange Friday to Rs.1,070 soon after opening be
Caraco Pharmaceutical Laboratories, in which the I
percent, was raided owing to its continued failure to
Indian company this year to have faced FDA's wrat
watchdog had taken regulatory action against Ranba
Himachal Pradesh on the ground it had falsified test

Text message
1/16
From iLab

News Alert
Officials raid Sun Pharma
subsidiary in Michigan
16:07:13,06\26\2009
Sent by iLab, Delhi

Fig. 2. Snapshots from the system



(a) (b)

Fig. 3. (a) Precision and Recall with varying topic cutoff and fixed alpha=0.5 (b) Precision and Recall with varying alpha and topic cutoff = 0.6

Since user preferences are given as topics, it is observed that the accuracy of recognizing new stories as relevant depends on (a) the size of the training set (b) the scaling factor α. Size of the training set is controlled using the topic cut off threshold. Fig. 3 shows system performance varies with these two parameters. With a fixed alpha, it is observed that recall falls slightly with higher values of topic cut-off, though precision does not suffer. Higher topic cut-off reduces the training set size causing the recall to fall. Sharper trends are observed with a fixed topic cut-off, and changing values of alpha. Recall falls drastically with higher values of alpha, since very few stories are now picked as relevant, though the precision is very high. Best results are observed for alpha and topic cut-off set to 0.5 and 0.6 respectively.

5 Conclusion

In this work, we have presented a news analysis system which generates user-preference based alerts. Preferences are for topics and concepts rather than predefined categories. The news repository is analyzed and indexed topically. We are currently working on integrating a mining and analytical platform to this system to enhance its prediction and tracking capabilities.

References

1. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: Proceedings of 22nd ACM SIGIR, California (1999)
2. Allan, J., Papka, R., Lavrenko, V.: On-line New Event Detection and Tracking. In: Proceedings of 21st ACM SIGIR, Melbourne (1998)
3. Lloyd, L., Kechagias, D., Skiena, S.: Lydia: A System for Large-Scale News Analysis. In: Consens, M.P., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 161–166. Springer, Heidelberg (2005)
4. Bacan, H., Pandzic, I.S., Gulija, D.: Automated News Item Categorization. In: JSAI (2005)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
6. Landauer, T.K., Foltz, P.W., Laham, D.: An Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)
7. Yamron, J.P., Carp, I., Gillick, L., Lowe, S., Van Mulbregt, P.: Topic Tracking in a News Stream. In: Proceedings of DARPA Broadcast News Workshop (1999)
8. Mori, M., Miura, T., Shioya, I.: Topic Detection and Tracking for News Web Pages. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 338–342 (2006)
9. Fukumoto, F., Suzuki, Y.: Topic tracking based on bilingual comparable corpora and semi-supervised clustering. ACM Transactions on Asian Language Information Processing 6(3) (2007)
10. Kuhns, R.J.: A News Analysis System. In: Proc. of 12th International Conference on Computational Linguistics, COLING 1988, vol. 1, pp. 351–355 (1988)

Anomaly Detection from Call Data Records

Nithi and Lipika Dey

TCS Innnovation Labs, Delhi
{nithi.1,lipika.dey}@tcs.com

Abstract. In this paper, we propose an efficient algorithm for anomaly detection from call data records. Anomalous users are detected based on fuzzy attribute values derived from their communication patterns. A clustering based algorithm is proposed to generate explanations to assist human analysts in validating the results.

Keywords: Anomaly detection, Fuzzy logic, DBSCAN Algorithm, Principal Component Analysis.

1 Introduction

Call Data Records (CDR) define a special kind of social network where nodes represent phone numbers and an edge represents a call between two numbers. Social networks derived from call data records model relationships among a group of mobile users. Investigative analysis has come to heavily rely on Call Data Record (CDR) analysis since these can provide major cues about temporally and spatially allied events as well as people involved.

Anomaly detection from call data records employs data mining techniques to identify abnormal behavioral patterns. As of today, major emphasis of CDR analysis has been towards visual analytics [1, 2]. The major drawback of such systems is the overemphasis on the analyst's capability to identify the regions of interest, particularly if the volume of data to be considered is prohibitively large.

In this paper we propose an automated anomaly detection mechanism that assigns anomaly score to nodes in a CDR network. It has been shown that rather than using the features of a single subscriber and his contacts, better results can be obtained by using attributes of links at greater depths. Efficiency is ensured through suitable feature transformation techniques.

2 Anomaly Detection in Social Networks – A Review

Anomaly detection is a mature area of research [3] and has been successfully applied to various areas like network intrusion detection, credit-card fraud detection, email based network analysis etc. They work on the premise that normal behavior is more pre-dominant than abnormal behavior and will be exhibited by majority of the network entities. Commonly used techniques for network analysis involves computation of Centrality measures [4]. [5] had proposed the use of indirect connections to detect

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 237–242, 2009.

anomalous subscribers from call data records. [6] had applied the algorithm to VAST dataset [9] to discover suspicious elements.

3 Feature Identification for Anomaly Detection

The properties typically used for Call Data Record analysis are those that capture individual user behavioral properties like number of calls made or received, number of contacts, towers used for making calls etc. Users can also be associated with statistical properties like average number of calls made per day, average number of contacts, etc. Interaction between any two users is described by the nature of calls between them. In our approach, behavior of an ID is derived from his direct and indirect interactions.

Call details from a service provider contain information about calls made or received by its subscribers. Each call is represented by a caller ID, a receiver ID, call duration, time of call initiation, tower used, call type (VOICE or SMS) and some details about the handset used. Each interaction is either of type OUTGOING or INCOMING for a given user. We represent the behavior of each user by his calling pattern and also his interaction patterns with other users. Interaction pattern takes into account the frequencies of different types of calls associated to the user.

The most important attribute of a call is its duration. Rather than using exact call duration times we characterize a call as “long” or “short” with a certain degree of fuzziness. They provide a robust characterization which is more tolerant to noise.

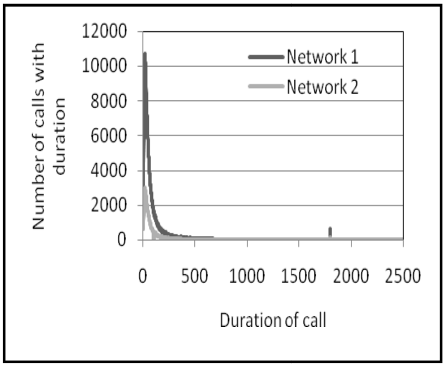


Fig. 1. Distribution of the duration of calls for two networks

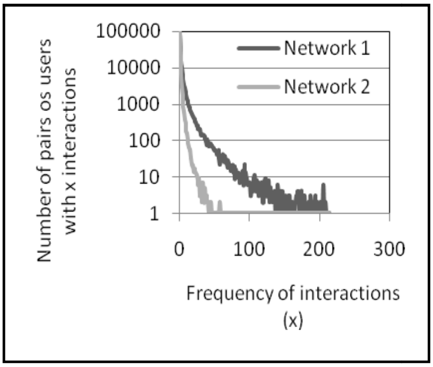


Fig. 2. Distribution of the number of pair-wise interactions

The relevant fuzzy membership functions were derived after analyzing various sets of call data records for different mobile networks. Figure 1 shows graphs for two networks - network 1 with 30866 users, 77691 calls over 12 days and network 2 with 31494 users, 207704 calls over 2 days. It shows that short duration calls are much more frequent than long duration calls. The exact cut-off value for identifying a call as long or short however can vary for different networks since they operate with different pulse rates. The proposed fuzzy membership function is designed to take this variation into account. It is as follows.

A long duration call is one which is longer than 95% or more of the other calls and a short duration call is one whose duration is less than the duration of 75% of the calls. A call of duration d is characterized as follows:

$\mu_S(d) = 1, \mu_L(d) = 0$ if the d < 75 percentile of duration of calls

$\mu_S(d) = 0, \mu_L(d) = 1$ if the d > 95 percentile of duration of calls

$$\mu_S(d) = \frac{0.95 - x}{0.95 - 0.75}, \mu_L(d) = \frac{x - 0.75}{0.95 - 0.75} \text{ where x denotes the percentile of d} \tag{1}$$

otherwise

where S denotes short, L denotes long.

While a call can be characterized by its duration, interactions between two users can be defined by the frequency of different types of calls between them. Frequency can be characterized as "high" or "low". Figure 2 shows that pair of users that interact infrequently is exceptionally more than pairs with frequent interactions.

Depending on the frequency f of interaction between two users, the interaction type is characterized as "high" or "low" using the following fuzzy functions:

$\mu_{low}(f) = 1, \mu_{high}(f) = 0$ if the f < 95 percentile of frequency of calls

$\mu_{low}(f) = 0, \mu_{high}(f) = 1$ if the f > 99 percentile of frequency of calls

$$\mu_{low}(f) = \frac{0.99 - x}{0.99 - 0.95}, \mu_{high}(f) = \frac{x - 0.95}{0.99 - 0.95} \tag{2}$$

where x denotes the percentile of f, otherwise

We can now combine call characteristics and interaction characteristics to have a more complete characterization of users. Two users can be connected through different types of calls with different frequencies. For example, two users A and B may frequently communicate through "short" calls, but rarely through "long" calls. It may be further observed that most of the "short" calls are initiated by B while the "long" calls are initiated equally by both. Consequently, the interaction pattern of A with B can be characterized as the following vector of fuzzy-membership values:

$<\mu_{LOW}(OS), \mu_{HIGH}(OS), \mu_{LOW}(OL), \mu_{HIGH}(OL), \mu_{LOW}(IS), \mu_{HIGH}(IS), \mu_{LOW}(IL), \mu_{HIGH}(IL)>$, where the symbol O is used for an OUTGOING call and I for an INCOMING call, S stands for short-duration call, L for long-duration call. $\mu_{LOW}(OS)$ is computed using the weighted average of fuzzy memberships of all OUTGOING calls to the class "short" or S. All other memberships are computed in a similar fashion.

It may be essential to consider a user's indirect connections also. In the current framework, if a user is connected to another through a chain of contacts, then the indirect interaction pattern between the two users is defined by the intermediate chain of direct interaction patterns. Let $LINK_1$ represent the direct interaction vector of A with B, and $LINK_2$ represent the direct interaction vector of B with C, then a vector $LINK_{12}$ that defines the indirect interaction pattern of A with C is computed as a 64 component vector. The $(ij)^{th}$ component of $LINK_{12}$ is computed as follows:

$LINK_{12}(ij) = \min(LINK_1(i), LINK_2(j))$, where i and j vary from 1 to 8.

Paths up to any predefined depth can be considered, though the impact of one node on another decreases with increase in depth. The dimensionality of the feature space also goes up. Weight for a pattern for an ID is the sum of weights of all interactions of that pattern for the ID. The behavior space is then normalized with respect to patterns to equalize the importance of all patterns.

For example, consider the call graph described in Figure 3. It is observed that A interacts with B through short duration calls only. Moreover, the frequency of short calls initiated by A is high but terminated at A is low. Consequently, the interaction pattern of A with B is described by the vector stated above with value 1 for μ_{HIGH}(OS) and μ_{LOW} (IS). Further, B interacts with D by initiating long duration calls with medium frequency. Value for μ_{LOW} (OL) is found to be 0.4 and for μ_{HIGH} (OL) is 0.6 in the vector describing interaction of B with D. The interactions of all other users are described in a similar way. Though A is not directly connected to D, but A is connected to D through B and C. Since A calls B by frequent short duration calls and B initiates medium long duration calls to D, therefore the value for the feature μ_{LOW}(OS) μ_{HIGH} (OL) in the interaction pattern of A and D via B is 0.6. Feature values for A's interaction with D through C can be similarly computed. The final values for such an interaction vector are computed using component-wise summation.

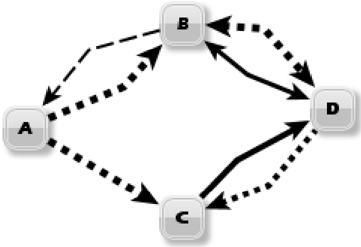


Fig. 3. Call graph. Dotted (solid) lines indicate low (high) duration calls. Frequency of calls is mapped to thickness of the edge.

4 Anomaly Detection

Due to the high dimension of the original feature space derived in the earlier section, we have applied Principal Components Generator to identify the dimensions along which the data set shows maximum variation. We then chose the three principal components only to represent the transformed dataset. An algorithm based on Ramaswamy's distance is designed to detect outliers [8]. This algorithm finds top *n* outliers based on their distance from their kth closest neighbor. This algorithm works as follows:

Step 1: Choose values of n and k.
Step 2: For each data point find its distance from its kth closest neighbor.
Step 3: Arrange the data points in descending order of the distance.
Step 4: Choose top n points as outliers.

5 Explanation of Interaction Patterns of Anomalous Users

Principal components are linear combinations of attributes. They do not retain original feature values. However, for intelligence analysis it is essential that explanation of anomalous behavior be generated in terms of original features. For small data sets one can generate the rules using decision trees, where top *n* anomalies are treated as elements of one class and the rest of the elements of another.

The decision tree is not suitable for generating rules from large datasets. In this case, for each of the top *n* anomalies detected, we find its p closest neighbors in the original feature space. We have set p to 1000. These neighbors are now clustered using the DBSCAN algorithm. Non-anomalous entities are much more in number and they form dense clusters. The features that distinguish the seed anomalous entity from all its neighboring clusters with more than 10 elements are extracted. For each cluster 5th and 95th percentile of the feature values are computed for all features. Each feature value of the seed is then compared with computed percentile of the corresponding features. The features with high variation are then extracted and reported.

6 Experimental Results

Table 1 shows that our system is able to better the results of [6] on VAST 2008 dataset [9]. This set has 400 entities with 10000 interactions. The challenge was to identify 12 members of a gang who conducted suspicious activities. All other results reported in VAST were for human-driven visual analysis [1]. Figure 4 shows that the anomalous entities become easily separable on the transformed space. Figure 5 shows a sample rule characterizing anomalous and non-anomalous entities. A feature characterizing interaction at depth 3 turns out to be most significant. It typically characterizes interaction patterns observed between key gang members and ground level workers.

Table 1. Performance results for various algorithms on VAST2008 dataset

Algorithm	Proposed Algorithm	Kojak[6]
Precision	80%	60%
Recall	80%	60%

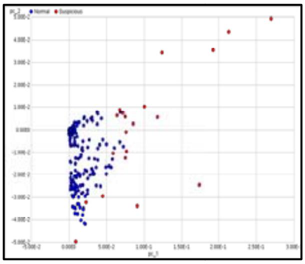


Fig. 4. Distribution of anomalous (RED) and non-anomalous (BLUE) entities (VAST 08

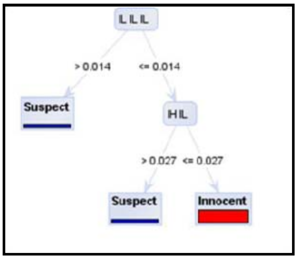


Fig. 5. Rule characterizing anomalous IDs

7 Conclusion

Call data record analysis have gathered momentum due to the role they play in investigative analysis. Most of the research till date has been directed towards visual analytics. We have presented here methods for finding anomalous behavior from large datasets, considering interaction up to any depths. We have presented the viability of using very complex feature spaces in conjunction with PCA to capture complex notions of behavior. The algorithms proposed here have been applied to large real-life data sets and found to scale extremely well.

References

1. Swing, E.: Solving the Cell Phone Calls Challenge with the Prajna Project. In: IEEE Symposium on Visual Analytics Science and Technology, Columbus, Ohio, USA (2008)
2. Payne, J., Solomon, J., Sankar, R., McGrew, B.: Grand Challenge Award: Interactive Visual Analytics Palantir: The Future of Analysis. In: IEEE Symposium on Visual Analytics Science and Technology, Columbus, Ohio, USA (2008)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. To Appear in ACM Computing Surveys (2009)
4. Wasserman, S., Faust, K.: Social Network Analysis: Methods & Applications. Cambridge University Press, Cambridge (1994)
5. Lin, S., Chalupsky, H.: Discovering and explaining abnormal nodes in semantic graphs. IEEE Transactions on Knowledge and Data Engineering 20 (2008)
6. Chalupsky, H.: Using KOJAK Link Discovery Tools to Solve the Cell Phone Calls Mini Challenge. In: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, Portugal (2008)
7. Ross, T.J.: Fuzzy Logic with Engineering Applications. Wiley, Chichester (2004)
8. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient Algorithms for Mining Outliers from Large Data Sets. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Texas, United States (2000)
9. IEEE VAST Challenge (2008), http://www.cs.umd.edu/hcil/VASTchallenge08

Mining Calendar-Based Periodicities of Patterns in Temporal Data

Mala Dutta and Anjana Kakoti Mahanta

Gauhati University, India
maladuttasid@gmail.com, anjanagu@yahoo.co.in

Abstract. An efficient algorithm with a worst-case time complexity of $\mathrm{O}(\mathrm{n} \log n)$ is proposed for detecting seasonal (calendar-based) periodicities of patterns in temporal datasets. Hierarchical data structures are used for representing the timestamps associated with the data. This representation facilitates the detection of different types of seasonal periodicities viz. yearly periodicities, monthly periodicities, daily periodicities etc. of patterns in the temporal dataset. The algorithm is tested with real-life data and the results are given.

Keywords: temporal datasets, pattern detection, periodicity mining.

1 Introduction

Electronic data repositories maintained across various domains usually contain data that are inherently temporal in nature. Some examples of these are credit-card transactions, market-basket datasets, stock prices, click-stream data etc. In large volumes of such data, patterns may appear periodically within the lifetime of the data. In [7], an algorithm with a worst-case time complexity of $\mathrm{O}(\mathrm{n}^3)$ is proposed for extracting calendar-based periodic patterns in a market-basket database. Here an algorithm is proposed that runs in $\mathrm{O}(\mathrm{n} \log n)$ time in the worst-case to detect periodicities of patterns in different kinds of temporal data. The algorithm works not only on a typical market-basket database, but also applies to any event-related dataset where events are associated with time-intervals marking their occurrence. The algorithm also successfully extends to any time-series database in which case at first, an attempt is made (usually by using the dynamic time-warping technique), to look for time-periods that may be significant in the context of the domain of the data. For example, in data-streams of stock prices, time-periods in which rising interest rates (peaks) are noticed might be of interest. In sales data, time-periods in which panic reversal of sales occur are usually significant. The proposed algorithm finds if there is any periodicity among the time-segments in which such domain-specific patterns are observed across the time-series.

In Sect. 2, recent works done in the field of pattern mining in temporal data are discussed. The work done in this paper and the algorithm proposed is discussed in Sect. 3. Experimental results are given in Sect. 4 followed by the conclusion and lines for future research in Sect. 5.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 243–248, 2009.

2 Related Works

Pattern mining in temporal data is an active research area. In [4], user-given calendar schemas are used to find a set of calendar-based patterns. In [2], an algorithm to find periodic patterns of unknown periods in a time-series database is proposed. [5] gives a linear distance based algorithm for discovering the potential periods regarding the symbols of a time-series. A similar algorithm has also been proposed in [9] where some pruning techniques have been used. However since both the algorithms consider the distance between adjacent intervals, they miss some valid periods. In [8], an algorithm is proposed to extract synchronous patterns having non-stationary offsets within each period. In [3], the problem of mining multiple partial periodic patterns in a parallel computing environment is addressed. The problem of mining frequently occurring periodic patterns with a gap requirement from sequences has been studied in [10]. In [7], the works in [1] and [6] are extended and a method for extracting calendar-based periodic patterns in a market-basket dataset is proposed. The algorithm proposed in this paper is more time-efficient than the one given in [7]. It easily extends to different types of temporal datasets and to other time hierarchies also.

3 Work Done

The problem that has been taken up in this paper is to find periodicities of patterns in different kinds of temporal data. The input consists of events (which are referred to as patterns) and with each event, there is associated a list of disjoint time intervals in which that event has occurred. The events could either be natural events such as earthquakes, cyclones etc. or could be events that are implicit in the dataset under consideration such as rise or downfall in interest rate in stock market data, steep rise or dip in temperature in weather-related time-stamped data etc. In the latter case, a preprocessing step is needed for identifying and subsequently extracting the relevant time intervals (again assumed to be disjoint) from the dataset under consideration.

If the timestamps associated with the temporal data are given as calendar dates, for example in the format hour-day-month-year, yearly or monthly or daily periodicities may be detected. Periodic patterns are detected starting from the highest level of the time-hierarchy. While finding yearly patterns, the year component of the time-hierarchy is removed from the dates. Similarly, appropriate components are stripped when monthly or daily periodicities are being searched.

For a pattern under study, a certainty function is defined on the timestamps stripped of the appropriate components. This is given by $c(t) = nintv/np$ where $nintv$ is the number of intervals containing the stripped timestamp t and np is the total number of years or months or days in the lifespan of the pattern (depending on whether yearly or monthly or daily periodicities are being detected). If the value of the certainty function becomes equal to 1 at any timestamp, it will imply that the pattern appears at that timestamp in each of the periods and hence the pattern is fully periodic at that timestamp. On the other hand, if the certainty value is positive and less than 1 at any timestamp, then this will imply that the pattern appears at that timestamp in some of the periods only and hence the pattern will be partially periodic at that timestamp.

3.1 Data Structure and Algorithm Proposed

A global variable *np* is maintained for the total number of periods in the lifespan of the pattern. The information of each discontinuity of the certainty function is captured in a structure *discontinuity* containing a timestamp *tmp* at which the discontinuity of the certainty function is detected, an integer *iv* which is the number of intervals to which timestamp *tmp* belongs and an integer *div* which gives the change that occured in the value of *iv* at timestamp *tmp*. Additionally *v*, the value of the certainty function at the timestamp *tmp*, which is actually nothing but iv/np can be retained. Information of each local maximum of the certainty function is kept in another structure *lmaxnode* which contains the timestamps *start*, *end*, *peakstart*, *peakend* and also the values *startvalue*, *peakvalue* and *endvalue* of the local maximum.

The processed information regarding the certainty function is kept in a structure containing an array *d* of *discontinuity*, an array *lmax* of *lmaxnode* and integers *m* giving the number of input intervals, *nd* giving the number of discontinuities and *nlmax* giving the number of local maxima found. Since each endpoint of the given intervals associated with the pattern potentially represents a discontinuity of the certainty function, the algorithm starts by populating the array of discontinuities *d* by the endpoints of the intervals given in the input. For a left endpoint, *tmp* is set to the timestamp of the endpoint and *div* is set to 1 (since a new interval starts at this timestamp). For a right endpoint, *tmp* is set to one more than the timestamp of the endpoint and *div* is set to -1 (since an interval ends just before the timestamp *tmp*). After this, the procedure *reorganize* now processes the information in the array of discontinuities *d*. The processing is based on the observation that the value of the certainty function at any timestamp is given by $v = (nl - nr)/np$ where *nl*, *nr* are the number of left and right endpoints respectively coming before this timestamp. It is easy to see that *v* is always non-negative. The *reorganize* procedure first sorts *d* in ascending order of timestamps. It then scans the array *d* in this order, collapsing all the potential discontinuities occurring at the same timestamp and then computes the *div*, *iv* and *v* values for timestamp *tmp* at each discontinuity.

Next, a procedure *locatemax* gathers information of each local maximum of the certainty function. If a single local maximum is found, it will imply piling up of overlapping subintervals at a single place. This indicates that the pattern appears once a year or month or day in that time-period, depending on whether yearly or monthly or daily patterns are being searched. Similarly, when two local maxima are detected, the time-intervals obviously overlap at two places indicating that the pattern holds twice a year/month/day. For detecting the local maxima, the procedure *locatemax* goes through the array *d* of discontinuities. A boolean variable *nondecreasing* is maintained which is initially set to *false*. For each discontinuity, if *nondecreasing* is *false* and the value of *iv* is found to be increasing at that discontinuity i.e. if $d[i].div > 0$ then the *end*, *endvalue* of the last local maximum, the *start*, *startvalue* of a new local maximum are set appropriately and *nondecreasing* is changed to *true*. If *nondecreasing* is *true* and the value of *iv* is decreasing at that discontinuity i.e. if $d[i].div < 0$ then the *peakstart*, *peakend* and the *peakvalue* of the peak of the current local maximum are set appropriately and *nondecreasing* is changed to false. Finally, if this is the last discontinuity then the *end* and the *endvalue* of the last local maximum are set appropriately. Once the whole structure is in place, the certainty of the pattern at any arbitrary point (i.e. date) can also be queried easily by doing a binary search on the array *d*.

3.2 Algorithm Complexity

The operation of populating the discontinuity array *d* will take O(n) time in the worst-case where *n* is the length of the input list of time-intervals. The reorganization procedure involves sorting the array *d* (this takes $\mathrm{O}(\mathrm{n}\log n)$ time in the worst-case) and then setting the values of certain fields/variables (this takes O(n) time in the worst-case). Scanning for local maxima takes O(n) time in the worst-case. Thus overall complexity in the worst case works out to be $\mathrm{O}(\ \mathrm{n} + \mathrm{n}\log n + \mathrm{n} + \mathrm{n}\ ) = \mathrm{O}(\mathrm{n}\log n)$. Once all this is done, to compute the certainty of the pattern at any arbitrary point (i.e. date) takes only $\mathrm{O}(\log n)$ time in the worst case. Hence there is a definite improvement in running time as compared to the algorithm proposed in [7] which takes $\mathrm{O}(\mathrm{n}^3)$ in the worst-case to detect seasonal periodicities of patterns in a market-basket database.

4 Experimental Results

The proposed algorithm is applied here to two temporal datasets, both containing real data. The first dataset is an event-related dataset. The second dataset is a time-series.

Working with an event-related dataset : This dataset contains time-periods of occurrence of tropical storms in the eastern-pacific region from 1949 to 2008. The source of this data is http://weather.unisys.com/hurricane/index.html. While looking for yearly periodicities, several time-periods are found during which tropical storms across the eastern-pacific region are likely to occur with a certainty of at least 20%. Some of these time-periods are shown in Table 1 along with the maximum certainty value reached across the given period.

Table 1. Periodicities for event-related data

Timestamp(s)	*Maximum certainty value reached across this span (in %)*
6th July to 8th July	25
16th July to 24th July	34
27th July to 30th July	30
21st August to 6th September	32
17th September to 4th October	32

Working with time-series : The time-series used is a dataset containing average daily temperatures for Delhi (capital of India) from 1^{st} January, 1995 to 31^{st} December, 2008. It is available at www.engr.udayton.edu/weather/. The fields in the input dataset are : month, day, year, average daily temperature (F).

From the time-series, at first, the dynamic time-warping technique is used to extract time-periods (spanning across five to twenty five days) in which a 10 F temperature rise is detected.

While looking for yearly periodicities, several timestamps are detected having a 35% or more certainty of being involved in a 10 F temperature rise in Delhi. Some of these timestamps are shown in Table 2 along with the maximum certainty value reached across different timespans.

Table 2. Periodicities for time-series data

Timestamp(s)	*Maximum certainty value reached across this span (in %)*
1st February to 7th February	50
12th February to 16th February	43
3rd March to 8th March	64
14th March to 17th March	57
1st April to 3rd April	50
7th June to 10th June	43

5 Conclusion and Future Work

An algorithm with a worst-case time complexity of $\mathrm{O}(\mathrm{n}\log n)$ was proposed and tested to extract seasonal periodicities of patterns in different kinds of temporal data. The proposed algorithm succeeds in improving on the worst-case time of the algorithm given in [7]. The calendar-based hierarchical representation of the timestamps allow the same algorithm to be used for finding both partial and fully periodic patterns with equal efficiency. Though the algorithm uses the time-hierarchy associated with calendar dates, it can be used with any other time-hierarchy also. Future works can include formulation of formal measures to quantify the interestingness of periodic patterns. Another natural extension of this work is to cluster patterns having similar periodicity. Finding the clusters could lead to interesting rules since patterns having similar periodicity might also share some other features.

References

1. Ale, J.M., Rossi, G.H.: An Approach to Discovering Temporal Association Rules. In: Proceedings of 2000 ACM symposium on Applied Computing, vol. 1, pp. 294–300. ACM, New York (2000)

2. Elfeky, M.G., Aref, W.G., Elmagarmid, A.K.: Periodicity detection in time series databases. IEEE Transactions on Knowledge and Data Engineering 17(7), 875–887 (2005)
3. Lee, G., Yang, W., Lee, J.: A parallel algorithm for mining multiple partial periodic patterns. Information Sciences 176(24), 3591–3609 (2006)
4. Li, Y., Ning, P., Wang, X.S., Jajodia, S.: Discovering Calendar-based Temporal Association Rules. Data and Knowledge Engineering 44(2), 193–218 (2003)
5. Ma, S., Hellestein, J.: Mining Partially Periodic Event Patterns with Unknown Periods. In: Proceedings of 17th International Conference on Data Engineering, pp. 205–214. IEEE Computer Society, Los Alamitos (2001)
6. Mahanta, A.K., Mazarbhuiya, F.A., Baruah, H.K.: Finding Locally and Periodically Frequent Sets and Periodic Association Rules. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 576–582. Springer, Heidelberg (2005)
7. Mahanta, A.K., Mazarbhuiya, F.A., Baruah, H.K.: Finding calendar-based periodic patterns. Pattern Recognition Letters 29(9), 1274–1284 (2008)
8. Nguyen, C.L., Nelson, D.E.: Mining periodic patterns from floating and ambiguous time segments. In: IEEE International Conference on Man and Cybernetics, vol. 4, pp. 3622–3627 (2005)
9. Yang, J., Wang, W., Yu, P.S.: Mining Asynchronous Periodic Patterns in Time Series Data. In: Proceedings of 6th International Conference on Knowledge Discovery and Data Mining, pp. 275–279 (2000)
10. Zhang, M., Kao, B., Cheung, D.W., Yip, K.Y.: Mining periodic patterns with gap requirement from sequences. ACM Transactions on Knowledge Discovery from Data (TKDD), article 7 1(2) (2007)

A Relation Mining and Visualization Framework for Automated Text Summarization

Muhammad Abulaish[1,*], Jahiruddin[1], and Lipika Dey[2]

[1] Department of Computer Science, Jamia Millia Islamia, New Delhi, India
abulaish@ieee.org, jahir.jmi@gmail.com
[2] Innovation Labs, Tata Consultancy Services, New Delhi, India
lipika.dey@tcs.com

Abstract. In this paper, we present a relation mining and visualization framework to identify important semi-structured information components using semantic and linguistic analysis of text documents. The novelty of the paper lies in identifying key snippet from text to validate the interaction between a pair of entities. The extracted information components are exploited to generate semantic network which provides distinct user perspectives and allows navigation over documents with similar information components. The efficacy of the proposed framework is established through experiments carried out on biomedical text documents extracted through PubMed search engine.

Keywords: Relation mining, Text mining, Text visualization, Text summarization, Natural language processing.

1 Introduction

The rapidly growing repository of text information on any topic necessitates the design and implementation of strategies that enables fast and efficient access to relevant content. While search engines provide an efficient way of accessing relevant information, the sheer volume of the information repository on the Web makes assimilation of this information a potential bottleneck in the way its consumption. Thus, in the age of increasing information availability, it is essential to provide users with a convenient way to understand information easily. To achieve this goal, many techniques, including document classification and document summarization have been developed [7]. However, most of these methods only provide ways for users to easily access the information; they do not help users directly capture the key concepts and their relationships with the information. Understanding key concepts and their relationships is critical for capturing the conceptual structure of document corpora and avoiding information overload for users. Besides, development of intelligent techniques to collate the information extracted from various sources into a semantically related structure can aid the user for visualization of the content at multiple levels of complexity. Such a visualizer provides a semantically integrated view of the underlying text repository

[*] To whom correspondence should be addressed.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 249–254, 2009.

in the form of a consolidated view of the concepts that are present in the collection, and their inter-relationships as derived from the collection along with their sources. The semantic net thus built can be presented to the user at arbitrary levels of depth as desired.

In this paper, we propose a relation mining and visualization framework which uses linguistic and semantic analysis of text to identify key information components from text documents. The information components are centered on domain entities and their relationships, which are extracted using natural language processing techniques and co-occurrence-based analysis. The novelty of the paper lies in identifying key snippet from text to validate the interaction between a pair of entities. For example, consider the following sentence: "*Diallyl Sulfite (DAS) prevents cancer in animals (PMID: 8636192)*". In this sentence, "*prevents*" is identified as relational verb relating the biological entities "*Diallyl Sulfite (DAS)*" and "*cancer*" while "*animals*" is identified as key snippet to validate the interaction *prevents*(*Diallyl Sulfite (DAS), cancer)*. Our system also extracts the adverbs associated with relational verbs, which plays a very important role especially to identify the negation in sentences. We have also presented a scheme for semantic net generation which highlights the role of a single entity in various contexts and thus useful for a researcher as well as a layman. The efficacy of the proposed framework is established through experiment over documents from biomedical domain in which information overload also exists due to exponential growth of scientific publications and other forms of text-based data as a result of growing research activities. The remaining paper is structured as follows. Section 2 presents a brief introduction to related work. Section 3 presents the proposed framework and details about functioning of different modules. Section 4 presents the evaluation result of the proposed framework. Finally, section 5 concludes the paper with future directions.

2 Related Work on Relation Mining and Visualization

Context of entities in a document can be inferred from analysis of the inter-entity relations present in the document. Sekimizu *et al.* [5] proposed a linguistic-based approach to mine relationship among gene and gene products. They have collected the most frequently occurring verbs in a collection of abstracts and developed partial and shallow parsing techniques to find the verb's subject and object. Thomas *et al.* [6] modified a pre-existing parser based on cascaded finite state machines to fill templates with information on protein interactions for three verbs – *interact with, associate with, bind to*. The PASTA system [1] is a more comprehensive system that extracts relations between proteins, species and residues. Ono *et al.* [2] reports a method for extraction of *protein-protein interactions* based on a combination of syntactic patterns. Rinaldi *et al.* [4] have proposed an approach towards automatic extraction of a predefined set of seven relations in the domain of molecular biology, based on a complete syntactic analysis of an existing corpus.

Visualization is a key element for effective consumption of information. Semantic nets provide a consolidated view of domain concepts and can aid in this process.

In the information visualization literature, a number of exploratory visualization tools are described in [3]. Zheng et al. [7] have proposed an ontology-based visualization framework, GOClonto, for conceptualization of biomedical document collections.

Although, some visualization methods extract key concepts from document corpora, most of them do not explicitly exploit the semantic relationships between these concepts. The proposed method differs from all these approaches predominantly in its use of pure linguistic techniques rather than use of any pre-existing collection of entities and relations. Moreover, the knowledge visualizer module is integrated with the underlying corpus for comprehending the conceptual structure of biomedical document collections and avoiding information overload for users. On selecting a particular entity or relation in the graph the relevant documents are displayed with highlighting the snippet in which the target knowledge is embedded.

3 Proposed Relation Mining and Visualization Framework

Fig. 1 highlights the functional details of the proposed framework, which comprises of the following four main modules – *Document Processor*, *Concept Extractor*, *Information Component Extractor*, and *Knowledge Visualizer*. The design and working principles of these modules are presented in the following sub-sections.

3.1 Document Processor and Concept Extractor

The *document processor* cleans text documents by filtering unwanted chunks for Parts-Of-Speech (POS) analysis which assigns POS tags to every word in a sentence. The POS tags help in identifying the syntactic category of individual words. For POS analysis we have used the Stanford parser to convert every sentence into a phrase structure tree which is further analyzed by the *concept extractor module* to extract noun phrases. In order to avoid long noun phrases, which generally contain conjunctional words, we have considered only those noun words that appear at leaf- nodes in phrase structure tree. In case, more than one noun phrase appears at leaf node as siblings, the string concatenation function is applied to club them into a single noun phrase. Thereafter, based on a standard list of 571 stopwords, the words that occur frequently but have no meaning are removed. A phrase is removed from the list if it contains a stopword either at beginning or at end position. After extracting all noun phrases from all text documents in the corpus their weights are calculated as a function of term frequency (tf) and inverse document frequency (idf) by using the equation $\omega(p) = tf \times \log(N/n)$, where $\omega(p)$ represents the weight of the noun phrase *P*, *tf* denotes the total number of times the phrase *P* occurs in the corpus, $\log(N/n)$ is the idf of *P*, *N* is the total number of documents in the corpus, and *n* is the number of documents that contain *P*. A partial list of noun phrases in descending order of their weights extracted from a corpus of 500 PubMed abstracts on "Breast Cancer" is: *breast cancer*, *breast carcinoma*, *prostate cancer*, *zoledronic acid*, *tumor*, *estrogen*, *prognosis*, *tomoxifen*, *protein*.

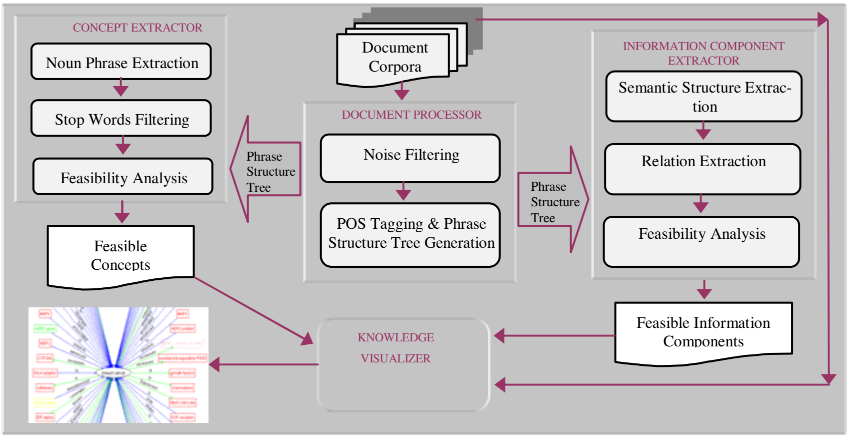


Fig. 1. Relation mining and visualization framework

3.2 Information Components Extractor

An information component can be defined formally as a 7-tuple of the form <$\mathcal{E}_i$, $\mathcal{A}$, $\mathcal{V}$, $\mathcal{P}_v$, $\mathcal{E}_j$, $\mathcal{P}_c$, $\mathcal{E}_k$> where, $\mathcal{E}_i$, $\mathcal{E}_j$, and $\mathcal{E}_k$ are biological entities identified as noun phrases; $\mathcal{E}_i$ and $\mathcal{E}_j$ forms the *subject* and *object* respectively for $\mathcal{V}$, $\mathcal{A}$ is adverb; $\mathcal{V}$ is relational verb, $\mathcal{P}_v$ is verbal-preposition associated with $\mathcal{V}$; $\mathcal{P}_c$ is conjunctional-preposition linking $\mathcal{E}_j$ and $\mathcal{E}_k$. The information component extraction module traverses the phrase structure tree and analyzes phrases and their linguistic dependencies in order to trace relational verbs and other constituents. Since entities are marked as noun phrases in the phrase structure tree, this module exploits phrase boundary and proximitivity to identify relevant information components. Initially all tuples of the form <$\mathcal{E}_i$, $\mathcal{A}$, $\mathcal{V}$, $\mathcal{P}_v$, $\mathcal{E}_j$, $\mathcal{P}_c$, $\mathcal{E}_k$> are retrieved from the documents. Thereafter, feasibility analysis is applied to filter out non-relevant verbs and thereby the corresponding information component.

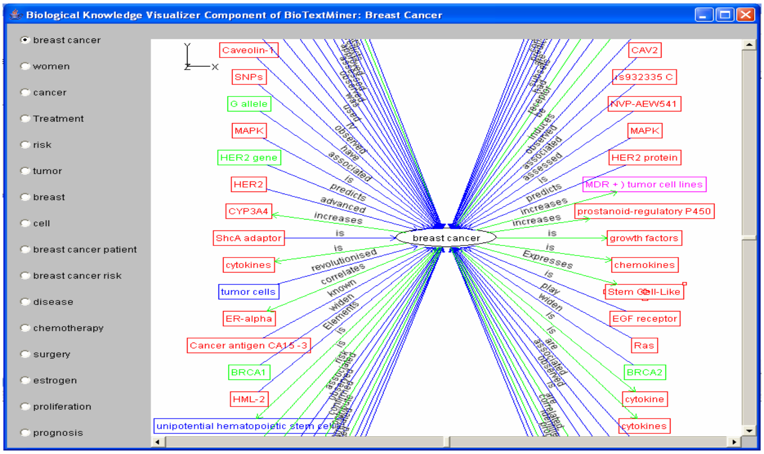
Rather than focusing only on root verbs, morphological variants of a relational verb and associated prepositions are also recognized by our system. Information component extraction process is implemented as a rule-based system. Four sample information components extracted from PubMed documents are presented in Table 1.

Table 1. Four sample information components extracted from PubMed documents

Left Entity	Adv.	Rel. Verb	Verb Prep	Right Entity	Conj Prep	Key Snippet
AO enzymes	---	associated	with	breast cancer and aging	---	---
the levels of MDC1 and BRIT1	---	correlated	with	centrosome amplification, defective mitosis and cancer metastasis	in	human breast cancer
oral glutamine (GLN)	---	inhibited	---	tumor growth	through	Stimulation of host
BMP-4	not	stimulate	---	cell proliferation	by	itself

3.3 Biological Knowledge Visualizer

The major idea of visualization using semantic net is to highlight the role of a concept in a text corpus by eliciting its relationship to other concepts. The nodes in a semantic net represent entities/concepts and links indicate relationships. The whole graph is centered around a selected entity by the user. For a relation triplet <$\mathcal{E}_i$, $\mathcal{R}$, $\mathcal{E}_j$>, which is a subset of information component (IC), the entities $\mathcal{E}_i$ and $\mathcal{E}_j$ are used to define classes and $\mathcal{R}$ is used to define relationships between them. The other constituents of the IC can be used to handle biomedical queries which design is one of our future works. While displaying $\mathcal{E}_i$ and $\mathcal{E}_j$, the biological entities are marked with ABNER – a biological entity recognizer. To define a relationship map, the user selects an entity, say ξ, around which the graph is to be created. The selected entity ξ is used to extract all those ICs which contain ξ either as a part of $\mathcal{E}_i$, or $\mathcal{E}_j$ or both. A snapshot of the semantic net generated around the biological entity "breast cancer" is shown in Fig. 2. The semantic net is also designed to facilitate users to navigate through the pile of documents in an efficient way. While double-clicking on a node all the ICs containing the entity stored in that node are selected. Thereafter, the documents containing these ICs are displayed in which the relevant parts of the sentences are highlighted. Similarly, on double-clicking an edge, all the ICs centered around the relation appearing as edge label are selected. Thereafter, the documents containing these ICs are displayed with properly highlighting the relevant snippet of text. The ICs are also displayed separately in the bottom pan of the same window.



Biological Knowledge Visualizer Component of BioTextMiner: Breast Cancer

● breast cancer
○ women
○ cancer
○ Treatment
○ risk
○ tumor
○ breast
○ cell
○ breast cancer patient
○ breast cancer risk
○ disease
○ chemotherapy
○ surgery
○ estrogen
○ proliferation
○ prognosis

Fig. 2. Semantic net generated around *"breast cancer"*

4 Evaluation of the Relation Extraction Process

The performance of the whole system is analyzed by taking into account the performance of the relation extraction process. A relation is said to be "correctly identified" if its occurrence within a sentence along with its left and right entities is grammatically correct, and the system has been able to locate it in the right context. To judge the performance of the system, it is not enough to judge the extracted relations only, but it is also required to analyze all the correct relations that were missed by the system. The system was evaluated for its *recall* and *precision* values for ten randomly selected relation triplets. The precision and recall value of the proposed system is found to be 88.77% and 84.96% respectively. On analysis, it was found that most of the incorrect identifications and misses occur when the semantic structure of a sentence is wrongly interpreted by the parser.

5 Conclusion and Future Work

In this paper, we have proposed the design of a relation mining and visualization framework which uses linguistic and semantic analysis of text to identify feasible information components from unstructured text documents. We have also proposed a method for collating information extracted from multiple sources and present them in an integrated fashion with the help of semantic net. Presently, we are refining the rule-set to improve the *precision* and *recall* values of the system. Moreover, a query processing module is being developed to handle biomedical queries on underlying repository of extracted information components.

References

1. Gaizauskas, R., Demetriou, G., Artymiuk, P.J., Willett, P.: Protein Structures and Information Extraction from Biological Texts: the PASTA System. Bioinformatics 19(1), 135–143 (2003)
2. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated Extraction of Information on Protein-Protein interactions from the Biological Literature. Bioinformatics 17(2), 155–161 (2001)
3. Plaisant, C., Fekete, J.-D., Grinsteinn, G.: Promoting Insight-based Evaluation of Visualizations: From Contest to Benchmark Repository. IEEE Transactions on Visualization and Computer Graphics 14(1), 120–134 (2008)
4. Rinaldi, F., Scheider, G., Andronis, C., Persidis, A., Konstani, O.: Mining Relations in the GENIA Corpus. In: Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy, pp. 61–68 (2004)
5. Sekimizu, T., Park, H.S., Tsujii, J.: Identifying the Interaction between Genes and Genes Products based on Frequently Seen Verbs in Medline Abstract. Genome Informatics 9, 62–71 (1998)
6. Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carroll, M.: Automatic Extraction of Protein Interactions from Scientific Abstracts. In: Pacific Symposium on Biocomputing, pp. 538–549 (2000)
7. Zheng, H.-T., Borchert, C., Kim, H.-G.: Exploiting Gene Ontology to Conceptualize Biomedical Document Collections. In: Domingue, J., Anutariya, C. (eds.) ASWC 2008. LNCS, vol. 5367, pp. 375–389. Springer, Heidelberg (2008)

Mining Local Association Rules from Temporal Data Set

Fokrul Alom Mazarbhuiya[1], Muhammad Abulaish[2,*], Anjana Kakoti Mahanta[3], and Tanvir Ahmad[4]

[1] College of Computer Science, King Khalid University, Abha, KSA
fokrul_2005@yahoo.com
[2] Department of Computer Science, Jamia Millia Islamia, Delhi, India
abulaish@ieee.org
[3] Department of Computer Science, Gauhati University, Assam, India
anjanagu@yahoo.co.in
[4] Department of Computer Engineering, Jamia Millia Islamia, Delhi, India
tanvir.ce@jmi.ac.in

Abstract. In this paper, we present a novel approach for finding association rules from locally frequent itemsets using rough set and boolean reasoning. The rules mined so are termed as *local association rules*. The efficacy of the proposed approach is established through experiment over retail dataset that contains retail market basket data from an anonymous Belgian retail store.

Keywords: Data mining, Temporal data mining, Local association rule mining, Rough set, Boolean reasoning.

1 Introduction

Mining association rules in transaction data is a well studied problem in the field of data mining. In this problem, given a set of items and a large collection of transactions, the task is to find relationships among items satisfying a user given support and confidence threshold values. However, the transaction data are temporal in the sense that when a transaction happens the time of transaction is also recorded. Considering the time aspect, different methods [1] have been proposed to extract temporal association rules, i.e., rules that hold throughout the life-time of the itemset rather than throughout the life-time of the dataset. The lifetime of an itemset is the time period between the first transaction containing the itemset and the last transaction containing the same itemset in the dataset and it may not be same as the lifetime of the dataset. Mahanta *et al.* have addressed the problem of temporal association rule extraction in [2]. They proposed an algorithm for finding frequent itemsets with respect to a small time-period not necessarily equal to the lifetime of the dataset or that of the itemset. They named such itemsets as *locally frequent itemsets* and corresponding rules as *local association rules*. In order to calculate the confidence value of a local association rule, say A ⇒ X – A, in the interval [t, t'] where X is a frequent itemset in [t, t'] and $A \subset X$, it is required to know the supports of both X and A in the same interval

[*] To whom correspondence should be addressed.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 255–260, 2009.

[t, t']. But, the way supports of itemsets are calculated in [2], the support of subsets of X may not be available for the same time interval rather, they may be frequent in an interval greater than [t, t']. So, they have loosely defined association rules, as confidence of the rule A ⇒ X – A cannot be calculated in interval [t, t'] directly.

Rough sets theory, proposed by Pawlak [3], seems to be a solution to this problem. Nguyen *et al.* [4] have presented a method of extracting association rules, based on rough set and boolean reasoning. They have shown a relationship between association rule mining problem and reducts finding problem in rough set theory. But, their works were mainly focused on non-temporal datasets.

In this paper, we present a novel approach for finding *local association rules* from locally frequent itemsets using rough set and boolean reasoning. For a given locally frequent itemset X in time interval [t, t'], all those transactions generated between t and t' are considered and mapped to decision table in line with [4]. Thereafter, we find the reducts using rough set theory and boolean reasoning to generate association rules that would be local to the time interval [t, t']. The rest of the paper is organized as follows. Section 2 presents the related works on temporal association rule mining. Basic concepts, definitions and notations are presented in section 3. The proposed local association rule mining method is described in section 4. The experimental setup is presented in section 5. Finally, section 6 concludes the paper.

2 Related Work

Temporal Data Mining is an important extension of conventional data mining. By taking into account the time aspect, more interesting patterns that are time dependent can be extracted. Hence, the association rule discovery process is extended to incorporate temporal aspects. Each temporal association rule has associated with it a time interval in which the rule holds. In [1], an algorithm for discovery of temporal association rules is described. For each item (which is extended to item set) a lifetime or life-span is defined as the time gap between the first occurrence and the last occurrence of the item in transaction database. Supports of items are calculated only during its life-span. Thus each rule has associated with it a time frame corresponding to the lifetime of the items participating in the rule. In [2], the works done in [1] has been extended by considering time gap between two consecutive transactions containing an item set into account. The frequent itemsets extracted by above method are termed as locally frequent itemsets. Although the methods proposed in [1] and [2] can extract more frequent itemsets than others; the methods did not address association rules extraction problem adequately. The relationship between the problem of association rules generation from transaction data and relative reducts finding from decision table using rough set theory is better presented in [4,5,6,7]. But, the temporal attribute which is naturally available in a transaction dataset is not taken into consideration.

3 Basic Concepts, Definitions and Notations

The *local support* of an itemset, say X, in a time interval [t_1, t_2] is defined as the ratio of the number of transactions in the time interval [t_1, t_2] containing the item set to the total number of transactions in [t_1, t_2] for the whole dataset D and is denoted by $\sup_{[t_1,t_2]}(X)$. Given a threshold σ, an itemset X is said to be frequent in the time interval [t_1, t_2] if $\sup_{[t_1,t_2]}(X) \geq (\sigma / 100) \times |D|$ where $|D|$ denotes the total number of transactions in D that are in the time interval [t_1, t_2]. The itemset X is termed as *locally frequent* in [t_1, t_2]. An association rule $X \Rightarrow Y$, where X and Y are item sets said to hold in the time interval [t_1, t_2] if and only if for a given threshold τ, $\sup_{[t_1,t_2]}(X \cup Y) / \sup_{[t_1,t_2]}(X) \geq \tau / 100$ and $X \cup Y$ is frequent in [t_1, t_2]. In this case we say that the confidence of the rule is τ.

An *information system* is a pair $S=(U, A)$, where U is a non-empty finite set called the universe and A is a non-empty finite set of attributes. Each $a \in A$ corresponds to the function $a{:}U \rightarrow V_a$, where V_a is called the value set of a. Elements of U are called *situations*, *objects* or *rows*, interpreted as, *cases*, *states*, *patients*, or *observations*.

A *decision table* is a special type of information system and is denoted by $S=(U, A \cup \{d\})$, where $d \notin A$ is a distinguishing attribute called the *decision*. The elements of A are called conditional attributes (conditions). In our case, each $a \in A$ corresponds to the function $a{:}U \rightarrow V_a = \{0, 1\}$, because we are considering only presence or absence of items in the transactions. In addition, A contains another attribute called time-stamp i.e. $A=A' \cup \{t\}$, where t indicates a valid time at which a transaction occurs.

4 Method of Generating Local Association Rules

In this section, we discuss the method of temporal template mining and thereafter local association rule mining from them using rough set and boolean reasoning.

4.1 Template as Patterns in Data

By template we mean the conjunction of descriptors. A descriptor can be defined as a term of the form $(a=v)$, where $a \in A$ is an attribute and $v \in V_a$ is a value from the domain of a. For a given template T the object $u \in U$ satisfies T iff all the attribute values of T are equal to the corresponding attribute values of u. In this way a template T describes the set of objects having common properties. The support of a template T is defined as: support(T)=|{u∈U: u satisfies T}|. A template T is called good template if the support(T) $\geq s$ for a given threshold value s. A template is called temporal template if it is associated with a time interval [t, t']. We denote a temporal template associated with the time-interval [t, t'] as $T[t, t']$. A temporal template may be “good” in a time-interval which may not be equal to the lifetime of the information table. The procedure of finding temporal template is discussed in [2]. From descriptive point of view, we prefer long templates with large support.

4.2 From Template to Optimal Association Rules

Let us assume that a temporal template $T[t, t'] = D_1 \wedge D_2 \wedge \ldots \wedge D_m$ with support s has been found using [2]. We denote the set of all descriptors occurring in template T by DESC(T[t, t']) which is defined as: DESC(T[t, t'])=$\{D_1 \wedge D_2 \wedge \ldots \wedge D_m\}$. Any set $P \leq$ DESC(T[t, t']) defines an association rule $R_P = \mathrm{def}(\wedge_{D_i \in P} D_i \Rightarrow \wedge_{D_j \notin P} D_j)$. For a given confidence threshold $c \in (0, 1]$ and a given set of descriptors P≤ DESC(T[t, t']), the temporal association rule R_P is called c-representative if (i) confidence(R_P)≥c, and (ii) for any proper subset P′ of P we have confidence($R_{P'}$)≤c. Instead of searching for all temporal association rules we search for c-representative temporal association rules because every c-representative temporal association rule covers a family of temporal association rules. Moreover the shorter is temporal association rule R, the bigger is the set of temporal association rules covered by R.

4.3 Searching for Optimal (Shortest) Local Association Rules

In order to find association rules from a locally frequent itemset, say *X*, in an interval [t, t'], all the transactions (say *A*) that happened between t and t' are considered to construct a decision table. Thereafter, α-reducts for the decision table which corresponds to the local association rules are found using rough set theory. The decision table A/X[t, t'] from the transactions falling between t and t', X[t, t'], can be constructed as follows:

A/X[t, t']={ a_{D_1} , a_{D_2} , ….., a_{D_m} } is a set of attributes corresponding to the descriptors of template X[t, t']. The values of a_{D_i} is determined using equation 1. The decision attribute d determines if a given transaction supports template X[t, t'] and its value is determined using equation 2.

$$a_{D_i}(t) = \begin{cases} 1, \text{if the transaction occurance time t} \in [t,t'] \\ 0, \text{otherwise} \end{cases} \tag{1}$$

$$d(t) = \begin{cases} 1, \text{if t} \in [t,t'] \text{ satisfies X} \\ 0, \text{otherwise} \end{cases} \tag{2}$$

4.3.1 The Approximate Algorithms

In this section, we present two algorithms - the first, shown in table 1, finds *almost shortest c-representative association rules*. After the algorithm in table 1 stops we do not have any guarantee that the descriptor set *P* is c-representative. But one can achieve it by removing from *P* all unnecessary descriptors. The second algorithm, shown in table 2, finds *k* short c-representative association rules where *k* and *c* are parameters given by the user.

Table 1. Short c-representative association rule algorithm

```
Algorithm: Short c-Representative Association Rule
Input:  Information table $A$, template $T[t_1, t_2]$, minimal confidence c.
Output: short c-representative temporal association rule.
Steps:
1.  Set := ∅; $U_P$: = $U$; min_support: = $|U|$ -1/c.support($T[t_1, t_2]$)
2.  Choose a descriptor $D$ from DESC($T[t_1, t_2]$)\$P$ which is satisfied by the smallest number of
    objects from $U_P$
3.  Set $P$ : = $P \cup \{D\}$
4.  $U_P$ := satisfy($P$); (i.e. set of objects satisfying all descriptors from $P$)
5.  If $|U_P|$ > min_support then go to Step 2 else stop
```

Table 2. *k* short c-representative association rules

```
Algorithm: k Short c-Representative Association Rules

Input: Information table $A$, template $T[t_1, t_2]$, minimal confidence c ∈ (0, 1], number of representative rules $k \in N$
Output: $k$ short c-representative temporal association rules $R_{P1}$,… $R_{Pk}$
Steps:
1.  For i : = 1 to k do
2.  Set $P_i$ := ∅; $U_{p_i}$ : = $U$
3.  End for
4.  Set min_support : = $|U|$-1/c.support($T$)
5.  Result_set : = ∅; Working_set := {$P_1$,…,$P_k$}
6.  Candidate_set := ∅
7.  for ($P_i$∈Working_set) do
8.  Chose k descriptors $D_1^i$,….,$D_k^i$ from DESC($T[t_1, t_2]$)\$P_i$ which is satisfied by smallest number
    of objects from $U_{p_i}$
9.  insert $P_i$ ∪{ $D_1^i$},….. $P_i$ ∪{$D_k^i$} to the Candidate_set
10. end for
11. Select k descriptor sets $P_1'$,…,$P_k'$ from the Candidate_set (if exist) which are satisfied by
    smallest number of objects from $U$
12. Set Working_set := {$P_1'$,…,$P_k'$}
13. for ($P_i$∈Working_set) do
14. Set $U_P$ := satisfy($P_i$)
15. if $|U_{Pi}|$< min_support then
16. Move $P_i$ from Working_set to the Result_set
17. End for
18. if |Result_set| > k or Working_set is empty then STOP else GO TO Step 4
```

5 Results

For experimentation purpose we have used a retail datasets that contains retail market basket data from an anonymous Belgian retail store. The dataset contains 88162 transactions and 17000 items. As the dataset in hand is non-temporal, a new attribute “time” was introduced. The domain of the time attribute was set to the calendar dates from 1-1-2000 to 31-3-2003. For the said purpose, a program was written using C++ which takes as input a starting date and two values for the minimum and maximum number of transactions per day. A number between these two limits are selected at random and that many consecutive transactions are marked with the same date so that many transactions have taken place on that day. This process starts from the first transaction to the end by marking the transactions with consecutive dates (assuming

Table 3. A partial view of generated association rules from *retail* dataset

Association Rules	Corresponding intervals where the rules hold
100%-representative Association Rules	
{38, 39}⇒{41}	[2-1-2000, 25-5-2001]
{38, 41}⇒{39}	[2-1-2000, 25-5-2001], [31-8-2002, 22-3-2003]
{41, 48}⇒{39}	[2-1-2000, 29-5-2001], [31-8-2002, 22-3-2003]
75%-representative Association Rules	
{39}⇒{41}	[2-1-2000, 29-5-2001], [31-8-2002, 22-3-2003]
{12935}⇒{39}	[13-2-2002, 22-3-2003]
{41, 48}⇒{39}	[2-1-2000, 29-5-2001]
50%-representative Association Rules	
{32}⇒{39}	[2-1-2000, 22-3-2003]
{32, 48}⇒{39}	[2-1-2000, 22-3-2003]
{41}⇒{39, 48}	[2-1-2000, 29-5-2001], [31-8-2002, 22-3-2003]
25%-representative Association Rules	
{41}⇒{32}	[2-1-2000, 25-5-2003]
{32}⇒{39, 48}	[2-1-2000, 22-3-2003]
{39, 41] ⇒{38}	[2-1-2000, 25-5-2001], [31-8-2002, 22-3-2003]

that the market remains open on all week days). This means that the transactions in the dataset are happened in between the specified dates. A partial view of the generated association rules from *retail* dataset is shown in table 3.

6 Conclusion

In this paper, we have proposed a novel approach for finding *local association rules* from locally frequent itemsets using rough set and boolean reasoning. To generate association rules from a locally frequent itemset in the interval $[t, t']$, first all transactions that occurs between t and t' are considered to form an information system. Later, the information system is converted into decision table to find the reducts and α-reducts.

References

1. Ale, J.M., Rossi, G.H.: An Approach to Discovering Temporal Association Rules. In: Proceedings of ACM symposium on Applied Computing, pp. 294–300 (2000)
2. Mahanta, A.K., Mazarbhuiya, F.A., Baruah, H.K.: Finding Locally and Periodically Frequent Sets and Periodic Association Rules. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 576–582. Springer, Heidelberg (2005)
3. Paulak, Z.: Rough Sets in Theoretical Aspects of Reasoning about Data. Kluwer, Netherland (1991)
4. Nguyen, H.S., Nguyen, S.H.: Rough Sets and Association Rule Generation. Fundamenta Informaticae 40(4), 383–405 (1999)
5. Nguyen, H.S., Slezak, D.: Approximate Reducts and Association Rules- Correspondence and Complexity Results. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 137–145. Springer, Heidelberg (1999)
6. Skowron, A., Rauszer, C.: The Dicernibility Matrices and Functions in Information Systems. In: Slowinski, R. (ed.) Intelligent Decision support, Handbook of Applications and Advances of the Rough Sets Theory, pp. 331–362. Kluwer, Dordrecht (1992)
7. Wróblewski, J.: Covering with Reducts - A Fast Algorithm for Rule Generation. In: Polkowski, L., Skowron, A. (eds.) RSCTC 1998. LNCS (LNAI), vol. 1424, pp. 402–407. Springer, Heidelberg (1998)

Multi-label Text Classification Approach for Sentence Level News Emotion Analysis

Plaban Kr. Bhowmick, Anupam Basu, Pabitra Mitra, and Abhishek Prasad

Indian Institute of Technology, Kharagpur, India-721302

Abstract. Multiple emotions are often evoked in readers in response to text stimuli like news article. In this paper, we present a novel method for classifying news sentences into multiple emotion categories using Multi-Label K Nearest Neighbor classification technique. The emotion data consists of 1305 news sentences and the emotion classes considered are disgust, fear, happiness and sadness. Words and polarity of subject, verb and object of the sentences and semantic frames have been used as features. Experiments have been performed on feature comparison and feature selection.

1 Introduction

The Internet has been one of the primary media for information dissemination shortly after the advent of Word Wide Web. Consequently, the amount of text data available online is enormous. The advent of new technologies makes way of new interaction possibilities and provides people with an array of social media like blog, chat, social network, news etc. As compared to traditional keyword based or topical access to the information, new social interactions require the information to be analyzed in social dimensions like emotion, sentiment, attitude, belief etc. Emotion analysis of text is one of such tasks that are gaining importance in text mining community in recent times.

Two types of emotion evaluation may be possible: writer perspective and reader perspective evaluation. In previous works [1], a few attempts towards writer perspective analysis of emotion in text data have been made. In all these studies, it has generally been assumed that the writer expresses only one emotion for a text segment. However, evocation of a blend of emotions is common in reader in response to a stimulus. For example, the following sentence may evoke *fear* and *sad* emotion in readers mind.

`Militant attack kills over 30 persons in Nigeria.`

Classifying emotion from reader's perspective is a challenging task and research on this topic is relatively sparse as compared to writer perspective analysis.

Affective text analysis was the task set in *SemEval-2007 Task 14* [2]. A corpus of news headlines extracted from Google news and CNN was provided. Some systems with different approaches have participated in solving the said task [2].

The work [3] provides the method for ranking reader's emotions in Chinese news articles from Yahoo! Kimo News. Eight emotional classes are considered

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 261–266, 2009.

in this work. Support Vector Machine (SVM) has been used as the classifier. Chinese character bigram, Chinese words, news metadata, affix similarity and word emotion have been used as features. The best reported system accuracy is 76.88%.

In this work, we perform reader perspective emotion analysis in text data where one text segment may evoke more than one emotion in reader. News is a media where certain facts in the articles are presented to the readers with the expectation that the articles evoke some emotional responses in the readers. So, this media is one potential data source for the computational study of reader perspective emotion.

2 Emotion Classification Problem and MLκNN

The problem of multi-label emotion classification is defined as follows: Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of emotional sentences and $\mathcal{E} = \{e_i | i = 1, 2, \ldots, |\mathcal{E}|\}$ be the set of emotion classes (e.g., happy, sad etc.). The task is to find a function $h : S \mapsto 2^{\mathcal{E}}$, where $2^{\mathcal{E}}$ is the powerset of $\mathcal{E}$.

This problem can be mapped to a multi-label text classification problem. In this work, we use Multi-Label κ Nearest Neighbor classifier (MLκNN) [4] for classifying sentences into emotion classes. MLκNN, a multi-label adaptation of single label κ Nearest Neighbor algorithm, is one of the state of the art high performance algorithm adaptation technique. In this technique, for each test instance t, its K nearest neighbors in the training set are identified. Then according to statistical information gained from the label sets of these neighboring instances maximum a posteriori (MAP) principle is utilized to determine the label set for the test instance. The entities that central to this classification technique are the prior probabilities $P(H_b^l)(l \in \mathcal{E}), b \in \{0, 1\})$ and the posterior probabilities $P(N_j^l | H_b^l)\,(j \in \{0, 1, \ldots, \kappa\})$. Here, H_1^l is the event that the test instance has label l, while H_0^l denotes the event that t has not label and $N_j^l\,(j = 1, 2, \ldots, \kappa\,)$ is denotes that among the κ nearest neighbors of t, there are exactly j instances which have label l. The probability values are estimated from training data set. Laplace smoothing is used for handling data sparsity problem.

3 Emotion Data

The emotion text corpus consists of 1305 sentences extracted from *Times of India* news paper archive[1]. The emotion label set consists of four emotions: disgust, fear, happiness and sadness. A sentence may trigger multiple emotions simultaneously. So, one annotator may classify a sentence into more than one emotion categories.

The distribution of sentences across emotion categories is as follows: Disgust = 307, Fear = 371, Happiness = 282 and Sadness = 735. The *label density* (LD = $\frac{1}{|S|}\sum_{i=1}^{|S|}\frac{|\mathcal{E}_i|}{|\mathcal{E}|}$, where $\mathcal{E}_i$ is the label set for S_i) and *label cardinality*

[1] `http://timesofindia.indiatimes.com/archive.cms`

(LC $= \frac{1}{|S|}\sum_{i=1}^{|S|}|\mathcal{E}_i|$) are the measures of how multi-label the data is. The LC and LD for this data set are computed to be 1.3 and 0.26 respectively.

4 Features for Emotion Classification

Three types of features have been considered in our work as given below:

- *Word Feature (W)*: Words sometimes are indicative of the emotion class of a text segment. For example, the word 'bomb' may be highly co-associated with fear emotion. Thus words present in the sentences are considered as features. Before creating the word feature vectors, stop words and named entities are removed and the words are stemmed using Porter's stemmer.
- *Polarity Feature (P)*: Polarity of the subject, object and verb of a sentence may be good indicators of the emotions evoked. The subject, object and verb of a sentence is extracted from its parse tree and the polarity for each phrase is extracted from manual word level polarity tagging with a set of simple rules.
- *Semantic Frame Feature (SF)*: The Berkeley FrameNet project[2] is a well-known resource of frame-semantic lexicon for English. Apart from storing the predicate-argument structure, the frames group the lexical units. For example, the terms 'kill', 'assassin' and 'murder' are grouped into a single semantic frame 'Killing'. In this work, we shall be exploring the effectiveness of the semantic frames feature in emotion classification. The semantic frame assignment was performed by SHALMANESER[3].

5 Evaluation Measures

We evaluate our emotion classification task with respect to different sets of multi-label evaluation measures:

- Example based measures: Hamming Loss (HL), Partial match accuracy (P-Acc), Subset accuracy (S-Acc) and F1. These measures are explained in the work [5].
- Ranking based measures: One Error (OE), Coverage (COV), Average Precision (AVP). The work [4] describes these measures in detail.

6 Experimental Results

In this section, we present results of experiments of emotion classification with MLkNN. 5-fold cross-validation has been performed in all the experiments and the number of neighbors considered is 10.

[2] http://framenet.icsi.berkeley.edu/

[3] http://www.coli.uni-saarland.de/projects/salsa/shal/

6.1 Comparison of Features

The comparison of the features is performed with respect to a baseline which considers only words as features. Table 1 summarizes the results of emotion classification with different features and their combinations with best results presented in bold face.

Table 1. Comparison of features (W = word feature, P = polarity feature, SF = Semantic frame feature)

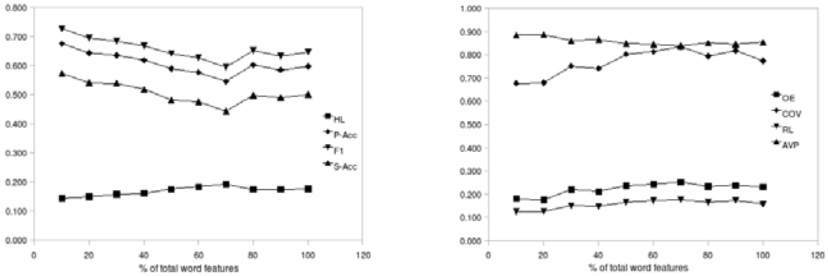
Measure Type	Measure	W	P	SF	W+P	P+SF	W+SF	W+P+SF
Example based measures	HL	0.175	0.223	**0.117**	0.156	0.118	0.131	0.126
	P-Acc	0.598	0.532	0.756	0.664	**0.764**	0.703	0.722
	F1	0.647	0.582	0.806	0.714	**0.817**	0.756	0.774
	S-Acc	0.499	0.428	0.653	0.563	**0.655**	0.594	0.618
Ranking based measures	OE	0.231	0.325	0.143	0.210	**0.129**	0.145	0.133
	COV	0.774	0.918	0.576	0.704	**0.538**	0.606	0.576
	RL	0.157	0.207	0.091	0.138	**0.079**	0.101	0.090
	AVP	0.853	0.801	0.911	0.871	**0.921**	0.906	0.915

General observations over the feature comparison experiment are as follows.

- The use of semantic frames (**SF**) as features improves the performance of emotion classification significantly ($\triangle$P-Acc = 15.8%, $\triangle$S-Acc = 15.4% and $\triangle$F1 = 15.9%) over the baseline. This significant improvement may be attributed to two different transformations over the word feature set.
  - *Dimensionality Reduction*: There is a significant reduction in dimension when (**SF**) feature is considered instead of (**W**) feature (SF feature dimension = 279 and W feature dimension = 2345).
  - *Feature Generalization*: Semantic frame assignment to the terms in the sentences is one of generalization technique where conceptually similar terms are grouped into a semantic frame. In semantic frame feature set, unification of these features are performed resulting in less skewedness in feature distribution.
- The **P+SF** feature combination performs best.
- The polarity feature (**P**) is inefficient as compared to other combinations but whenever coupled with other feature combinations (i.e., **W** vs. **W+P**, **SF** vs. **SF+P** and **W+SF** vs. **W+SF+P**), the performance improves.
- Whenever **W** feature combines with **SF**, degradation in performance have been observed (i.e., **SF** vs. **W+SF**, **P+SF** vs. **W+P+SF**).

6.2 Feature Selection

The plot word feature χ^2 value vs. rank follows the Zipfian distribution (power law fit with equation $y = \alpha x^{-\beta}$ where $\alpha = 236.43$, $\beta = 0.8204; R^2 = 0.89$) having a long tail which is strong indication of feature sparseness problem. To



(a) Example based measures Vs. Percentage of total word features

(b) Ranking based measures Vs. Percentage of total word features

Fig. 1. Variation of multi-label measures with Percentage of total word features

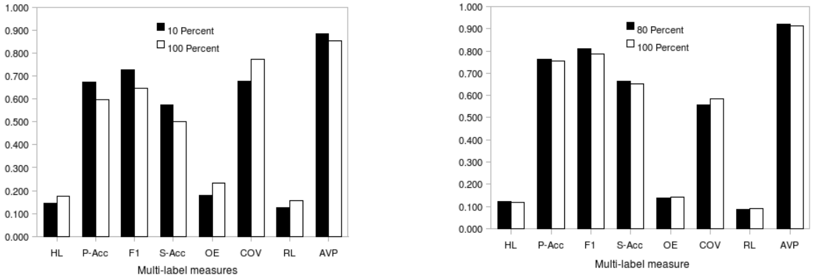
alleviate this problem, we have performed χ^2 feature selection [6] on the **W** and **SF** feature sets.

We performed experiment on selecting optimal **W** feature set size based on their χ^2 values. We present the variations of the performance measures with **W** feature set size in Fig. 1.

Top 10% of the total **W** feature set is found to be optimal feature set. The relative performance after feature selection for **W** is shown in Fig. 2(a). In case of **SF** feature, top 80% out of the total set was selected as optimal feature set for **SF** feature. The relative performance with the selected **SF** feature set is presented in Fig 2(b).

It is evident from Fig. 2 that the there is a slight improvement in performance after adopting feature selection strategy for both the feature sets.

With **P+SF** feature combination being the close competitor, best performance is achieved with **P+80%SF**(HL = 0.115, P-Acc = 0.773, F1 = 0.827 and S-Acc = 0.666). As the emotion analysis task is modeled in a multi-label framework, comparison with other systems can only be made with micro-average



(a) Relative performance after χ^2 word feature selection

(b) Relative performance after χ^2 semantic frame feature selection

Fig. 2. Performance after χ^2 feature selection

Table 2. Comparison of the proposed system with others

Measure	Accuracy	Precision	Recall	F1
UPAR7	89.43	27.56	5.69	9.43
UA-ZBSA	85.72	17.83	11.27	13.81
SWAT	88.58	19.46	8.62	11.95
Li and Chen	76.88	–	–	–
Our System	**88.2**	**84.42**	**79.93**	**82.1**

measures like accuracy, precision, recall and F1. The comparison with other systems is presented in Table 2.

7 Conclusions

We have presented an extensive comparison of different features for multi-label reader perspective emotion classification with MLkNN. Feature selection on word and semantic frame features is found to be fruitful for better performance. The classifier performs better with semantic frame (SF) features compared to word features (W) as SF helps in dealing with feature sparseness problem. Improvements have been noticed when the polarity feature is combined with other features.

References

1. Abbasi, A., Chen, H., Thoms, S., Fu, T.: Affect analysis of web forums and blogs using correlation ensembles. IEEE Transactions on Knowledge and Data Engineering 20(9), 1168–1180 (2008)
2. Strapparava, C., Mihalcea, R.: Semeval-2007 task 14: Affective text. In: Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007), Prague, Czech Republic (June 2007)
3. Lin, K.H.Y., Yang, C., Chen, H.H.: Emotion classification of online news articles from the reader's perspective. In: Web Intelligence, pp. 220–226 (2008)
4. Zhang, M., Zhou, Z.h.: Ml-knn: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2007 (2007)
5. Tsoumakas, G., Vlahavas, I.: Random *k*-labelsets: An ensemble method for multilabel classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007)
6. Zheng, Z., Wu, X., Srihari, R.: Feature selection for text categorization on imbalanced data. Explorations Newsletter 6(1), 80–89 (2004)

Semantic Relation between Words with the Web as Information Source

Tanmay Basu and C.A. Murthy

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
{tanmaybasu_r,murthy}@isical.ac.in

Abstract. Semantic relation is an important concept of information science. Now a days it is widely used in semantic web. This paper aims to present a measure to automatically determine semantic relation between words using web as knowledge source. It explores whether two words are related or not even if they are dissimilar in meaning. The proposed formula is a function of frequency of occurrences of the words in each document in the corpus. This relationship measure will be useful to extract semantic information from the web . Experimental evaluation on ten manually selected word pairs using the WebKb data as information source demonstrates the effectiveness of the proposed semantic relation.

Keywords: semantic relation, page count.

1 Introduction

Web mining[2] has introduced a new era in the field of information science. It is the process of extracting knowledge such as patterns, relations from web data. The world wide web is a huge, diverse and dynamic information source. As the web contents grow, it becomes increasingly difficult to manage and classify its information. One of the most important applications of web mining is the web search engine. Semantic relation[1] is required to improve the query search results of the web search engine. It has several applications in fields such as natural language processing, text mining, semantic web [6][7].

Semantic relation between two terms means how much two terms are related even if they are dissimilar in meaning. For example *c programming* is a very frequent word pair in the web though they have no similarity in meaning. In this word pair *c* is a letter and *programming* is a word in English language. Thus meaning wise *c* and *programming* are very different. But, in the web they occur frequently. We intend to find this semantic relation between two words. The page count of *c programming* in Google[1] is 2,090,000[2] whereas the same for *u programming* is 2,810, where page count[4] of a word is the number of pages that contain the query words. This means that *u* and *programming* are not semantically as much related as *c* and *programming*.

[1] http://www.google.com

[2] This Google page count is taken on June 2009.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 267–272, 2009.

This measure is suggested to determine semantic relation between two words using the web as information source. The frequency of the words in each document are measured rather than page count and normalize it over all documents. The other methods available in the literature are mainly based on page count of a word, such as *Normalized Google Distance*[3]. But page count is not sufficient to measure the semantic relation between two words. Page count provides the number of pages in which a word (or words) occurs. It does not indicate the number of times a word has occurred in each of this page. A word may appear many times in a document and once in the other document, but page count will ignore this.

This paper is described as follows. In section 2 we discuss our proposed semantic relation and its properties . Section 3 describes related work and comparison with the proposed measure whereas the last section describes the conclusions and future works.

2 Proposed Measure of Semantic Relation

Various semantic similarity measures are available in the literature. Different methods have been found to be useful in several applications. After extensively investigating a number of methods this measure is proposed to determine the semantic relation between two words which is applicable to any set of documents.

2.1 Formulation

Given two words x and y, we need to find the semantic relation between them on the basis of a corpus(here web pages). The proposed semantic relation named as *Semantic Word Distance*(SWD) is defined as

$$\mathrm{SWD(x,y)} = \begin{cases} -1 & \text{if } f_i(x) = f_i(y) = 0 \; \forall\, i \in \mathrm{M} \\ |\dfrac{\mathrm{wr(x,y)} + \mathrm{wr(y,x)}}{2} - 1| & \text{otherwise} \end{cases} \tag{1}$$

where M is the total number of pages from which we will find the relation between the words e.g. the number of pages indexed by a web search engine and *Word Ratio*(wr) is defined as,

$$\mathrm{wr(x,y)} = \frac{1}{M} \sum_{i=1}^{M} max(I_i(x), I_i(y)) \, \frac{f_i(x) + 1}{f_i(y) + 1}$$

Here $f_i(x)$ denotes the number of times the word x occurs in the ith page and

$$\mathrm{I}_i(x) = \begin{cases} 0 & \text{if } f_i(x) = 0 \\ 1 & \text{otherwise} \end{cases}$$

That is we are neglecting the pages where none of the words occur. We are measuring the semantic relation between two words depending on the score of

SWD. Note that if $f_i(x) = f_i(y) \neq 0 \; \forall \, i \in$ M then SWD(x,y) = 0 i.e. we can say the semantics of x and y are exactly same on a particular information source. SWD(x,y) = 0 means the semantic relation between x and y is the strongest. As the value of SWD approaches zero, the relationship between the word pair becomes stronger. Essentially the above measure finds, for each page, the ratios of the number of occurrences of the word pair and averages it over the corpus. If the ratios are closer to one, the two words are more semantically related except for $f_i(x) = f_i(y) = 0 \; \forall \, i$. This idea will be appropriate for the word pair like *c programming* which exists simultaneously on the web pages but has no similarity in meaning.

2.2 Properties

Now we will discuss some properties of Semantic Word Distance.
• SWD is symmetric. For every pair of word x, y we have SWD(x,y) = SWD(y,x).

• The value of SWD must be greater than or equal to 0 except for the case of $f_i(x) = f_i(y) = 0 \; \forall \, i$ and irrespective of the values of $\max(I_i(x), I_i(y)) \; \forall \, i$. As we know that $\mathrm{a}+\frac{1}{\mathrm{a}} \geq 2$ *where* $\mathrm{a} \in N$, a set of positive integers. Then we can state from equation (1) that SWD(x,y) $=\frac{1}{2}\left[\mathrm{wr(x,y)} + \frac{1}{\mathrm{wr(x,y)}}\right] - 1 \geq 0$

• If SWD(x,y) = 0 then we have

$$\left[\frac{\mathrm{wr(x,y)} \; + \; \mathrm{wr(y,x)}}{2}\right] - 1 = 0$$

$$\text{i,e} \quad \frac{1}{M}\sum_{i=1}^{M}\left\{\left(\frac{f_i(x)+1}{f_i(y)+1} + \frac{f_i(y)+1}{f_i(x)+1}\right)\max(I_i(x), I_i(y))\right\} = 2$$

If $\max(I_i(x), I_i(y)) = 1 \; \forall \, i$ then $f_i(x)+1 = f_i(y)+1 \; \forall \, i$ as we know that $\mathrm{a}+\frac{1}{\mathrm{a}} = 2$ only when a = 1. But it is not necessarily true that $\max(I_i(x), I_i(y)) = 1 \; \forall \, i$. Therefore SWD(x,y) = 0 does not imply $f_i(x) = f_i(y) \; \forall \, i = 1, 2, ..., M$.

•So SWD is not a metric. Also it need not satisfy the triangular inequality i.e. SWD(x,y) + SWD(y,z) ≥ SWD(x,z) $\forall \, x, y, z$ need not be true.

Choose $f_i(x)+1 = X_i$, $f_i(y)+1 = Y_i$, $f_i(z)+1 = Z_i$ and $\max(I_i(x), I_i(y)) = 1 \; \forall \, i$. Therefore the triangular inequality becomes

$$\frac{1}{2M}\sum_{i=1}^{M}\left(\frac{X_i}{Y_i} + \frac{Y_i}{X_i}\right) - 1 + \frac{1}{2M}\sum_{i=1}^{M}\left(\frac{Y_i}{Z_i} + \frac{Z_i}{Y_i}\right) - 1 \; \geq \; \frac{1}{2M}\sum_{i=1}^{M}\left(\frac{X_i}{Z_i} + \frac{Z_i}{X_i}\right) - 1$$

$$i, e \;\; \sum_{i=1}^{M}\left(\frac{X_i}{Y_i} + \frac{Y_i}{X_i} + \frac{Y_i}{Z_i} + \frac{Z_i}{Y_i} - \frac{X_i}{Z_i} - \frac{Z_i}{X_i}\right) - 2M \; \geq \; 0 \qquad (2)$$

Let $X_i \leq Y_i \leq Z_i \; \forall \, i$ and $Y_i = pX_i$ and $Z_i = qX_i$ where p ≥ 3 and q = 2p . This violates the condition of inequality (2). Hence swd is not a metric.

2.3 Similarity Measurement

The lower limit of SWD is -1 which indicates, this is beyond the scope of SWD to determine the relation between the words on a particular corpus. This occurs only when no two words appear in any of the documents of the corpus. In the other cases SWD always provides non negative values. If the SWD value is 0 then it indicates that the semantic relation between the word pair is strongest. It has no upper limit. The relationship between the words decreases as long as the value grows from 0. For simplicity we assume 2 as the upper limit. If two words appear in 1:11 ratio or more than that on an average, then SWD value of two words is greater than 2 and we will just ignore it i.e. two words are dissimilar when their SWD value is greater than 2.

3 Comparison with Other Measures

Semantic similarity measures have proved their worth in web related tasks. A number of methods on semantic relation have been proposed in the last few years. Some of them are subjective i.e. based on some specific corpus e.g. Wordnet[3] [9] or Wikipedia[8]. No measure based on a specific corpus is considered here for comparison. Some of the corpus independent measures are WebJaccard, WebOverlap, WebDice, WebPMI [4], Normalized Google Distance(NGD)[3]. WebJaccard, WebOverlap, WebDice, WebPMI are the modifications of the popular co-occurrence measures Jaccard, Overlap, Dice and PMI respectively. NGD is also a popular measure and is not confined to a particular web search engine. The range of WebJaccard, WebDice and WebOverlap are in between 0 and 1. According to these three formulae two words are similar if the value is 1 and the relationship decreases when the value decreases to 0. But these three measures do not incorporate the total number of pages of the corpus(for example total number of pages indexed by the search engine) which is very important in measuring the semantic relation. NGD normally takes the values between 0 and 1 though the value lies in between 0 and ∞ [3]. For NGD two words are similar if the value is 0 and the relationship decreases as it grows to 1.

All these measures are mainly based on page count. Page count of a word may not provide the full notion of the word e,g Google returns 12,700,000[4] as the page count for *electronic* AND *calculator* whereas 131,000,000 for *electronic* AND *food* although *electronic* is strongly related to *calculator* than *food*. In SWD the frequency of a word in a document is measured rather than page count and normalize it over the corpus. Hence SWD has more proficiency in finding the relation. For example if one word occurs for 45, 30, 22, 16, 53 times in five documents and other word occurs for 2, 3, 4, 1 ,6 times in the same five documents, SWD will produce a result different from zero i,e semantically the words are not that much related. But page count based methods generally produce a result which indicate that the words are close semantically. This is

[3] http://wordnet.princeton.edu

[4] This count is taken on May 2009.

the main difference between the proposed definition and the page count based definition.

The performance of the proposed Semantic Word Distance has been evaluated and compared with the other methods mentioned earlier. In the next two section we will discuss about the data sets used and the methods of evaluation.

3.1 The Data Source

Ten word pairs are selected manually to measure semantic relation between each of them. For this the WebKb data set [5] is used. This data set contains WWW pages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base (WebKb) project of the CMU text learning group. The 8,282 pages were manually classified into seven categories. For each category the data set contains pages from the four universities Cornell (867), Texas (827), Washington (1205) and Wisconsin (1263).

3.2 Evaluation

Here WebKb data set is used as the information source. Basically it will be determined, how much two words are related depending on this corpus. It may again be mentioned here, that two words are more related if the SWD, NGD, WebPMI values are close to zero and the WebJaccard, WebOverlap, WebDice values are close to one.

It is observed from Table 1 that SWD is seemingly better than NGD and WebPMI for almost all the ten word pairs. *Research* and *paper* are very related words as WebKb is a university data set. SWD determines this word pair is very related whereas the other measures fail to show this. They have shown *research* and *paper* are related but not so much as it is shown by SWD. Again the word pairs *university-graduate*, *web-page*, *programming-assignment* and *course-coordinator* have the higher NGD and WebPMI values though they are very

Table 1. Performance of Semantic Relation Between the Selected Word Pairs

Selected Word Pair	Web Jaccard	Web Overlap	Web Dice	Web PMI	NGD	SWD (Proposed)
research - paper	0.250	0.641	0.400	0.651	0.732	0.168
university - graduate	0.271	0.776	0.426	0.698	0.716	0.129
web - page	0.281	0.813	0.439	0.487	0.780	0.001
programming - assignment	0.202	0.625	0.336	0.716	0.747	0.030
course-coordinator	0.041	0.417	0.080	0.678	0.868	0.429
computer - science	0.760	0.975	0.864	0.598	0.379	0.076
home - page	0.676	0.899	0.806	0.632	0.415	0.171
sound - vehicle	0.021	0.064	0.041	1.694	0.748	0.955
book - culture	0.031	0.315	0.061	1.629	0.749	0.818
speaker - player	0.018	0.038	0.037	1.924	0.714	0.908

related (which is determined by SWD with a lower value). This is because the co-existence of these three word pairs in each document of the corpus is very low. The word pairs *sound-vehicle*, *book-culture* and *speaker-player* are weakly related which is shown by SWD like the other measures according to the general understanding of the meaning of the words and their interrelations. It may be noted that WebJaccard, WebDice and WebOverlap are not measuring the total number of pages. Even then the proposed measure is giving better results except for the word pairs *computer-science* and *home-page*. For these two word pairs WebOverlap and WebDice are giving slightly better results than SWD because the co-existence of these word pairs are very high and those measures do not count the total number of pages.

4 Conclusion

A method is suggested to determine semantic relation between two words. This is not based on the traditional page count method. The method counts the number of occurrences of each word in each document and normalizes it over the corpus. By this the entire knowledge of a word pair in a document can be obtained without measuring the co-existence of the words. The co-existence of two words is automatically measured in the formula. The main application of SWD is to measure the semantic relation between words using the web search engine for better query search results. In future we will apply SWD in web page classification and to measure the similarity between web pages.

References

1. Gracia, J.L., Mena, E.: Web Based Measure of Semantic Relatedness. In: Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., Wang, X.S. (eds.) WISE 2008. LNCS, vol. 5175, pp. 136–150. Springer, Heidelberg (2008)
2. Kosala, R., Blockeel, H.: Web Mining Research: A Survey. ACM SIGKDD Explorations 2, 1–15 (2000)
3. Cilibrasi, R.L., Vitanyi, P.M.: The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
4. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring Semantic Similarity Between Words Using Web Search Engines. In: Proceedings of WWW 2007, Banff, Canada (2007)
5. http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/webkbdata.gtar.gz
6. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, 34–43 (May 2001)
7. Catherine, M.C., Frank, S.M.: Which Semantic Web. In: Proceedings of ACM Hypertext, pp. 57–66 (2003)
8. Strube, M., Ponzetto, S.P.: Wikirelate! Computing Semantic Relatedness Using Wikipedia. In: AAAI. AAAI Press, Menlo Park (2006)
9. Pedersen, T., Banerjee, S.P.: Maximizing Semantic Relatedness to Perform Word Sense Disambiguation. Research Report. UMSI 2005 (2005)

Automatic Keyphrase Extraction from Medical Documents

Kamal Sarkar

Computer Science & Engineering Department,
Jadavpur University,
Kolkata – 700 032, India
jukamal2001@yahoo.com

Abstract. Keyphrases provide semantic metadata that summarizes the documents and enable the reader to quickly determine whether the given article is in the reader's fields of interest. This paper presents an automatic keyphrase extraction method based on the naive Bayesian learning that exploits a number of domain-specific features to boost up the keyphrase extraction performance in medical domain. The proposed method has been compared to a popular keyphrase extraction algorithm, called Kea.

Keywords: Domain specific keyphrase extraction, Medical documents, Text mining, Naïve Bayes.

1 Introduction

Medical Literature such as research articles, clinical trial reports, medical news available on the web are the important sources to help clinicians in patient care. The pervasion of huge amount of medical information through WWW has created a growing need for the development of techniques for discovering, accessing, and sharing knowledge from medical literature. The keyphrases help readers rapidly understand, organize, access, and share information of a document. Document keyphrases provide a concise summary of the document content. Medical research articles published in the journals generally come with several author assigned keyphrases. But, medical articles such as medical news, case reports, medical commentaries etc. may not have author assigned keyphrases. Sometimes, the number of author-assigned keyphrases available with the articles is too limited to represent the topical content of the articles. So, an automatic keyphrase extraction process is highly desirable.

A number of previous works has suggested that document keyphrases can be useful in a various applications such as retrieval engines [1], [2], [3], browsing interfaces [4], thesaurus construction [5], and document classification and clustering [6].

Turney [7] treats the problem of keyphrase extraction as supervised learning task. Turney's program is called Extractor. One form of this extractor is called GenEx, which is designed based on a set of parameterized heuristic rules that are fine-tuned using a genetic algorithm.

A keyphrase extraction program called Kea, developed by Frank et al. [8], uses Bayesian learning for keyphrase extraction task. In both Kea and Extractor, the candidate

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 273–278, 2009.

keyphrases are identified by splitting up the input text according to phrase boundaries (numbers, punctuation marks, dashes, and brackets etc.). Kea and Extractor both used supervised machine learning based approaches. Two important features such as distance of the phrase's first appearance into the document and TF*IDF (used in information retrieval setting), are considered during the development of Kea. Frank et al. [8] compares performance of the kea to Turney's work and shows that performance of Kea is comparable to GenEx proposed by Turney. Moreover, Frank et al. [8] claims that training Naïve Bayes learning technique is quicker than training GenEx that employs the special purpose genetic algorithm for training.

Compared to the previous works, our work differs in several ways: (1) we use POS tagger based noun phrase identification method, (2) our approach exploits a number of new domain-specific features and statistical features for keyphrase extraction and (3) a glossary database has been incorporated to discriminate between more-domain-specific and less-domain-specific phrases.

The paper is organized as follows. In section 2 we discuss how to build and use domain knowledge. In section 3, the proposed keyphrase extraction method has been discussed. We present the evaluation and the experimental results in section 4.

2 Domain Knowledge Preparation

One domain specific vocabulary is built up by using MeSH (Medical Subject Headings), which is NLM's (U.S. National Library of Medicine) controlled vocabulary thesaurus. All the MeSH terms are treated as domain specific key phrases. We prepare one table for medical keyphrases (MeSH terms), which is treated as a glossary database. We use this glossary database in our wok to determine domain specificity of a phrase.

In addition to this glossary database, we use another vocabulary of natural language terms to identify novel medical term. To decide whether a term is novel, we used two vocabularies: glossary database (medical vocabulary) and natural language vocabulary because absence of a word in the medical vocabulary is not a sufficient condition to consider it as a novel term. The vocabulary of natural language words has been constructed from a corpus of natural language texts (not related to the medical domain) downloaded from the site under the heading Yahoo news coverage. If a term is not available in these two vocabularies, we treat the term as novel term.

3 Proposed Keyphrase Extraction Method

The proposed keyphrase extraction method consists of three primary components: document preprocessing, noun phrase identification and keyphrase identification. The preprocessing task includes conversion from the pdf format to text format, formatting the document for removing non-textual content.

3.1 Noun Phrase Identification

We treat the noun phrases in the document as the candidate keyphrases [1]. To identify the noun phrases, documents should be tagged. We use GENIA[1] tagger for

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

tagging the input document. GENIA is part-of-speech tagger for biomedical text [9]. We use GENIA tagger 3.0.1. This version can assign tags to the terms in the medical text and separately mark terms which at the beginning and inside a noun phrase. By checking these markers of the noun phrases, we can easily identify noun phrases.

3.2 Keyphrase Extraction Using Naïve Bayes

To extract keyphrase from the medical documents, the Naïve Bayes classifier is trained on a set of medical documents and author assigned keyphrases available with those documents. Based on the features of the noun phrases discussed below, the classifier is trained to classify the candidate noun phrases as the keyphrases (positive examples) or not (negative example). Preparation of training and test instances, training procedure of the Naïve Bayes classifier are also discussed later in the subsequent subsections.

Features. To characterize the noun phrases in the medical documents we have designed and adopted a number of features discussed in full below:

Phrase Frequency and Positional Information. We adopt these two features used in [8] to our Keyphrase extraction task. Phrase frequency means number of times a phrase P occurs in a document. We also consider position of the first occurrence of the phrase in the document as a feature.

Frequency of Component Words in the Phrase. If a phrase frequency is not very high, but frequency of the individual words of the phrase is very high, we consider this phrase as an important one. The log-normalized value of the sum of the frequency of the component words is considered as a feature value.

Presence of the Phrase in the Glossary Database. We apply the following formula to compute the value of this feature.

G = 1 if the phrase is present in the glossary database
=1/n if the phrase partially matches with MeSH terms, where n is the number of partial matches.

If the number of partial matches increases, we assume that the phrase consists of more common words and the entire phrase is assumed to be less domain specific.

Acronyms and Novel Terms. A phrase gets a score based on the number of novel terms and acronyms it contains. In medical articles, authors frequently use acronyms for important complex medical terms, perhaps it helps them memorize the things better. Following two rules are used to detect novel term and acronym:

If the length of the term is greater than a threshold (5 characters) and the term is not found in any of two vocabularies (discussed in section 2), we consider the term as novel term.

If some letters (at least two letters) of a term are capital, we treat the term as an acronym (gene names, medical instruments etc.). For example, fMRI is a term, which is found in our test document, is recognized as an acronym by this rule.

Average Word Length. We also consider the average length of words in the phrase as a feature.

Keyphrase Identification. Training Naïve Bayesian learning algorithm for keyphrase extraction requires document noun phrases to be represented as feature vectors. Author assigned keyphrases are removed from the original document and stored in the different files with document identification number. For each candidate noun phrase in the given document we extract the feature values from the source document using the measures discussed above. If the noun phrase under consideration is found in list of author assigned keyphrases corresponding to the document, we label the phrase as "Positive" example and if it is not found we label the phrase as "negative" example. Thus the feature vector for each noun phrase looks like {<a_1 a_2 a_3 ….. a_n>, <label>} which becomes a training instance (example) for Naïve Bayesian learning algorithm where a_1, a_2 . . .a_n, indicate feature values for a noun phrase. After preparation of training data set, the Naïve Bayesian learning algorithm is trained on the training set to learn to classify candidate noun phrases as one of two categories: "Positive" (class 1) or "Negative" (class 0).

For our experiment, we use Weka (www.cs.waikato.ac.nz/ml/weka) machine learning tools. To build up the model based on Naïve Bayes learning algorithm, we used Weka's Simple CLI utility, which provides a simple command-line interface that allows direct execution of WEKA commands. Since all the features that we consider in this work are real numbers (feature values are continuous), we use Fayyad and Irani's [10] discretization scheme, which is based on the Minimum Description Length principle (MDL). The trained classifier is applied on the test document. During testing, we use –p option. With this option we can generate a probability estimate (posterior probability) for the class of each vector. This is required when the number of noun phrases classified as positive by the classifier is less than the desired number of the keyphrases. For a given document, if the user specifies that K keyphrases are desired, then we select the K vectors that have the highest estimated probability of being in class 1.

4 Evaluation and Experimental Results

There are two usual practices for evaluating the effectiveness of a keyphrase extraction system. One method is to use human judgment, asking human experts to give scores to the keyphrases generated by the system. Another method, less costly, is to measure how well the system-generated keyphrases match the author-assigned keyphrases. We prefer the second approach [7][8] [11] to evaluate the proposed keyphrase extraction system by computing its precision and recall using author-provided keyphrases for medical documents. In this experiment, precision is defined as the proportion of the extracted keyphrases that match the keyphrases assigned by a document's author(s). Recall is defined as the proportion of the keyphrases assigned by a document's author(s) that are extracted by the keyphrase extraction system.

To train and test our keyphrase extraction system, 75 journal articles have been downloaded from a number of online medical journals. The downloaded research articles are basically available as PDF files. All PDF files are converted to text files. Only the text content is considered, non-textual content is removed. Author assigned keywords are separated from the articles and stored in the different file with document identification number. Out of 75 medical research articles, 50 documents and the associated author provided keyphrases are randomly selected for training and the rest 25 documents are used for testing.

Kea [8] is now a publicly available keyphrase extraction system based on Naïve Bayes learning algorithm. Kea uses a limited number of features such as positional information and TF*IDF feature for keyphrase extraction. We download version 5.0 of Kea[2] and install it on our machine. Then it is trained with the same set of medical documents, which are used to train our system. After training Kea, a model is built based on Naïve Bayes. This pre-built model is used to extract keyphrases from the test set consisting of 25 documents.

We calculate the precision and recall for both systems when the number of extracted keyphrases is 5, 10 respectively. We also conduct the statistical significance test on the difference between precisions of the two systems, as well as their recalls, using a paired t test. From table 1, we can find that, in respect to precision and recall, the proposed system performs better than Kea. The results are significant at 95% confidence level.

Table 1. Precision and Recall for the proposed keyphrase extraction system and Kea in medical domain. P-values in the middle column indicates sigificance test on precision difference and P-values in the last column indicates sigificance test on recall difference.

Number of keyphrases	Average Precision ± SD		p-value	Average Recall ± SD		p-value
	Proposed System	Kea		Proposed System	Kea	
5	0.47± 0.20	0.28±0.21	<0.01	0.57±0.24	0.33±0.24	<0.01
10	0.28±.0.11	0.23±0.11	<0.01	0.66±0.24	0.54±0.22	<0.01

To interpret the results shown in table 1 we should mention some important points: some author-provided keyphrases may not occur in the document they are assigned to. According to Turney [7], about only 75% of author-provided keyphrases appear somewhere in the documents. This implies that the highest possible average recall for a system could only be 0.75, even when all the phrases are extracted from the documents. In our experiment, the average number of author-provided keyphrases for all the documents is only 4.33, so the precision would not be high even when the number of extracted keyphrases is large. For example, when the number of extracted keyphrases for each document is 10, the highest possible average precision is around 0.32475 (4.33 * 0.75/10 = 0.32475).

5 Conclusion

This paper discusses a keyphrase extraction method in medical domain. The proposed method uses Naïve Bayes learning algorithm that exploits a number of domain specific features and a number of statistical features for keyphrase extraction from medical documents. The experimental results also suggest that the proposed keyphrase

[2] http://www.nzdl.org/Kea/

extraction method is effective in medical domain and incorporation of domain specific features boosts up the system performance.

References

1. Wu, Y.B., Li, Q.: Document keyphrases as subject metadata: incorporating document key concepts in search results. Journal of Information Retrieval 11(3), 229–249 (2008)
2. Li, Q., Wu, Y.B., Bot, R., Chen, X.: Incorporating document keyphrases in search results. In: Proceedings of the tenth American conference on information systems, New York (2004)
3. Jones, S., Staveley, M.: Phrasier: A system for interactive document retrieval using keyphrases. In: Proceedings of SIGIR 1999, Berkeley, CA (1999)
4. Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., Frank, E.: Improving browsing in digital libraries with keyphrase indexes. Journal of Decision Support Systems 27(1-2), 81–104 (2003)
5. Kosovac, B., Vanier, D.J., Froese, T.M.: Use of keyphrase extraction software for creation of an AEC/FM thesaurus. Journal of Information Technology in Construction 5, 25–36 (2000)
6. Jonse, S., Mahoui, M.: Hierarchical document clustering using automatically extracted keyphrase. In: Proceedings of the third international Asian conference on digital libraries, Seoul, Korea, pp. 113–120 (2000)
7. Turney, P.D.: Learning algorithm for keyphrase extraction. Journal of Information Retrieval 2(4), 303–336 (2000)
8. Frank, E., Paynter, G., Witten, I.H., Gutwin, C., Nevill-Manning, C.: Domain-specific keyphrase extraction. In: Proceeding of the sixteenth international joint conference on artificial intelligence, San Mateo, CA (1999)
9. Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 382–392. Springer, Heidelberg (2005)
10. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of 13th International Joint Conference on Artificial Intelligence, pp. 1022–1027. Morgan Kaufmann, San Francisco (1993)
11. Jones, S., Paynter, G.W.: Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. Journal of American Society of Information Science and Technology 53(8), 653–677 (2000)

An Approach for Preparing Groundtruth Data and Evaluating Visual Saliency Models

Rajarshi Pal, Jayanta Mukherjee, and Pabitra Mitra

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur, India
{rajarshi,jay,pabitra}@cse.iitkgp.ernet.in

Abstract. Evaluation is a key part while proposing a new model. To evaluate models of visual saliency, one needs to compare the model's output with salient locations in an image. This paper proposes an approach to find out the salient locations, i.e., groundtruth for experiments with visual saliency models. It is found that the proposed human hand-eye coordination based technique can be an alternative to costly human pupil-tracking based systems. Moreover, an evaluation metric is also proposed that suits the necessity of the saliency models.

Keywords: Evaluation, Visual saliency model, Groundtruth.

1 Introduction

Visual saliency models try to emulate human vision. Hence, the ideal way to evaluate these models is to estimate how similar (or dissimilar) the results obtained by these models are with the salient regions detected by human. Eye-tracking based systems are used to produce groundtruth data for the evaluation of saliency models.

The eye-tracking technology is costly and few research groups having access to it. The lack of technology of recording salient regions from human observers is evident as many of the works ([1], [2], [3]) have skipped the proper objective evaluation procedure. According to [4] the results obtained from the model was shown to a group of volunteers and they were asked to assess the result as either good, acceptable or failed. This evaluation is quite subjective. In [5], the model's performance is compared against randomly generated locations, not against salient locations reported by humans.

Therefore, technological bottleneck for collecting groundtruth data is a hindrance to the correct evaluation of visual saliency models. As an alternative to human eye-tracking based systems, in this paper, an approach to collect groundtruth data is discussed where volunteers' opinion will be recorded through hand-eye coordination.

Evaluation metric is equally important as collecting groundtruth data. Some of these metrics operate only if the groundtruth data is converted into a human attention map. An attention map (from human fixation) is formed by summing 2D Gaussian patches around each fixation point or using some variation of it.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 279–284, 2009.

Correlation coefficient between the human attention map and model predicted saliency map [6], difference between model predicted saliency map and human attention map [7], and area under ROC curve [8] are used as evaluation metrics. Another class of metrics may be formulated if the groundtruth is specified as a set of points. Average [9] and summation [10] of saliency values at these points are two such metrics. Though all these estimates are shown to work fine as a metric for saliency models, they are not tuned to the specific purpose of using visual saliency. They are very general in nature and are also applied in other fields of computer vision. In this paper, an alternative metric for evaluating saliency models is proposed. The proposed metric is very much tuned to the fact that real-time tasks process only a few salient locations. This metric discourages any model that selects less salient locations.

The outline of the remaining paper is as follows: Section 2 discusses the collection and compilation of groundtruth data. An evaluation strategy is proposed in section 3. Section 4 reports some experimental results to support the proposed approach and finally, section 5 draws the conclusion.

2 Preparation of Groundtruth Data

Let N be the number of image for which groundtruth data, i.e., the salient locations to be recorded. Let V be the number of volunteers recording the salient locations. Not to burden them with much load, each of them is shown n images. To ensure that opinions of a good number of volunteers are taken for each image, the values of these variable are set to satisfy $V \times n >> N$. The ratio $(V \times n)/N$ indicates the strength of evaluation. Typical values of N, V and n, for our experiments, are 50, 62 and 12, respectively.

2.1 A Session with an Individual

At first, the volunteer is shown a very prominent spot at a randomly chosen location within a white background. The spot remains for τ_1 (typically 100 ms in our experiments) time in the screen. The volunteer is asked to click the mouse to the location where the point appeared. The point disappears so quickly that the volunteer has to mark on the white background after its disappearance. The same process is repeated for m_1 (typically 12 in our experiments) times. The Euclidean distance between the actual position of appearance of the point and the position where the volunteer clicked is recorded for the last m_2 (typically 8 in our case) out of those m_1 instances. Mean μ (denoted as mean offset) and standard deviation σ (denoted as standard deviation of offset) of these recorded distances are estimated. The record for first (m_1 - m_2) cases are ignored to give the volunteer some time to get familiar with the procedure. One needs to be very much attentive to have a good hand-eye coordination. This phase serves two purposes. Firstly, it helps the volunteer to be attentive before the recording of salient locations for test images commences. Secondly, an estimate of how far she clicks from the true position is found by taking the mean and standard deviation for the last m_2 instances.

As soon as the first phase completes, the second phase starts where the volunteer's opinion about the salient locations for n images are accumulated. Each image is shown to her for a very short period τ_2 (typically, 100 ms). Human vision has two components. When confronted to a scene, of which it has no prior clue, at first it will be guided to the visually salient locations in the scene. This is called *bottom-up component* of vision. With time the familiarity of the scene increases and the contents of the scene begin to be recognized/interpreted. Then this gradually enhanced understanding of the scene guides our vision. This is called *top-down component* of vision. As the objective here is to collect the groundtruth data to evaluate visual saliency model, i.e., bottom-up attention model, the time period τ_2 for which the image is shown to the volunteer needs to be very small. Smaller τ_2 indicates, lesser influence of top-down component of our vision and better capturing of salient locations. A white screen follows each image. The volunteer is asked to mark at the centers of each location that seems salient to her. Like the first phase, here too she marks on white screen that follows the image. This process is repeated for n images.

2.2 Combining Individual's Opinion to Form Salient Locations

Good hand-eye coordination is crucial for deriving groundtruth data. Therefore, the mean and standard deviation of both μ and σ (obtained in section 2.1) of all V volunteers are calculated. Let μ_μ and σ_μ be respectively mean and standard deviation for the mean distances μ and μ_σ and σ_σ are respectively mean and standard deviation for the standard deviations of distances σ of all the volunteers. The opinions of the volunteers, whose mean offset μ and variance of offsets σ are less than or equal to $(\mu_\mu + \sigma_\mu)$ and $(\mu_\sigma + \sigma_\sigma)$ respectively, are taken into consideration for groundtruth data preparation. Others opinions are discarded as their hand-eye coordination is poorer than the rest. Let S_V and S_v be the set of all volunteers and selected volunteers, respectively. Therefore,

$$S_v = \{i | \, (\mu_i \leq (\mu_\mu + \sigma_\mu)) \land (\sigma_i \leq (\mu_\sigma + \sigma_\sigma)) \land (i \in S_V)\} \tag{1}$$

Now, for each image the following procedure is applied to prepare the groundtruth data using the opinions of only the chosen set of volunteers S_v. Let, P_v be the set of points marked by volunteer v for an image *I*. Set of points P is found by taking together all the volunteers' responses for the image *I*.

$$P = \bigcup_{\forall v \in S_v} P_v \tag{2}$$

Let, N_v be the number of points marked by volunteer v for the image I. A set N is formed comprising of all N_v's.

$$N = \{N_v | \forall v\} \tag{3}$$

Next k-means clustering is performed on the set of points P. Number of clusters k is set to be the mode (i.e., the value that occurs most number of times) of the set N. The mean μ_C and standard deviation σ_C for each cluster C are also computed. As each volunteer marks at the approximate center of each location that seemed salient to her, a circular disk centering at μ_C of radius σ_C covers some portion around the center of each cluster. Collection of these circular disks (one or many for a particular image) constitute the groundtruth data for that image. It may be noted that all these circular regions are non-overlapping, as the clusters form a partitioning in the space. A binary image B (of same size as I) is constructed where all the pixels belonging to these disks (salient locations) are set to 1 and the remaining pixels are set to 0.

3 Evaluation Metric

It is checked whether there exist other locations which are more salient than the groundtruth indicated in B (obtained in previous section). Let S be the saliency map obtained for image I. Moreover, let m_i represent the maximum saliency value for each salient region R_i (in the form of a circle as discussed in previous section) in B.

$$m_i = max(S(R_i)) \tag{4}$$

where $S(R_i)$ represent the collection of values in the saliency map S corresponding to the region R_i.

It is checked whether there are other pixels that have saliency value greater than m_i and do not belong to any other salient region R_j in B. Let Γ is the set of such pixels. If no such pixel exists, then Γ becomes a null set.

$$\Gamma = \{x| \, (S(x) > m_i) \wedge (x \notin R_j) \wedge (\forall R_j \in B)\} \tag{5}$$

If Γ is not the null set, the error measure E_i for R_i is the sum of normalized distances for all pixels in Γ from R_i. Distances are normalized with respect to $\sqrt{(L^2 + M^2)}$, where image I is of size L-by-M.

$$E_i = \sum_{y \in \Gamma} mindist(y, R_i)/\sqrt{(L^2 + M^2)} \tag{6}$$

where mindist(a,A) is the minimum of all the Euclidean distances of a pixel a from a group of pixels A.

Error estimate E for the saliency model is the summation of all error estimates E_i for all values of i.

$$E = \sum_i E_i \tag{7}$$

In a nutshell, the summation of minimum distance of each pixel in the locations that are more salient (according to the computational model) from salient regions indicated in B is the proposed evaluation metric.

4 Experimental Validation

In our experiments, 50 images are chosen from a larger set of databases taken from iLab image database ([1], [11]), UCID([12]), Zurich natural image database ([13] and the Internet. A well-known saliency model [1] is used to validate the procedures described in above two sections. Groundtruth data is prepared using the procedure described in section 2. This data along with the input images is given in `http://www.facweb.iitkgp.ernet.in/~jay/VS/Groundtruth.html`. The groundtruth is represented in a binary image with regions filled with 1 belong to salient locations. Let for an image I, the set of salient locations obtained by the proposed procedure is denoted by R.

On the other hand, for each of the input images, a set of circular regions are chosen with randomly selected center positions. Let, for the image I the set of randomly chosen circular regions is denoted by T. For each image, the number of randomly selected circular regions, i.e., cardinality of T, is kept equal to the number of salient regions obtained in the groundtruth (cardinality of R, i.e., which is equal to k). Moreover, radius of each of the circular salient locations in R are maintained in T. The purpose, here, is to show that scores at locations obtained according to the proposed procedure (R) is higher than the scores at randomly selected locations (T). As in [9], average of saliency values at these locations are used as evaluation metric. Scatter diagram in figure 1 shows that average saliency value at the volunteer specified locations (average over 50 images is 140) is higher than average saliency value at randomly selected locations (average over 50 images is 72.16). It strengthens the fact that locations obtained by proposed procedure are salient.

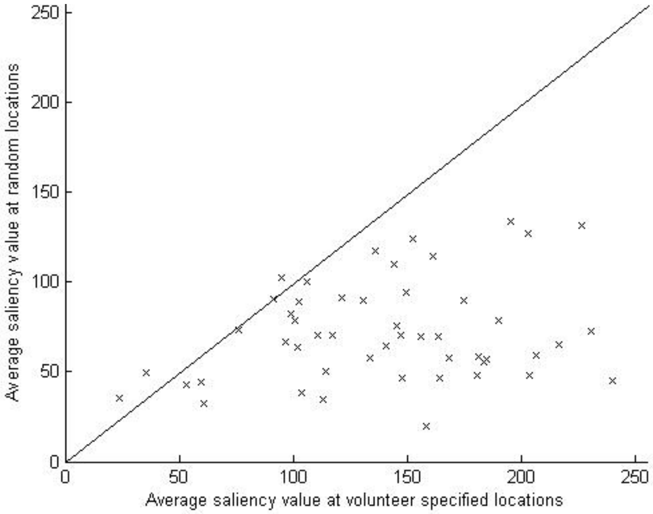


Fig. 1. Scatter Diagram of average saliency values at locations obtained by proposed procedure versus that of randomly chosen locations

It is also experimentally observed, as expected, that the average saliency value (measure of similarity) and the proposed metric (measure of dissimilarity) are negatively correlated. Their correlation coefficient is measured to be -0.58. This statistics justifies the proposed metric as a error measure.

5 Conclusion

In this paper, a hand-eye coordination based procedure is proposed to compile groundtruth data. Experimental results show that this can be an alternative to using eye-tracking systems. An evaluation metric is also proposed which is an error measure by definition. The fact that it negatively correlates with another evaluation metric (average saliency value) strengthens this proposed metric.

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
2. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision 45(2), 83–105 (2001)
3. Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artificial Intelligence 146, 77–123 (2003)
4. Yu, Z., Wong, H.S.: A rule based technique for extraction of visual attention regions based on real-time clustering. IEEE Transactions on Multimedia 9(4), 766–784 (2007)
5. Minut, S., Mahadevan, S.: A reinforcement learning model of selective visual attention. In: Proceedings of 15th International Conference on Autonomous Agents, pp. 457–464 (2001)
6. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5), 802–817 (2006)
7. Bruce, N.D.B.: Features that draw visual attention: an information theoretic perspective. Neurocomputing 65-66, 125–133 (2005)
8. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: Proceedings of IEEE 11th International Conference on Computer Vision, pp. 1–6 (2007)
9. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Research 42, 107–123 (2002)
10. Meur, O.L., Thoreau, D., Callet, P.L., Barba, D.: A spatio-temporal model of the selective human visual attention. In: Proceedings of IEEE International Conference on Image Processing, pp. III–1188–1191 (2005)
11. Itti, L., Koch, C.: Feature combination strategies for saliency based visual attention systems. Journal of Electronic Imaging 10(1), 161–169 (2001)
12. Schaefer, G., Stich, M.: Ucid - an uncompressed color image database. In: SPIE Storage and Retrieval Methods and Applications for Multimedia, vol. 5307, pp. 472–480 (2004)
13. Frey, H.P., Konig, P., Einhauser, W.: The role of first- and second- order stimulus features for human overt attention. Perception and Psychophysics 69, 153–161 (2007)

Evaluation of Segmentation Techniques Using Region Size and Boundary Information

D.P. Dogra[1], A.K. Majumdar[1], and S. Sural[2]

[1] Department of Computer Sc. & Engineering, Indian Institute of Technology,
Kharagpur, 721302, India
{dpdogra,akmj}@cse.iitgp.ernet.in

[2] School of Information Technology, Indian Institute of Technology,
Kharagpur, 721302, India
shamik@sit.iitgp.ernet.in

Abstract. Image segmentation quality evaluation is a key element when comparing segmentation algorithms. In computer vision, unsupervised segmentation algorithms, although of great interest, often suffer from lack of a well-defined measure to evaluate. This paper presents a novel idea for evaluating such algorithms. A measure is proposed to evaluate four well referred segmentation algorithms. The metric proposed in this work is composed of both size and boundary of segments. When compared with some of the existing techniques, it is found that the proposed scheme can approximate the segmentation error in a better way.

Keywords: Segmentation Evaluation, Area Matching Index, Boundary Matching Index, Combined Matching Index.

1 Introduction

Image segmentation is an important ingredient that is used in many image analysis and computer vision applications. In spite of focusing extensively in designing segmentation techniques, relatively less attention has been given in evaluating those algorithms. Thus, development of a suitable evaluation technique that can be used to compare efficacy of segmentation algorithms.

The existing evaluation methods can broadly be divided into two categories: analytical and empirical methods [17]. A review of analytical based measures can be found in [11] [15]. Empirical methods indirectly judge the segmentation algorithms by comparing them with gold standard segmentation of the images under consideration. Though, most of these evaluation methods are subjective or tied to specific applications [16], a number of unsupervised objective evaluation measures [1] [8] [9] [10] [12] [14] have received considerable attention. Features like, region boundary, region size, region uniformity, region contrast etc. are used in unsupervised objective evaluation. Furthermore, these techniques can be classified into two groups, i.e. region boundary and size based techniques. The former evaluates segmentation in terms of accuracy of region's shape, while the latter approach assesses the segmentation quality in terms of overlap size.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 285–290, 2009.

The evaluation methodology developed by Martin et al. [9] is a classic example of size based technique. Recently, a new size based measure using Normalized Probabilistic Rand (NPR) index is proposed [14]. Alternatively, the boundary based approach proposed by the authors of [10] has been used in many applications. Also, precision and recall based method proposed by the authors of [5] [6] are popular. A drawback of precision / recall based scheme is that, given a correspondence between two segmentations, it is possible to change the locations of the unmatched boundary almost arbitrarily and retain the same precision and recall score.

Thus, a measure that considers both size and boundary of regions can be of great use in defining segmentation accuracy. We propose a technique that combines a recently proposed region size based measure [1] with a new boundary based measure. Rest of the paper is organized as follows. A detail description of the proposed measure is described in section 2. Evaluation strategy and results are reported in Section 3 while we conclude in Section 4 with some discussions.

2 Proposed Method of Evaluation

LCE-GCE method introduced by Martin et al. [9], NPR method proposed by Unnikrishnan et al. [14] and Segmentation Covering (SC) [1] are size based measures whereas methods proposed by the authors of [5] [6] [10] are boundary based. We explain the limitations of these methods with the example shown in Fig.1. Let, $G = \{G_1, G_2, ...G_m\}$ and $S = \{S_1, S_2, ...S_n\}$ be two sets of segmentations where G_i and S_j are i^{th} and j^{th} segments of gold and machine segmentations respectively. Fig. 1(d) shows that G_2 has maximum intersection with S_2 ($\forall\ j,\ S_j \in S$) and it is a subset of S_2. These measures do not consider the common boundary information. NPR based method too does not incur

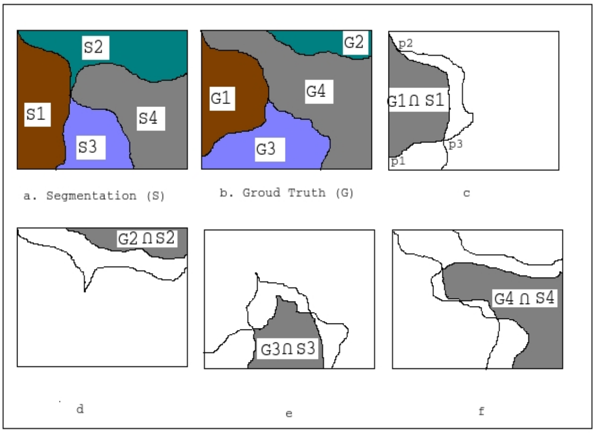


Fig. 1. (a) Computer Generated Segmentation Map of an Image. (b) It's Ground Truth Map. (c-f) Individual Intersection Map.

additional penalty despite a high degree of dissimilarity between the boundaries. Thus, a measure that takes into account both size and boundary based errors, will be of great help in evaluating segmentation algorithms. To define the proposed measure, two separate indices are used, i.e. Area Matching Index (AMI) [1] and a new Boundary Matching Index (BMI). Out of the three size based measures discussed earlier, as SC [1] based approach is most recent and it performs reasonably well for natural images, we have adopted it in our work.

$$AMI_{overall} = C(G \rightarrow S) = \frac{1}{N} \sum_{S_j \in S} |S_j|. \max_{G_i \in G} \{O(S_j, G_i)\} \tag{1}$$

where $S = \{S_1, S_2, ..., S_n\}$ and $G = \{G_1, G_2, ..., G_m\}$ and N is total number of pixels of the image and $O(S_j, G_i)$ is called the overlap between two regions S_j and G_i. It is estimated by:

$$O(S_j, G_i) = \frac{|S_j \cap G_i|}{|S_j \cup G_i|} \tag{2}$$

A new boundary based measure that considers neighborhood information is proposed here. We call it as Boundary Matching Index (BMI). It is computed is as follows. Let, the closed arc P $=\{p_1, p_2, p_3\}$ in Fig. 1(c) describes the boundary of $G_1 \cap S_1$. Two situations may arise: (i) A point on the arc P, say p, is a common boundary point of both G_1 and S_1 (e.g. $x \in \{p_1, p_2\}$) (ii) The point on P is either on boundary of G_1 or S_1 (e.g. $x \in \{p_1, p_3\}$ or $x \in \{p_3, p_2\}$). A score (BMI_{ij}) based on boundary points (P) of the region $G_1 \cap S_1$ is estimated by:

$$BMI_{ij} = \frac{\sum_{p=1}^{k} C_p}{k} \qquad \text{for } i = 1 \text{ and } j = 1$$

where k is the total number of points on P and C_p is defined by:

$$C_p = \begin{cases} 1 \text{ , If } |p - Boundary(G_i)| < d \text{ and } |p - Boundary(S_j)| < d \\ \frac{N_p}{8} \text{ , Otherwise} \end{cases}$$

where N_p is the number of neighboring pixels of p that belong to $G_i \cap S_j$ using 8-connectivity rule and d is a strictness measure normally set to 3. A value of $d = 1$ will over-penalize boundary points those are close but not coincident. Taking the average BMI_{ij} of all machine generated segments, final score of the index as: $BMI_{overall} = \frac{BMI_{ij}}{n}$ for $i \in m$ and $j \in n$ is estimated. Finally, a Combined Matching Index (CMI) is produced by:

$$CMI = \frac{W_{AMI} X AMI_{overall} + W_{BMI} X BMI_{overall}}{2} \tag{3}$$

where W_{AMI} and W_{BMI} are weights for area and boundary index respectively and satisfy the condition $W_{AMI} + W_{BMI} = 1$.

3 Evaluation Strategy and Results

We choose four well referred segmentation techniques to evaluate the performance of the proposed measure namely, Mean Shift (MS) [2] [3], Normalized Cut (NC) [13], Efficient Graph Based Method (GB) [4] and Color Based Salient Region Segmentation (SR) [7]. Berkeley Segmentation Dataset (BSDS) [9] that contains 300 natural images including multiple manual segmentations, is used for evaluation of the proposed measure. Fig. 2 depicts the segmentations of an image taken from BSDS with five ground truth results. To represent the result shown in Fig. 2 in a better way, we have plotted the values of five measures including the proposed one using the graph shown in Fig. 3. It is understood from the graph that the proposed measure (CMI) rates the SR based segmentation algorithm with a significantly low matching index when compared to other measures. CMI performs equally well as compared to NPR for MS, NC and GB algorithms while a better approximation of segmentation error is evident particularly for SR algorithm.

We have applied four segmentation algorithms on the images of the BSDS dataset and recorded the values of five evaluation measures and found that the proposed measure does a better approximation of segmentation error. From Fig. 4, it is evident that the segmentation result produced by NC, GB and MS are almost similar for most of the images and all measures including CMI agree in most of cases. Segmentation produced by SR algorithm for the second image is not satisfactory and CMI detects it with a high degree of confidence. Though all five measures evaluate the segmentation of the rock image (Fig. 4, 3rd row) in a consistent manner, CMI performs good approximation in case of bird image (Fig. 4, 1st) and old man image (Fig. 4, 2nd).

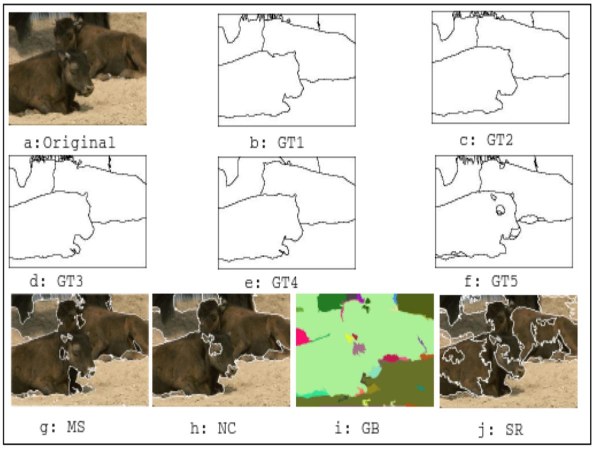


Fig. 2. Output of MS, NC, GB, SR Schemes with Five Ground Truth Segmentations

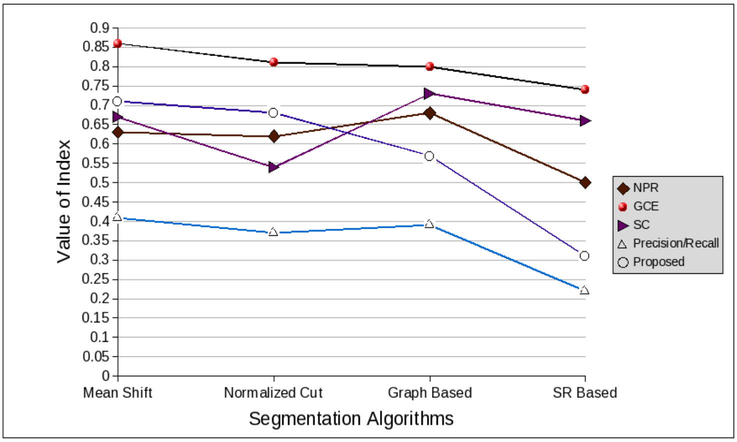


Fig. 3. Graphical Representation of Five Evaluation Measures for the Image of Fig. 2

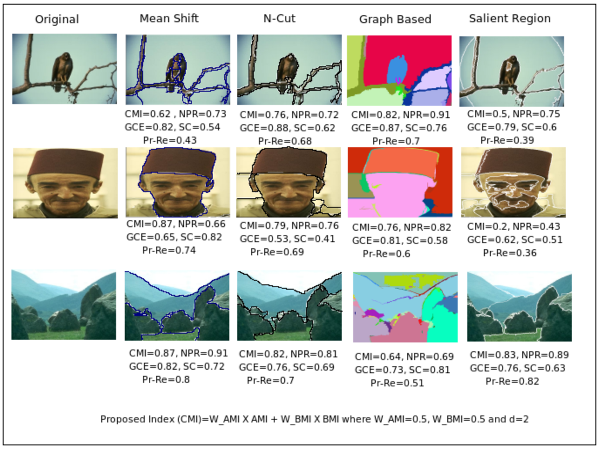


Fig. 4. Results using MS, NC, GB, SR Schemes and Recorded Evaluation Measures Applied on BSDS dataset

4 Conclusion and Future Work

Evaluation of segmentation schemes is an interesting problem in computer vision and image processing area. In this work, a new objective evaluation measure is proposed to compare segmentation algorithms. While most recent segmentation quality evaluation methods only deal with either of area or boundary, the proposed one uses both region size and boundary of the segmentation partition in a clean and comprehensive way. As a low level measure, it produces valid results under varying conditions where reference segmentation is available. This technique could also be used as a building block in more complex and application specific evaluation schemes. In conclusion, though efforts have been put on and encouraging results have been obtained in the last few years, new ideas and procedures for the evaluation methodology and their practical implementation are still required.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: From Contours to Regions: An Empirical Evaluation. In: CVPR (in press, 2009)
2. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Trans. on PAMI 24(5), 603–619 (2002)
3. Comaniciu, D., Meer, P.: Mean Shift Image Segmentation Software, http://www.caip.rutgers.edu/riul/research/code/EDISON/index.html
4. Felzenszwalb, D.: Efficient Graph-based Image Segmentation. IJCV 59(2), 167–181 (2004)
5. Freixenet, J., Muñoz, X., Raba, D., Martí, J., Cufí, X.: Yet Another Survey on Image Segmentation: Region and Boundary Information Integration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 408–422. Springer, Heidelberg (2002)
6. Huang, Q., Dom, B.: Quantitative Methods of Evaluating Image Segmentation. In: ICIP, pp. 53–56 (1995)
7. Kuan, Y., Kuo, C., Yang, N.: Color-Based Image Salient Region Segmentation Using Novel Region Merging Strategy. IEEE Trans. on MM 10(5), 832–845 (2008)
8. Martin, D.: An Empirical Approach to Grouping and Segmentation. PhD Dissertation, Univ. of California, Berkeley (2002)
9. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. ICCV 2, 416–425 (2001)
10. Martin, D., Fowlkes, C., Malik, J.: Learning to Detect Natural Image Boundaries Using Local Brightness, Color and Texture Cues. IEEE Trans. on PAMI 26(5), 530–549 (2004)
11. Pal, N.R., Pal, S.K.: A Review on Image Segmentation Techniques. Jour. of PR 26(9), 1277–1294 (1993)
12. Rand, W.: Objective Criteria for the Evaluation of Clustering Methods. Journal of ASA 66, 846–850 (1971)
13. Shi, J., Malik, J.: Normalized Cuts and Image Segmentation. IEEE Trans. on PAMI 22(8), 888–905 (2000)
14. Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward Objective Evaluation of Image Segmentation Algorithms. IEEE Trans. on PAMI 29(6), 929–944 (2007)
15. Weszka, J.S., Rosenfeld, A.: A.: Threshold Evaluation Techniques. IEEE Trans. on SMC 8(3), 622–629 (1978)
16. Zhang, H., Frittb, J.E., Goldman, S.A.: Image Segmentation Evaluation: A Survey of Unsupervised Methods. Jour. of CVIU 110(2), 260–280 (2008)
17. Zhang, Y.J.: A Survey on Evaluation Methods for Image Segmentation. Jour. of PR 29(8), 1335–1346 (1996)

Unsupervised Color Image Segmentation Using Compound Markov Random Field Model

Sucheta Panda[1] and P.K. Nanda[2]

[1] IPCV Lab., Department of Electrical Engineering, National Institute of Technology, Rourkela-769008, Orissa, India
[2] Department of Electronics and Telecommunication Engineering, C.V Raman College of Engineering, Bhubaneswar.-752054, Orissa, India
pandasucheta06@gmail.com, pknanda_d13@rediffmail.com

Abstract. In this paper, we propose an unsupervised color image segmentation scheme using homotopy continuation method and Compound Markov Random Field (CMRF) model. The proposed scheme is recursive in nature where model parameter estimation and the image label estimation are alternated. Ohta (I_1, I_2, I_3) model is used as the color model for image segmentation and we propose a compound MRF model taking care of intra-color and inter-color plane interactions. The CMRF model parameters are estimated using Maximum Conditional Pseudo Likelihood (MCPL) criterion and the MCPL estimates are obtained using homotopy continuation method. The image label estimation is formulated using Maximum *a Posteriori* criterion and the MAP estimates are obtained using hybrid algorithm. In the context of misclassification error, the proposed unsupervised scheme with CMRF model exhibited improved segmentation accuracy as compared to MRF model and Kato's method.

Keywords: Color Image,Color Model,Segmentation,Simulated Annealing and MRF model.

1 Introduction

Image segmentation is a basic early vision problem which serves as precursor to many high level vision problems. Color image segmentation provides more information while solving high level vision problems such as, object recognition, shape analysis etc. Therefore, the problem of color image segmentation has been addressed more vigorously for more than one decade. Different color models such as $RGB, HSV, YIQ, Ohta(I_1, I_2, I_3), CIE(XYZ, Luv, Lab)$ are used to represent different colors [1]. From the reported study, HSV and I_1, I_2, I_3 have been extensively used for color image segmentation. Ohta color space is a very good approximation of the Karhunen-Loeve transformation of the RGB, and is very suitable for many image processing applications [2].

Stochastic models, particularly MRF models, have been successfully used as the image model for image restoration and segmentation [3],[4]. MRF model has

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 291–296, 2009.

also been successfully used as the image model while addressing the problem of color image segmentation both in supervised and unsupervised framework. The model parameters can be estimated in both supervised and unsupervised framework [5]. Homotopy continuation methods are globally convergent methods that have been used to trace the zeros of a function and hence determines the solution of functions [6],[7].

In this work, a Compound MRF model based color image segmentation scheme is proposed in unsupervised framework. We have used Ohta(I_1, I_2, I_3) color space to model the color images. In the proposed scheme, the compound MRF model parameters and the image labels are estimated concurrently. Since the image label estimates and the estimates of model parameters are dependent on each other, obtaining global estimates of label as well as model parameters is very hard. Hence, we have proposed a recursive scheme for estimation of image labels and model parameters. The recursive scheme yields partial optimal solutions as opposed to optimal solutions. The MRF model parameter estimation problem is formulated in Maximum Conditional Pseudo Likelihood (MCPL) framework and the MCPL estimates are obtained using homotopy continuation bases algorithm. The image label estimation problem is formulated in Maximum a Posteriori (MAP) framework and the MAP estimates are obtained using the proposed hybrid algorithm. The proposed unsupervised algorithm, has been successfully tested on different images, however, for the sake of illustration we have presented two results and a comparison is made with the Kato *et al* [5] method.

2 Compound MRF Model

MRF modeling for color is more complex than the gray image modeling in the sense that it has to take care of the different color components of a color space. In this work, we have employed (Ohta(I_1, I_2, I_3)) color model. MRF model is used to model images in both RGB and Ohta color space. The proposed compound MRF model is based on the following notion: (i)Intra-color-plane (I_1 or I_2 or I_3) enities of each color plane are modeled as MRF model (ii)Inter-color-plane interactions among color planes for e.g.(I_1 and I_2 and I_3), are also modeled as MRF. This process of interaction(intra-plane and inter-plane) is shown in Fig. 1(a) and Fig. 1(b).

We assume all images to be defined on discrete rectangular lattice M1 x M2. Let Z denote the label process corresponding to the segmented image and z is a realization of the label process i.e the segmented image. It is known that if Z is assumed MRF, then the prior probability distribution $P(Z = z)$ is Gibb's distributed that can be expressed as $P(Z = z|\theta) = \frac{1}{Z'}e^{-U(z,\theta)}$, where $Z' = \sum_z e^{-U(z,\theta)}$ is the partition function, θ denotes the clique parameter vector, the exponential term $U(z,\theta)$ is called the energy function and is of the form $U(z,\theta) = \Sigma_{c \in C} V_c(z,\theta)$, with $V_c(z,\theta)$ being referred as the clique potential function. Since the inter-plane process is viewed to be MRF, we know that $P(Z^{I_2}_{i,j} = z^{I_2}_{i,j}|Z^{I_1}_{k,l} = z^{I_1}_{k,l}, (k,l) \neq (i,j), \forall(k,l) \in I_1,) = P(Z^{I_2}_{i,j} = z^{I_2}_{i,j}|Z^{I_1}_{k,l} = z^{I_1}_{k,l}, (k,l) \neq (i,j), (k,l) \in \eta^{I_1}_{i,j})$, Where I_1 and I_2 denotes I_1 and I_2 color planes respectively. In other words

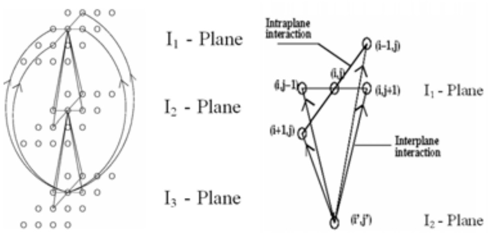


Fig. 1. (a) I_1, I_2, I_3 Plane Interaction (b) Interaction of one pixel of I_1-plane with I_2-plane

a pixel in one plane (say for e.g I_1-plane) is assumed to have interaction with pixels of I_2 and I_3 planes, the interaction process of each color plane is shown in Fig 1(a). Thus the energy function, $U(z, \theta) = \sum_{c \in C_{in}} V_c(z^{(1)}, z^{(2)}, z^{(3)}) + \sum_{c \in C_{ir}} V_c(n^{(1)}, n^{(2)}, n^{(3)})$, where $V_c(z^{(1)}, z^{(2)}, z^{(3)})$ correspond to the intra-color-plane pixels and $V_c(n^{(1)}, n^{(2)}, n^{(3)})$ correspond to the inter-color-plane pixels.

3 Unsupervised Image Segmentation

In unsupervised segmentation scheme, the labels are the MAP estimates assuming the estimates of the associated model parameter $\hat{\theta}$ are available. In this scheme, the MAP estimates of the labels and the estimates of the model parameters need to be carried out concurrently. Thus, an estimation strategy needs to be developed, which using the observed image, X,will yield an optimal pair (Z^{opt}, θ^{opt}). The following joint optimality criterion is considered,

$$(Z^{opt}, \theta^{opt}) = arg \quad \max_{z,\theta} P(Z = z|X = x, \theta) \tag{1}$$

The estimated pair (Z^{opt}, θ^{opt}) satisfying (1) are the global optima of $P(Z = z/X = x, \theta)$ with respect to Z and θ. Since both the entities Z and θ are unknown and interdependent, the problem is a very hard problem. Therefore, it is necessary to opt for strategies for suboptimal solution. In (1), z, θ could be viewed as a set of parameter of the given function $P(Z = z/X = x, \theta)$. For such kind of problems in deterministic framework, Wendell and Horter[8] have proposed an alternate approach that would yield suboptimal solutions instead of optimal solution. Their approach is based on splitting the variables followed by recursively estimating the parameters. They have proved that, the final estimate in this process is called as the partial optimal solution. In our case, in stochastic framework, we in the same spirit, venture to split the original problem into estimation of labels(z) and parameters θ to obtain the partial optimal solutions The splitting of the variables can be expressed as follows

$$(Z^*) = arg \quad \max_{z} P(Z = z|X = x, \theta^*) \tag{2}$$

$$(\theta^*) = arg \quad \max_{\theta} P(Z = z^*|X = x, \theta) \tag{3}$$

These partial optimal solutions Z^* and θ^* are not global maxima, rather they are almost always local optimal solutions [8]. But with $\theta = \theta^*$, the estimate z^* is global optimal satisfying equation (2) and analogously for $z = z^*$, θ^* is global optimal satisfying equation(3). Since neither θ^* nor z^* is known, evaluating Z^* and θ^* is also hard and hence, a recursive scheme is adopted where the model parameter estimation and segmentation is alternated. Let at the k^{th} iteration $\theta^k = [\alpha^k, \beta^k]^T$ be the estimate of model parameters and z^k be the estimate of the labels of the observed image. We adopt the following recursion

$$(Z^{k+1}) = arg \quad \max_{z} P(Z = z | X = x, \theta^k) \tag{4}$$

$$(\theta^{k+1}) = arg \quad \max_{\theta} P(Z = z^{k+1} | X = x, \theta) \tag{5}$$

The first problem of equation (4) is solved using Bayesian approach [3]. The optimal value of θ^k is obtained by the proposed Homotopy Continuation method. The MAP estimates are obtained by the proposed hybrid algorithm. One estimate of z^k and θ^k constitute *one combined iteration*. This recursion is continued for finite number of steps to obtain z^* and θ^*. Thus, the partial optimal solutions are obtained. The model parameter θ has been estimated using Homotopy Continuation method proposed by Panda *et al* [9]. In the following we briefly explain the estimation of image label.

4 Image Label Estimation

The segmentation problem is cast as the pixel labelling problem. Each pixel can assume a label from the set of labels $\{0 - L\}$. In a given image of size $L = M1$ x $M2$, let $Z_{i,j}$ denote the random variable for $(i, j)^{th}$ pixel, $\forall (i, j) \in L = M1$ x $M2$. Z denotes the label process and z denotes a realization of the process. The label estimates $\hat{z}$ is obtained by maximizing the posterior probability $P(Z = z | X = x, \theta)$. Thus, the optimality criterion can be expressed as follows,

$$(\hat{z}) = \underset{z}{arg\, max}\, P(Z = z | X = x, \hat{\theta}) \tag{6}$$

where, $\hat{\theta}$ denotes the associated parameter vector of the Compound MRF model Z.

After carrying out simplification, the problem reduces to the following minimization problem,

$$\hat{z} = \underset{z}{arg\, min} \{ \sum_{i=1}^{3} \frac{(x^{(i)} - z^{(i)})^2}{2\sigma^2} + \\ \sum_{c \in C_{in}} V_c(z^{(1)}, z^{(2)}, z^{(3)}) + \sum_{c \in C_{ir}} V_c(n^{(1)}, n^{(2)}, n^{(3)}) \} \tag{7}$$

where $V_c(z^{(1)}, z^{(2)}, z^{(3)})$ and $V_c(n^{(1)}, n^{(2)}, n^{(3)})$ corresponds to the clique potential function of intra-color-plane pixels and inter-color-plane pixels respectively.

In this work, we have considered Weak Membrane MRF model of Geman and Geman with line field [3]. The MAP estimates of the image labels $\hat{z}$ are obtained using the Hybrid algorithm. The algorithm is a combination of Simulated Annealing (SA) and Iterated Conditional Mode (ICM) algorithm. SA is run till the energy reaches a threshold and thereafter ICM is run to obtain the MAP estimates of labels.

5 Simulation

One outdoor images and one indoor image are considered in simulation. The first original image, an indoor image with three objects on a table, is shown in Fig.2(a). In order to compute the percentage of misclassification error, the Ground Truth image, as shown in Fig.2(b), is constructed manually. The estimated MRF model parameters are, $\alpha = 0.1025$, $\beta = 2.28$ and $\sigma = 0.5$. However σ is chosen by trial and error and is fixed at 0.5. The results obtained by basic MRF model is shown in Fig.2(c), where it is observed that some portions of the stapler are not prominent. CMRF model based approach could prominently preserve these portions. In Fig.2(d) it is observed that the edges and the shapes could be recovered. Observing the result obtained by Kato's method, as shown in Fig.2(e), the edges are dithered and some portion of the stapler has been misclassified. However, the percentage of misclassification error in case of MRF model is 6.5% which is close to that of CMRF model i.e. 2.91%. The observations are different in case of real outdoor images. Fig.3(a) shows an image with

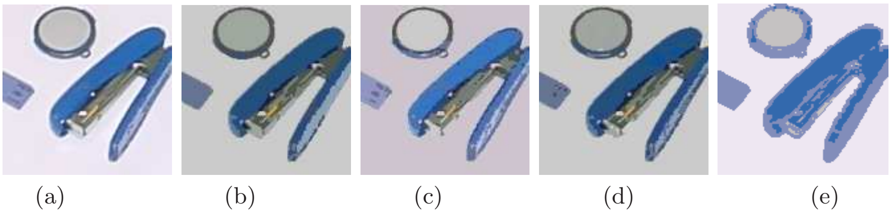


Fig. 2. (a)*Stapler(Indoor) image(128 x 128)* (b)Ground Truth (c)MRF optimized using Hybrid (d)CMRF optimized using Hybrid (e)MRF_KATO

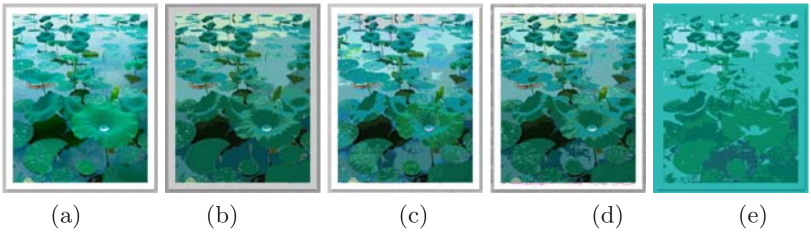


Fig. 3. (a)*Leaves-Water image(512 x 627)* (b)Ground Truth (c)MRF optimized using Hybrid (d)CMRF optimized using Hybrid (e)MRF_KATO

leaves and water body and the corresponding ground truth is shown in Fig.3(b). The misclassification error for MRF model is 9.2% which is higher than that of CMRF model which is 3.5%. The estimated MRF model parameters are $\alpha = 0.003$, $\beta = 8.98$ and $\sigma = 0.35$. Thus, the scheme with CMRF model yielded satisfactory results for indoor as well as outdoor scenes.

6 Conclusion

An unsupervised color image segmentation scheme is proposed with homotopy continuation method and CMRF model. Because of the globally convergent property of the homotopy continuation method, the algorithm can start from a arbitrary set of model parameters and converges to the partial optimal sets. In order to speed up the MAP estimation process, hybrid algorithm is used. The only limitation of the scheme is to choose a proper value of σ for the degradation process. However, estimation of σ together with the model parameters is currently focused.

References

1. Cheng, H.D., Jiang, X.H., Sun, Y., Wang, J.: Color Image Segmentation: Advances and prospects. Pattern Recog. 34, 2259–2281 (2001)
2. Ohta, Y.I., Kanade, T., Sakai, T.: Color information for region segmentation. Comp. Grap. Image. Process. 62, 222–241 (1980)
3. Geman, S., Geman, D.: Stochastic relaxation,Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on PAMI 6, 721–741 (1984)
4. Zhang, J., Modestino, J.W.: A Model-Fitting Approach to Cluster Validation with Application to Stochastic Model-based Image Segmentation. IEEE Trans. on PAMI 12(10), 1009–1017 (1990)
5. Kato, Z., Pong, T.C., Lee, J.C.M.: Color image segmentation and parameter estimation in a markovian framework. Pattern Recognition Letters 22, 309–321 (2001)
6. Chow, N., Mallet-Paret, J., Yorke, J.A.: Finding zeros of maps: homotopy methods that are constructive with probability one. Math. Computation 32(143), 887–899 (1978)
7. Stonick, V.L., Alexander, S.T.: A Relationship between recursive least square update and homotopy continuation methods. IEEE Trans. Signal Processing 39(2), 530–532 (1991)
8. Wendell, R.E., Horter Jr., A.P.: Minimization of a non-separable objective function subject to disjoint constraints. Operations Research 24(4), 643–657 (1976)
9. Sucheta, P., Nanda, P.K.: Color Image using Constrained Compound Markov Random Field Model and Homotopy Continuation Method. In: Proc. of the first International Conference on Distributed Frameworks and Applications, Universiti Sains Malaysia, Penang, Malaysia, pp. 151–158 (2008)

A New Statistical Restoration Method for Spatial Domain Images

Arijit Sur[1,⋆], Piyush Goel[2], and Jayanta Mukherjee[2]

[1] Department of Computer Science and Engineering,
Indian Institute of Technology, Guwahati-781039, India
[2] Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur-721302, India
arijit@iitg.ernet.in, {piyush,jay}@cse.iitkgp.ernet.in

Abstract. In this paper[1], a new algorithm is proposed for restoring first order statistics of the cover image after steganographic embedding in spatial domain on gray scale images. The main motivation of this paper is to prevent first order statistics based steganalytic attack in spatial domain. We provide experimental results to show that the proposed scheme gives better performance than existing restoration methods especially for non-Gaussian cover distribution. A few state of art steganalytic attacks based on first order image statistics are considered and it is experimentally shown that proposed scheme has outperformed the existing schemes against those attacks.

1 Introduction

Statistical undetectability is one of the main aspects of a steganographic algorithm. In steganographic research several algorithms have been proposed for preserving statistical features of the cover for achieving more security. Recently Solanki et al. [1] have proposed a statistical restoration method where a portion of cover coefficients is allocated for embedding and another portion is used to restore the statistics. To ensure Minimum Mean Square Error (MMSE) criteria while modifying the histogram, all the bins of the target histogram are compensated in an increasing order by mapping the input data with values in increasing order [2]. To the best of our understanding, histogram restoration in this fashion limits the application of the algorithm to only Gaussian cover. Further the algorithm shows some erratic behavior in low probability tail areas. The proposed scheme tries to overcome this limitation of a Gaussian cover assumption and provides better restoration of image histogram for non-Gaussian cover distribution as well. The rest of the paper is organized as follows: In section 2, we formalize the important definitions used in the restoration scheme, the detailed algorithm and experimental results are given in section 3, the application of the scheme towards the steganographic algorithms as well as a new steganographic scheme is proposed in section 4. The paper is concluded in section 5.

[⋆] Corresponding author.

[1] The first author gratefully acknowledge the support received from Infosys Technologies Ltd., Bangalore, under the Infosys Fellowship Award.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 297–302, 2009.

2 Mathematical Formulation of Proposed Scheme

The proposed restoration scheme is dependent on the embedding scheme. The whole idea of embedding and restoring is that some of image pixels are used for embedding and rest are used for restoration.

Let the cover image, stego image (i.e. embedded but not yet compensated) and compensated stego image (stego image after compensation) be defined by C, S and R respectively. Suppose C_{ij}, S_{ij} and R_{ij} represent the $(i,j)^{th}$ pixel of C, S and R images respectively ($0 < i < m$, $0 < j < n$, m is number of rows and n is number of columns of image matrices). Histogram ($h(I)$) of an image I can be represented as $h(I) = \{h(0), h(1), h(2), \ldots, h(L-1)\}$

Definition 1 (*Embed Matrix*(Ψ)**).** *It is a $m \times n$ characteristic matrix representing whether a pixel has been used for embedding or not. $\Psi(i,j) = 1$ if $(i,j)^{th}$ pixel is used for embedding and $\Psi(i,j) = 0$ if $(i,j)^{th}$ pixel is not used for embedding.*

Definition 2 (*Compensation Vector*(Ω)**).** *It is a one dimensional vector with length L where L is number of existing gray levels in the cover image (C). $\Omega(k)$ $= u$ means that u number of pixels with gray value k can be used for restoration.*

Definition 3 (*Changed Matrix*(Γ)**).** *It is a $L \times L$ matrix where L is number of existing gray levels in the cover image (C). $\Gamma(x,y) = \lambda$ means during embedding λ number of pixels are changed from gray value x to gray value y.*

Changed Matrix (Γ) is computed as given below:

$$\Gamma(x,y) = \sum_{i=0}^{m} \sum_{j=0}^{n} eq(C_{ij}, x) \times eq(S_{ij}, y) \times \epsilon_{ij} \tag{1}$$

where $eq(a,b) = 1$ if a = b and $eq(a,b) = 0$ if $a \neq b$.

Definition 4 (*Compensation Matrix*(ξ)**).** *It is a $L \times L$ matrix where L is number of existing gray levels in the cover image (C). $\xi(x,y) = \lambda$ means during embedding number of times x is changed to y minus number of times y changes to x is λ.*

Compensation Matrix (ξ) has been formed as following:

$$\xi = UT(\Gamma - \Gamma^T) \tag{2}$$

where $UT(M)$ means upper triangulation of matrix M.

3 Proposed Restoration Scheme

3.1 Algorithm

The detailed restoration algorithm is described below using the matrices and vector defined in the previous section.

Algorithm Restoration
begin
for each element $\xi(i,j)$ of compensation matrix
do
{
Step1: k = $\xi(i,j)$

Step2: If $k > 0$, k number of pixels with gray value i from the set of pixels used for compensation are changed to gray value j for full compensation. If $k < 0$, reverse operation is done i.e. k number of pixels with gray value j from the set of pixels used for compensation are changed to gray value i for full compensation.

Step3: Modify the $Compensation\ Vector$ (Ω) to reflect the pixel changs under taken in step 2 as in Eq.(3) below

$$\Omega(i) = \begin{cases} \Omega(i) - k & \text{if } \Omega(i) > k \\ \\ 0 & \text{if } \Omega(i) \leq k \end{cases} \tag{3}$$

/* Here we assume that for $\Omega(i) < k$, full compensation is not possible. Further research can be possible to improve this situation.*/
}
End Algorithm Restoration

3.2 Restoration with Minimum Distortion

The additional noise added due to compensation is an important issue. The goal is to design a restoration procedure in such a way that additional noise should be kept minimum. In the proposed compensation procedure, the noise introduced depends on the embedding algorithm used. The total noise (η) introduced at the time of restoration can be estimated by

$$\eta = \sum_{i=0}^{L-1} \sum_{j=1}^{abs[\hat{h}(i)-h(i)]} abs(i - k_j) \tag{4}$$

where $\hat{h}(i)$ and $h(i)$ is the histogram of the stego and cover images respectively. $L - 1$ is the no. of bins in the histogram. k_j ($0 \leq k_j \leq L - 1$) is a bin that is used to repair atleast one unit of data in i^{th} bin.

Lemma 1. *With any restoration scheme the minimum total noise* $\sum_{i=0}^{L-1} abs$ $[\hat{h}(i) - h(i)]$.

Proof. The total noise (η) introduced at the time of restoration is

$$\eta = \sum_{i=0}^{L-1} \sum_{j=1}^{abs[\hat{h}(i)-h(i)]} abs(i - k_j) \tag{5}$$

where $1 \leq abs(i - k_j) \leq L - 1$. η is minimum when $abs(i - k_j) = 1$. Putting $abs(i - k_j) = 1$ in Eq. 5 we get

$$\eta = \sum_{i=0}^{L-1} abs[\hat{h}(i) - h(i)] \tag{6}$$

□

Lemma 2. *With proposed restoration scheme the total noise (η) is minimum if maximum noise per pixel due to embedding is* 1.

Proof. Since the proposed restoration scheme is based on simple pixel swapping strategy i.e. if a the gray level value α of a pixel is changed to β during steganographic embedding, at the time of restoration, a pixel with gray level value β is changed to α.

During embedding with ±1 embedding, the gray level value of a pixel, x can be changed into either $x + 1$ or $x - 1$. Hence during restoration the proposed scheme restores bin x value is repaired from either bin $x + 1$ or $x - 1$ according to embedding. It is to be noted that maximum noise that can be added during restoration for one member of a bin is atmost 1 since we are using only the neighboring bins for compensation. Hence, with ±1 embedding scheme (or any other steganographic scheme where noise added during embedding per pixel is atmost 1), the proposed scheme increments or decrements gray value by 1 i.e. $abs(i - k_i) = 1$.

From Eq. (6), the total noise (η) introduced at the time of restoration is

$\eta = \sum_{i=0}^{L-1} \sum_{j=1}^{abs[\hat{h}(i)-h(i)]} abs(i - k_j)$

since in the proposed restoration scheme $abs(i - k_i) = 1$, putting this value in above equation, we get

$\eta = \sum_{i=0}^{L-1} \sum_{j=1}^{abs[\hat{h}(i)-h(i)]} (1)$

$= \sum_{i=0}^{L-1} abs[\hat{h}(i) - h(i)]$ □

So from *Lemma* 1 *and* 2, we can conclude that with proposed restoration scheme minimum amount of noise is added during restoration if maximum noise per pixel due to embedding is atmost 1.

4 Experimental Results

4.1 Comparison with Existing Method(s)

In our experiments, LSB matching at embedding rate 0.125 *bpp* is used as the steganographic embedding method on two standard images as *Dinosaur* and *Baboon*. In Fig. 1 and 2, Histogram, Difference Histogram Before Compensation, Difference Histogram After Compensation by Solanki's Method and Difference Histogram After Compensation by Proposed Method are given for Dinosaur and Baboon images respectively. It may be seen from Figs. 1 and 2 that the proposed scheme provides better restoration than Solanki. et. al's scheme.

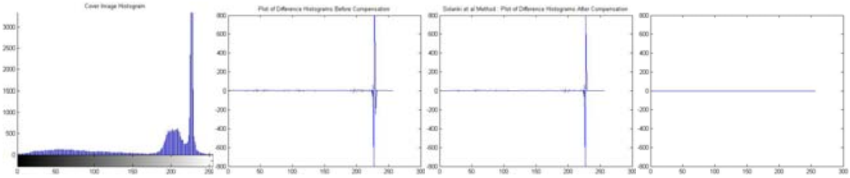


Fig. 1. Histogram, Difference Histogram Before Compensation, Difference Histogram After Compensation by Solanki's Method and Difference Histogram After Compensation by Proposed Method respectively for Dianosaur Image

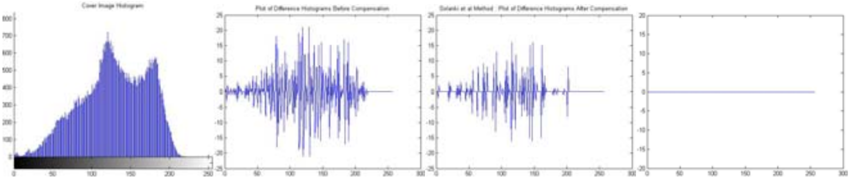


Fig. 2. Histogram, Difference Histogram Before Compensation, Difference Histogram After Compensation by Solanki's Method and Difference Histogram After Compensation by Proposed Method respectively for Baboon Image

4.2 Application towards Steganography

To analyze the applicability of the proposed scheme towards more secure steganographic algorithm, the LSB matching is used as steganographic embedding scheme on images images from UCID dataset [5]. Some of the state of art targeted attacks like Ker's calibrated HCF [3], HCF of Adjacency Histogram based attacks [3] and Jun Zhang et al 's high frequency noise based attack [4] are considered for experimentation. To evaluate the steganographic security using our proposed scheme, we have used Area under the Receiver Operating Characteristic Curve (A_{ROC}) and the Detection accuracy (P_{detect}) [6].

In Figs. 3 and 4, we have shown the P_{detect} and (A_{ROC}) plots for comparing the proposed restoration scheme with the restoration scheme proposed in [1]. It

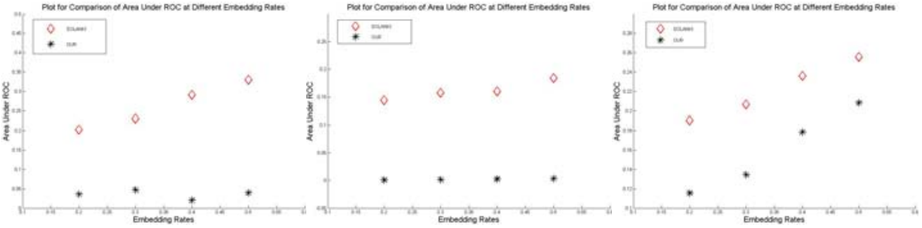


Fig. 3. Comparison of Area under ROC against Jun Zhang et al 's targeted attack, HCF Calibration Attack and HCF Adjacency Attack

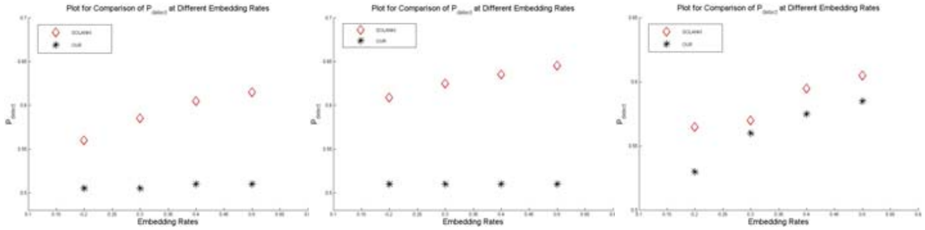


Fig. 4. Comparison of P_{detect} against Jun Zhang et al 's targeted attack, HCF Calibration Attack and HCF Adjacency Attack

can be observed from Fig. 3 and 4, our proposed method geartly outperforms the competing scheme against mentioned targeted attacks.

5 Conclusion

In this paper we have proposed a new statistical restoration scheme which can be used for preserving the histogram i.e. first order statistics of the cover image after embedding and thus making the data hiding process robust against first order statistic based steganalytic attack. Moreover the proposed scheme does not assume any particular distribution for the cover image and hence gives better performance than scheme given in [1] especially for non-Gaussian covers. It must be mentioned that the additional noise added during restoration depends on the embedding algorithm for proposed scheme and is a topic of future research.

References

1. Solanki, K., Sullivan, K., Madhow, U., Manjunath, B.S., Chandrasekaran, B.S.: Statistical Restoration for Robust and Secure Steganography. In: Proc. IEEE International Conference on Image Processing, Genova, Italy (September 2005)
2. Tzschoppe, R., Bauml, R., Eggers, J.J.: Histogram modifications with minimum MSE distortion. Tech. rep., Telecom. Lab., Univ. of Erlangen-Nuremberg (December 2001)
3. Ker, A.: Steganalysis of LSB matching in grayscale images. IEEE Signal Process. Lett. 12(6), 441–444 (2005)
4. Zhang, J., Cox, I.J., Doerr, G.: Steganalysis for LSB Matching in Images with High-frequency Noise. In: Proc. IEEE 9th Workshop on Multimedia Signal Processing, MMSP 2007, October 1-3, September 16-October 19, pp. 385–388 (2007)
5. Schaefer, G., Stich, M.: UCID - An Uncompressed Colour Image Database. In: Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia, vol. 5307, pp. 472–480 (2004)
6. Solanki, K., Sarkar, A., Manjunath, B.S.: YASS: Yet Another Steganographic Scheme that Resists Blind Steganalysis. In: Furon, T., Cayre, F., Doërr, G., Bas, P. (eds.) IH 2007. LNCS, vol. 4567, pp. 16–31. Springer, Heidelberg (2008)

Prediction of Cloud for Weather Now-Casting Application Using Topology Adaptive Active Membrane

Sitansu Kumar Das, Bhabatosh Chanda, and Dipti Prasad Mukherjee

Electronics and Communication Sciences Unit
Indian Statistical Institute, Kolkata, India

Abstract. Prediction of meteorological images from a given sequence of satellite images is an important problem for weather now-casting related application. In this paper, we have concentrated on the dominant object of a meteorological image, namely cloud, and predicted its topology within a short span based on currently available sequence images. A topology adaptive active membrane is used to model the intensity profile of the cloud mass and based on a series of such membranes a future membrane is extrapolated under linear regression method and affine shape constraint. The proposed scheme is tested with ISRO satellite (Kalpana 1) images.

1 Introduction

For weather now-casting, meaningful, short-term atmospheric conditions need to be predicted using extrapolated meteorological images. The extrapolation is done based on a few already available sequence images. Recent studies in weather now-casting have shown promising results [1][2]. In this paper we concentrate on the most dominant object of a meteorological image, namely cloud and its extrapolation scheme.

Assuming a square finite element grid over the entire cloud image surface, minimizing energy function involving approximation of finite element grid to the intensity of cloud and following derivation of [3] we get,

$$\begin{aligned} &\nabla X_{i,j}^{t-1} = \beta\left(I\left(X_{i,j}^{t-1}, Y_{i,j}^{t-1}\right)\right)\triangle P_X, \nabla Y_{i,j}^{t-1} = \beta\left(I\left(X_{i,j}^{t-1}, Y_{i,j}^{t-1}\right)\right)\triangle P_Y,\\ &\nabla Z_{i,j}^{t-1} = \beta\left(I\left(X_{i,j}^{t-1}, Y_{i,j}^{t-1}\right)\right)\triangle P_Z, \end{aligned} \tag{1}$$

where, ∇ denotes Laplacian in discrete domain. The FEM vertices are indexed using (i, j) and the corresponding image pixel value at time t is given by $I(X_{i,j}^t, Y_{i,j}^t)$. The external force components along different axes are defined as,

$$\begin{aligned} &\triangle P_X = -\textstyle\sum_\eta \sum_{k=0}^q N(k)\left|\triangle G_x(I)\right|, \triangle P_X = -\sum_\eta \sum_{k=0}^q N(k)\left|\triangle G_y(I)\right|,\\ &\triangle P_Z = -\rho(1+\exp(|\rho|)). \end{aligned} \tag{2}$$

In (2), ρ is defined as $\rho = (I(X_{i,j}, Y_{i,j}) - Z_{i,j})$. $|\nabla G_x(I)|$ and $|\nabla G_y(I)|$ are Gaussian convolved image gradients along x and y axes respectively. The external energy at one vertex is calculated summing external energies of $4N$ vertices

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 303–308, 2009.

specified by domain η and $N(k)$ is the weight of the external energy at the kth point out of all the q number of discrete points in the inter-vertex distance.

In the first frame, the weight $\beta(I)$ is taken as linear function, $\beta(I(X,Y)) = (1-\gamma)\beta_{high}$, where $\gamma = \frac{I(x,y)}{Dyn.Range(I)}$, where $Dyn.Range(I)$ denotes the dynamic range of gray level in Image I. β_{high} is set experimentally. From (1), we ultimately get the active membrane evolving equation for segmentation [3] as $V^t = V^{t-1} + \Delta t(\beta. * F_V + AV^{t-1})$, where A is the stiffness matrix. V is position matrix. Fv is force matrix. The operation '$\cdot *$' denotes element wise multiplication. We assume that we have *a priori* estimation V^{t-1} at $(t-1)$th iteration for the current iteration t. I and $\triangle t$ denote identity matrix and time step respectively.

For each iteration we compare the inter-vertex Euclidean distance between a pair of neighbouring vertices (i,j) and (k,l) with the $D^{k,l}_{i,j}$ given as,

$$D^{k,l}_{i,j} = g_D + \frac{I(X_{i,j}, Y_{i,j}) + I(X_{k,l}, Y_{k,l})}{2 \times Range(I)} \times (D_{high} - g_D), \qquad (3)$$

where, g_D is the initial inter-vertex distance of the membrane and D_{high} are application dependent preset constants. Each inter-vertex distance or link have two neighbouring finite elements. While checking the connection between two neighbouring vertices, we also check the status or existence of its two neighbourhood elements. If inter-vertex distance exceeds preset limit or there exists no neighbourhood element in either side of the link then connectivity or link between the vertices is deleted. After this, we delete all the vertices having zero connected neighbour. Simultaneously, we modify matrix A by deleting corresponding row and column of matrix A and then start the next iteration. Hence, the size of matrix A gets reduced compared to the previous iteration of evolution of the membrane. We stop the evolution of membrane when for same number of membrane vertices in previous and current iteration, $\left\|V^t - V^{t-1}\right\| < \epsilon$, where ϵ is a very small preset number.

Once the membrane evolution stops, we get different membrane pieces representing the clouds in the first cloud image. In the second cloud image, membrane is evolved using template evolution scheme in [4] to track the clouds. We describe it next briefly.

2 Tracking of Cloud Mass

Assuming the square finite element grid over cloud resulting from segmentation of clouds described in Section 1, minimizing energy function involving approximation of finite element grid to the intensity of cloud and following derivation of [4] we get,

$$\sum_{(k,l)} \left(1 - \frac{(d^{i,j}_{k,l})^0}{(d^{i,j}_{k,l})^{t-1}}\right)(V^t_{i,j} - V^t_{k,l}) = \delta_{i,j}\triangle I(V^t_{i,j}) + \triangle\delta_{i,j}\triangle^2 I(V^t_{i,j}), \qquad (4)$$

where $\delta_{i,j}$ is equal to $\left(I(V^t_{i,j}) - I_p(V_{i,j})\right)$. In (4), $(d^{i,j}_{k,l})^0$ is defined as the distance between the vertices (i,j) and (k,l). We take $(d^{i,j}_{k,l})^0$ equal to $\|V_{i,j} - V_{k,l}\|$ in the starting of the evolution. We use superscript 0 in (4) for denoting initial membrane parameters. $I_P(V_{i,j})$ and $\nabla I_P(V_{i,j})$ are cloud intensity and gradient respectively found in the first cloud image at membrane position $V_{i,j}$. From (4) we ultimately get the active membrane evolving equation for tracking [4] as, $(\mathtt{I} + \triangle t \times A)V^t = (V^{t-1} + \triangle t \times FTv)$, where $\mathtt{I}$ and Δt denote identity matrix and time step respectively. FTv are the force matrix at the membrane vertices. With the help of evolving equation for tracking, we track the cloud in the consecutive cloud images. After that, extrapolating the cloud tracking result, we predict the cloud for the next image. The prediction methodology is described in the next section.

3 Now-Casting scheme

Given $I^1, ..., I^{\tau-1}, I^{\tau}$ meteorological satellite images obtained at time instants $1, ..., \tau$ respectively, being a sufficiently small finite integer (typically 30 minutes), the objective of the extrapolation scheme is to predict the cloud mass at $I^{\tau+1}, ...,$ $I^{\tau+n}$ for a positive integer n.

With the help of linear regression scheme, the shape and location of the cloud in the image $I^{\tau+n}$ can be deduced as $V^{\tau+n} = (\tau + n)\xi(V^{\tau}) + \zeta(V^{\tau})$, where, $\xi(V^{\tau}) = \frac{\Sigma\kappa\Sigma V^{\kappa} - \tau\Sigma V^{\kappa}\kappa}{(\Sigma\kappa)^2 - \tau\Sigma\kappa^2}$ and $\zeta(V^{\tau}) = \frac{1}{\tau}(\Sigma V^{\kappa} - \xi(V^{\tau})\Sigma\kappa)$, for $\kappa = 1...\tau$. The expressions of $\xi(V^{\tau})$ and $\zeta(V^{\tau})$ are found out minimizing the sum of distances from the membrane points $V^1, ..., V^{\tau}$ with respect to linear regression straight line. We can discuss it with the help of an example given in Fig.1. Suppose membranes in Fig.1(a) and (b) represent V^1 and V^2 respectively, as τ equals to 2. We denote the (i, j)th vertex in these two membrane are denoted by $V^1_{(i,j)}$ and $V^2_{(i,j)}$. So following the linear regression scheme, we get the (i, j)th vertex of $V^3_{(i,j)}$ as $V^3_{(i,j)} = \zeta(V^2_{(i,j)}) + 3\xi(V^2_{(i,j)})$.

The intensity in the image $I^{\tau+1}$ at the membrane vertices are generated following the similar principle. For example the predicted intensity at the vertex point $V^{\tau+1}_{(i,j)}$ is define as,

$$I^{\tau+n}(X^{\tau+n}_{(i,j)}, Y^{\tau+n}_{(i,j)}) = (\tau + n)\xi(I^{\tau}(X^{\tau}_{(i,j)}, Y^{\tau}_{(i,j)})) + \zeta(I^{\tau}(X^{\tau}_{(i,j)}, Y^{\tau}_{(i,j)})). \quad (5)$$

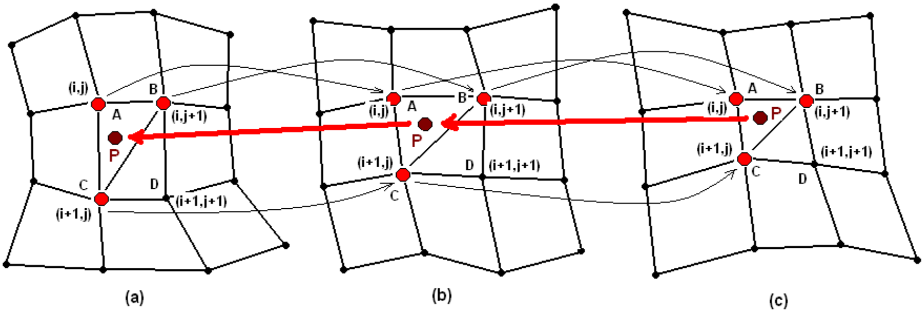


Fig. 1. (a): Membrane V^1, (b): Membrane V^2, (c): Membrane V^3

Next we divide each finite element into two triangles. For example, the finite element with grid indices (i, j), $(i, j + 1)$, $(i + 1, j + 1)$ and $(i + 1, j)$ is divided into two triangles ABC and BCD with grid indices (i, j), $(i, j+1)$, $(i+1, j)$ and $(i, j + 1)$, $(i + 1, j + 1)$, $(i + 1, j)$ respectively as shown in Fig.1. P is the point inside the triangle ABC. Here we assume that the change in shape of the triangle ABC occurs maintaining the affine shape constraint in the time of cloud tracking and extrapolation. Therefore, the point inside the triangle also moves following the affine shape constraint. So, to find out the location of the point P in I^2 we first find out the affine shape constraint coefficients for the transformation of the triangle ABC from image I^3 to I^2. We know the three vertices of triangle ABC in both the images I^2 and I^3. Therefore, we can write with the help of affine shape constraint as $[X^2_{(k,l)}; Y^2_{(k,l)}] = [a, b; c, d][X^3_{(k,l)}; Y^3_{(k,l)}]+[e; f]$, where a, b, c, d, e and f are six unknown affine shape constraint coefficients and $(k, l) \in \{(i, j), (i, j + 1), (i + 1, j)\}$. Affine shape constraint equation gives six equations from where six unknown affine transform coefficients are solved. Now, P is a common point of the membrane inside the triangle ABC and we denote its coordinate in image I^3 as V^3. The corresponding point inside the triangle ABC in image I^2 can be define as $[X^2; Y^2] = [a, b; c, d][X^3; Y^3] + [e; f]$. With above described procedure, we can find out previous locations of all the points of extrapolated cloud in the images I^2 and I^3. Therefore, using (5), we can generate all the point inside the cloud I^3. In the next section we show the experimental results of our method.

4 Results

We have implemented the proposed methodology to track and predict cloud base on Infra-red images of Kalpana-1 satellite of the western part of India received in interval of half an hour (between 5:30am to 9:00am) on 16th June 2009. Fig. 2(a) shows the segmentation of a cloud image at 5:30am with the help of the procedure described in Section 1. Then, Figs. 2(b)-(d) show the final result of tracking of cloud at 6:00am, 6:30 am and 7:00am respectively by the membrane representing the segmented cloud in Figs. 2(a). Finally, based on the cloud we get in Figs. 2 (a)-(d), we have generated cloud at 7:30am, 8:00am, 8:30am and 9:00am as shown in Figs. 2(e)-(h) respectively, using the extrapolation scheme described in Section 3.

In all the above examples, we take $D_{high} = 3g_d$, $g_d = 5$ and $\beta_{high} = 0.015$. The entire approach was implemented in Matlab 6.5 in Pentium 4, 2.1 GHz PC. In the next section we compare our method with the predicted cloud based on pixel based linear extrapolation method.

4.1 Comparison

To show the efficacy of our proposed method we evaluate the root mean square error of intensity difference between extrapolated clouds and actual cloud images shown in Figs. 2(m)-(p). Then we evaluate the root mean square errors of intensity difference between extrapolated clouds based on the pixel based linear

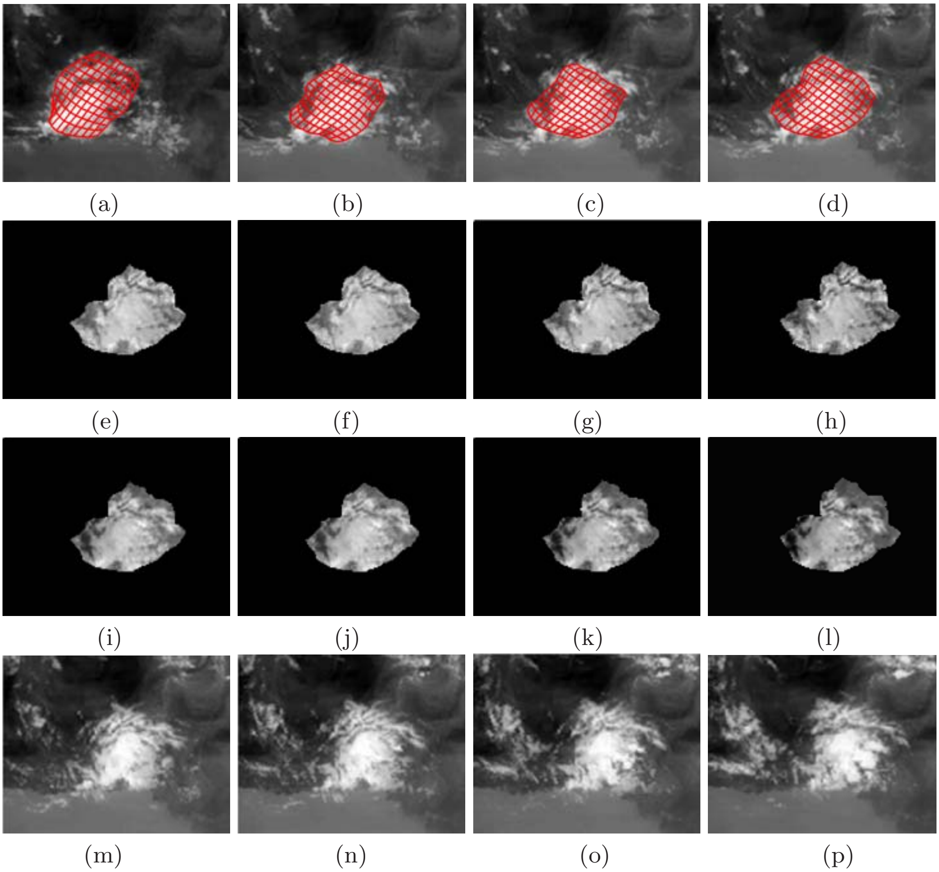


Fig. 2. (a): Segmentation of cloud at 5:30am, (b)-(d): Tracking of the cloud at 6:00am, 6:30am and 7:00am respectively. (e)-(h): Predicted cloud at 7:30am, 8:00am, 8:30am and 9:00am respectively by proposed method. (i)-(l): Predicted cloud at 7:30am, 8:00am, 8:30am and 9:00am respectively by pixel based linear regression method. (m)-(p): Actual cloud image at 7:30am, 8:00am, 8:30am and 9:00am respectively.

Table 1. RMS error of the proposed approach and pixel based linear regression method

	RMS error with Figs. 2(m)-(p)			
	7:30am	8:00am	8:30am	9:00am
Proposed Mehtod Figs. 2(e)-(h)	16.9	27.3	35.52	40.9
Pixel based Method Figs. 2(i)-(l)	24.4	34.6	41.7	47.7

regression model and the actual cloud images of Figs. 2(m)-(p). Table 1 shows the proposed approach performs better than the pixel based linear regression model.

5 Conclusions

We have developed an active membrane and affine transformation based cloud prediction framework for weather now-casting. The initial membrane is independent of any a priori knowledge of the cloud. The framework can track and predict cloud efficiently.

Acknowledgments. The support of ISRO in sponsoring this work is gratefully acknowledged.

References

1. Grose, A.M.E., Simth, E.A., Chung, H.S., Ou, M.L., Sohn, B.J., Turk, F.J.: Possibility and limitations for quantitative precipitation forecasts using nowcasting methods with infrared geosynchronous satellite imagery. Journal of Applied Meteorology 41, 763–785 (2002)
2. Roberts, R.D., Rutledge, S.: Nowcasting storm initiation and growth using goes-8 and wsr-88d data. Weather and Forecasting 18, 562–584 (2003)
3. Das, S.K., Mukherjee, D.P.: Topology adaptive active membrane. In: Ghosh, A., De, R.K., Pal, S.K. (eds.) PReMI 2007. LNCS, vol. 4815, pp. 95–102. Springer, Heidelberg (2007)
4. Das, S.K., Mukherjee, D.P.: Multiple objects tracking via topology adaptive active membrane. In: Procedings Sixth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 665–672 (2008)

Image Enhancement Using Multi-objective Genetic Algorithms

Dinabandhu Bhandari, C.A. Murthy, and Sankar K. Pal

Center for Soft Computing Research, Indian Statistical Institute,
Kolkata 700108, India
dinabandhu.bhandari@gmail.com, {murthy,sankar}@isical.ac.in

Abstract. Given an image, there is no unique measure to quantitatively judge the quality of an image enhancement operator. It is also not clear which measure is to be used for the given image. The present work expresses the problem as a multi-objective optimization problem and a methodology has been proposed based on multi-objective genetic algorithm (MOGA). The methodology exploits the effectiveness of MOGA for searching global optimal solutions in selecting an appropriate image enhancement operator.

Keyword: Image Enhancement, Multi Objective Genetic Algorithms, Ambiguity Measures.

1 Introduction

There are many problems in the area of pattern recognition and image processing [1], where we need to perform efficient search in complex spaces in order to achieve an optimal solution. One such problem is contrast enhancement by gray-level modification, where the purpose is to improve the picture quality, more specifically, to improve the quality for visual judgment and/or machine understanding. In this case two major tasks are involved.

1. Selection of an appropriate transformation/mapping function (operator) for obtaining a desired output. Usually, a suitable nonlinear functional mapping is used to perform this task.
2. Selection of an evaluation function to define a quantitative measure of an enhanced image.

Not every kind of nonlinear function will produce a desired (meaningful) enhanced version [2]. Given an image, it is difficult to select a functional form which will be best suited without prior knowledge of the image statistics. Even if we are given the image statistics it is sometimes possible only to estimate approximately the function required for enhancement [2,3]. The selection of the exact functional form still needs human interaction in an iterative process. To make process of evaluation of the quality of an image (picture) objective, it is necessary to define an objective function which will provide a quantitative

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 309–314, 2009.

measure for enhancement quality. Various evaluation functions are available in literature such as entropy, compactness, index of area coverage (IOAC), Divergence etc. to measure automatically the quality of the enhanced image. It has also been observed that all these measures are suitable for different kinds of images.

The effectiveness of multi-objective genetic algorithm (MOGA) is exploited to find a number of solutions in the Pareto-optimal front. Then depending on the image characteristics, an optimal solution is used as enhancement operator. Identifying the Pareto front is useful because it can be used to make well-informed decisions that balance trade-offs between the objectives. The problem here is to select automatically an optimum set of parameter values of 4 basic enhancement functions and their combining weightages that optimizes the fitness value (evaluation measure). The algorithm uses both spatial and grayness ambiguity measures to qualitatively evaluate the enhanced image. The methodology is demonstrated on a wide variety of images and obtained satisfactory results. Due to scarcity of space only a few have been presented in this paper. The problem of selecting an image enhancement operator and objective functions are described in section 2. Proposed methodology based on MOGA is explained in Section 3. Section 4 presents the results.

2 Image Enhancement Operator and Objective Functions

In the problem of gray-level re-scaling, each pixel is directly quantized to a new gray level in order to improve the contrast of an image. The simplest form of the functional mapping may be expressed as

$$x'_{mn} = x_{max}.f(x_{mn}) \tag{1}$$

where, x_{mn} = gray value of the (m,n)th pixel of the input image (original), x'_{mn} = transformed value of the (m,n)th pixel (enhanced), x_{max} = maximum value of the gray level dynamic range and $f(x)$ is the prescribed transformation function.

Most commonly used transformation functions [2,3,4] are shown in fig. 1. The mapping function $f_1(.)$ depicted in Figure 1 (a) increases the contrast within the darker area of the image, while the application of a function $f_3(.)$ as in Figure 1 (c) will produce effects exactly opposite to that of function $f_1()$. The function $f_2(.)$ shown in Figure 1 (b) will result in stretching of the middle range gray levels and the function $f_4(.)$, in Figure 1 (d), will drastically compress the middle range values, and at the same time it will stretch the gray levels of the upper and lower ends. The mathematical forms of the above mentioned mapping functions are given below.

$$\begin{aligned} f_1(x) &= \tfrac{Ax^2}{1+Ax^2} \qquad = \tfrac{x^2}{par_1+x^2} \quad \text{or} \\ &= par_1\ log(x) \end{aligned} \tag{2}$$

where, par_1 and A are positive constants.

$$f_2(x) = [1 + (\frac{x_{max} - x_{min}}{par_6})^{par_7}]^{-1} \tag{3}$$

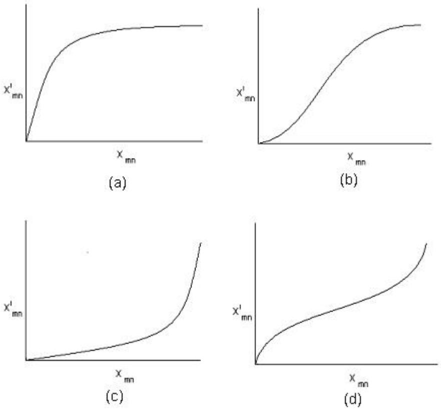


Fig. 1. Mapping functions commonly used for image enhancement

where, par_6 and par_7 are positive constants, and x_{min} and x_{max} are the minimum and maximum gray levels in the image.

$$\begin{aligned} f_3(x) &= par_2\ [G(x)]^2 + par_3\ x + par_4 && \text{where } 0 < par_2, par_3, par_4 < 1, \\ G(x) &= x - par_5 && \text{for } x > par_5 \\ &= 0 && \text{Otherwise} \end{aligned} \tag{4}$$

where, $x_{min} < par_5 < x_{max}$. Finally, the function in Figure 2 (d) is given as:

$$f_4(x) = \frac{1}{x_{max}}[x_{max} - par_6\{(\frac{x_{max}}{x} + par_8) - 1\}^{-par_7}] \tag{5}$$

where, par_6 and par_7 are positive constants, and par_8 is the value of $f(x)$ for $x = 0$.

Not all nonlinear functions will produce desired (meaningful) enhanced versions [2] of a particular image. Given an image, it is very difficult to decide a nonlinear functional form that will be best suited for highlighting its object without prior knowledge of image statistics. Even if the image statistics is known, it is possible only to estimate approximately the function required for enhancement and the selection of the exact functional form still needs human interaction in an iterative process. Since we do not know the exact function which will be suited for a given image, it seems appealing and convenient to use one general functional form which will yield the four functions mentioned above as special cases and possibly others. As an illustration one may consider a convex combination of these four functions e.g.,

$$\begin{aligned} f(.) = par_9 f_1(.) + par_{10} f_2(.) + par_{11} f_3(.) + par_{12} f_4(.) \text{ where} \\ par_9 + par_{10} + par_{11} + par_{12} \ = \ 1 \end{aligned} \tag{6}$$

Here, the multipliers (par_9, par_{10}, par_{11}, par_{12}) are to be chosen according to the importance (suitability) of a function for a given image. On the other hand, parameters $(par_1, par_2, ..., par_8)$ of the respective functions, are to be defined according to the quality of enhancement desired. It may be noted that this combination will enable one to stretch/compress any region of an image one may desire. Therefore, the selection of a transformation function boils down to determining an optimum set of values of these 12 parameters in order to achieve a desired enhancement.

Once the enhancement function is known, to quantify the enhancement quality of the image individual judgment is required that makes the optimal decision subjective. Therefore, we need an evaluation function for quantifying the enhanced output, i.e., to objectively evaluate the subjective evaluation. One can consider entropy (H), compactness, IOAC as they have been successfully used as grayness and spatial ambiguity measures for image enhancement and segmentation problems [5,6,7,4]. Entropy of an image (X) considers the global information and provides an average amount of fuzziness in grayness of X, i.e., the degree of difficulty (ambiguity) in deciding whether a pixel would be treated as black (dark) or white (bright). Compactness and IOAC on the other hand, take into account the local information and reflect the amount of fuzziness in shape and geometry (spatial domain) of an image. Therefore, the concept of minimization of these ambiguity measures may be considered as the basis of a fitness (evaluation) function. In [4], a composite measure (e.g., product of both grayness and spatial ambiguity measures) is used as the evaluation function so that minimization of this composite measure implies achieving minimum ambiguity (fuzziness) of an image. However, the composite measure may not be an ideal choice as each measure is given equal importance. For example, for an image consisting of a compact object, compactness would be given higher importance than IOAC. On the other hand, for an image consisting of an elongated object, IOAC would be given more importance. Instead this can be viewed as a multi-objective optimization problem. One can find the Pareto optimal solutions and finally, select the best transformation function based on the image characteristics.

3 Proposed Methodology

In this work, the real coded genetic algorithm is adapted, where a string of 12 real numbers is considered as chromosome. The domains of the parameters considered are shown in the following table.

Parameters	Range
$par_1 - par_4$ and $par_8 - par_{12}$	$[0, 1]$
par_5 and par_6	$[x_{min}, x_{max}]$
par_7	$[1, 3]$.

The single point crossover operation is adapted here with crossover probability 0.8 and mutation probability is varied from 0.3 to 0.01 with iteration. The concept of minimization of Entropy, Compactness and index of area coverage is considered as the basis of a fitness (evaluation) functions.

The non-dominated sorting algorithm NSGAII, proposed by Srinivas and Deb [8,9], is adopted in implementing the proposed methodology. The population size is assumed to be 100. Once the child population is generated using the parent population, both the populations are combined together. Among these 200 offspring and parent solutions, 100 are being selected for the selection, crossover and mutation operations and a child population of same size is created. The non-dominated sorting is then used to classify the entire population. Once the non-dominated sorting is over, the new population of size 100 is generated taking solutions from the best non-dominated front and continues with second front and so on. The algorithm ensures the diversity among the selected solution in the parent population. However, during the later stages of the algorithm, it is likely that most solutions in the population lie in the best non-dominated front.

4 Implementation and Results

The proposed methodology has been tested on several images, having bimodal as well as multi-modal histograms. The images consist of compact as well as elongated objects. The algorithm uses the standard S-function of Zadeh to calculate the fuzziness of a pixel. Three measures entropy, compactness and IOAC of the images are used as the evaluation functions. The algorithm is executed for 500 iterations. The Pareto front obtained after 500 iterations contains 100 solutions. Then depending on the image characteristics, a solution set is taken to transform the image. For the simulation, gray scale [0, 255] images of size 256×256 pixels are used. In order to evaluate the performance of the proposed methodology, results are compared with the methodology developed using simple genetic algorithm (SGA) with elitist model. In the implementation of SGA, composite evaluation measure (product of entropy, compactness and IOAC) is minimized to obtain the optimum solution. The algorithm is also executed 500 iterations with a population size 100.

Fig 2(a) illustrates a low contrasted image. 2(b) and 2(c) depict the output obtained using SGA and MOGA based methodologies. Out of the several solutions obtained in the Pareto front using MOGA, the solution producing medium entropy, and low compactness and IOAC measures are found to produce better results for enhancement. This may be due to the reason that entropy tends to 0 when there is least ambiguity that means all the pixels lie either in object or background. Similar results are also found for the image provided in Fig 3(a).

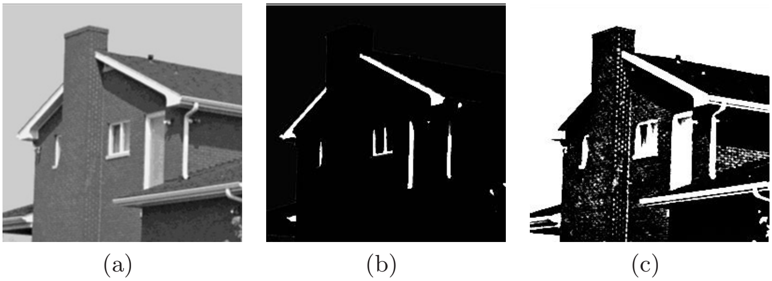
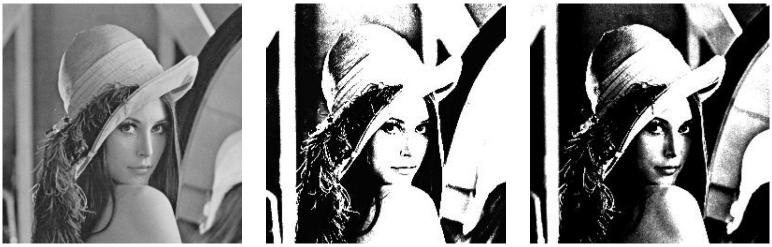

(a) (b) (c)

Fig. 2. The House image (a), (b) and (c)



(a) (b) (c)

Fig. 3. Leena image (a), (b) and (c)

Comparing the images of figs 2 and 3, one can observe the efficiency of the proposed method in enhancing the images.

5 Conclusions

The effectiveness of MOGA in the automatic selection of an optimum image enhancement operator is tested for various bimodal and multimodal images. Although fuzziness measures have been used as fitness values, one may use other measures depending on the problems. Moreover, one can use more measures as objective functions. The algorithm determines number of optimum parameter sets rather than a single set in selecting an appropriate enhancement function.

References

1. Rosenfeld, A., Kak, A.C.: Digital picture processing. Academic Press, New York (1982)
2. Ekstrom, M.P.: Digital image processing techniques. Academic Press, New York (1984)
3. Kundu, M.K., Pal, S.K.: Automatic selection of object enhancement operator with quantitative justification based on fuzzy set theoretic measure. Pattern Recognition Letters 11, 811–829 (1990)
4. Pal, S.K., Bhandari, D., Kundu, M.K.: Genetic algorithms for optimal image enhancement. Pattern Recognition Letters 15, 261–271 (1994)
5. Vlachos, I.K., Sergiadis, G.D.: Parametric indices of fuzziness for automated image enhancement. Fuzzy Sets and Systems 157, 1126–1138 (2006)
6. Cheng, H.D., Li, J.: Fuzzy homogeneity and scale-space approach to color image segmentation. Pattern Recognition 36, 1545–1562 (2003)
7. Munteanu, C., Rosa, A.: Gray-scale image enhancement as an automatic process driven by evolution. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 34(2), 1292–1298 (2004)
8. Srinivas, N., Deb, K.: Multiobjective function optimization using nondominated sorting genetic algorithms. Evolutionary Computation Journal 2(3), 221–248 (1995)
9. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. John Wiley, Chichester (2001)

A Robust Object Tracking Method Using Structural Similarity in Daubechies Complex Wavelet Domain

Anand Singh Jalal and Uma Shanker Tiwary

Indian Institute of Information Technology, Allahabad, India
{asjalal,ust}@iiita.ac.in

Abstract. Many of the existing algorithms for object tracking that are based on spatial domain features, fail in the presence of illumination variation or change in appearance or pose or in the presence of noise. To overcome these problems, in this paper, we have proposed a new method of object tracking using structural similarity index in complex wavelet transform domain, which is approximately shift-invariant. The reference object in the initial frame is modeled by a feature vector in terms of the coefficients of Daubechies complex wavelet transform. A similarity measure based on structural similarity index is used to find the object in the current frame. The advantage of using structural similarity index in complex wavelet domain is that it allows small spatial translations, rotations and scaling changes, which are depicted in fig. 1. Experimental results illustrate that the proposed algorithm has good performance in noisy video with significant variations in object's pose and illumination. The search for the candidate subframe is made fast by using the motion prediction algorithm.

1 Introduction

Tracking an object in a complex environment is a challenging task. The problem becomes hard when applied to real life situations, such as, sport video analysis to extract highlights, surveillance system to know the suspicious activity, traffic monitoring, and human computer interface to assist visually challenged people [1]. Most of the existing algorithms are unable to track objects in the presence of variations in illumination, appearance and camera angle, as most of these algorithms working in spatial domain use features which are sensitive to these variations. The mean shift algorithm widely used for object tracking proposed by [2], is based on color features. They use Bhattacharyya coefficient as a measure of comparability and optimize the search using mean shift technique. One of the problems of using color or a similar quantity as feature is that it is sensitive to the illumination or appearance changes [3]. In recent years the wavelet feature based techniques have gained popularity in object tracking. One of the features of discrete wavelet transform (DWT) is that the spatial information is retained even after decomposition of an image into four different frequency coefficients. However, one of the major problems with real wavelet transform is that it suffers from shift-sensitivity [4]. In [3] an undecimated wavelet packet transform (UWPT) has been used to overcome the problem of shift sensitivity. However, the UWPT expansion is redundant and computationally intensive. A Daubechies complex wavelet transform can be a better solution which is also approximately

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 315–320, 2009.

shift-invariant. Not many researchers have explored the application of complex wavelet transform (CxWT) for tracking problems. Recently [5] has shown the applicability of CxWT to denoising and deblurring.

In this paper, we present a new object tracking algorithm using Daubechies complex wavelet coefficients as features to represent the object. We have used a similarity measure based on structural similarity metrics in the complex wavelet domain. The advantage of using structural similarity metrics in complex wavelet domain is to allow imperceptible spatial translations, and also small rotations and scaling changes. By applying all computations in the complex wavelet domain, we can attain high immunity to noise and can work at higher levels to reduce the computations as per requirement of the application.

The remaining part of the paper is organised as follows. Section 2 gives an overview of Daubechies complex wavelet transform. Section 3 presents the structural similarity metrics. Section 4 describes the proposed algorithm. Section 5 contains results over real world video sequences, and finally, Section 6 concludes and discusses the open issues for future research.

2 Overview of Daubechies Complex Wavelet Transform

In object tracking we require a feature which remains invariant by translation and rotation because the different video frames may contain a translated and rotated version of the moving object. As Daubechies CxWT is approximately shift-invariant, it can be a good candidate for object tracking.

Any function $f\,(t)$ can be decomposed into complex scaling function $\phi(t)$ and a mother wavelet $\psi(t)$ as:

$$f(t) = \sum_k c_k^{j_0} \phi_{j_0,k}(t) + \sum_{j=j_0}^{j_{\max}-1} d_k^j \psi_{j,k}(t) \tag{1}$$

where, j$_0$ is a low resolution level, { $C_k^{j_0}$ } and { d_k^j } are known as approximation [$\phi(t) = 2\sum_n a_n \phi(2t-n)$] and detail coefficients [$\psi(t) = 2\sum_n (-1)^n \overline{a_{1-n}} \phi(2t-n)$].

Where $\psi(t)$ and $\phi(t)$ shares the same compact support [-N, N+1] and a_ns are coefficients. The a_n s can be real as well as complex valued and $\sum a_n = 1$.

The Daubechies wavelet bases $\{\psi_{j,k}(t)\}$ in one dimension are defined through the above scaling function and multiresolution analysis of $L^2(\Re)$. During the formation of solution if we relax the Daubechies condition for a_n [5], it leads to complex valued scaling function. We have used this symmetric Daubechies complex wavelet transform for tracking.

3 Structural Similarity Metrics

Most similarity measures are unable to capture the perceptual similarity of images/video frames under the conditions of varied luminance, contrast or noise. The

Structural Similarity Measure (SSIM) overcomes these problems [6], which is defined as the distance between two images or sub-images by jointly comparing the mean and variance characteristics. In [7] a complex wavelet domain image similarity measure has been proposed, which is simultaneously insensitive to small luminance change, contrast change and spatial translation.

Due to the symmetric nature of Daubechies complex wavelet, in this paper we are measuring the similarity of the object feature vector using a structural similarity measure in Daubechies complex wavelet domain.

Suppose in the complex wavelet transform domain, $c^x = \{c^x_{i,j} \mid i = 1,\dots,M,\, j = 1,\dots,N\}$ and $c^y = \{c^y_{i,j} \mid i = 1,\dots,M,\, j = 1,\dots,N\}$ are two sets of wavelet coefficients corresponding to two images. Then structural similarity in CxWT domain (CW-SSIM) is given by

$$S_c(c^x, c^y) = \frac{2\left|\sum_{i=1}^{M}\sum_{j=1}^{N} c^x_{i,j} c^{y*}_{i,j}\right| + k}{\sum_{i=1}^{M}\sum_{j=1}^{N} |c^x_{i,j}|^2 + \sum_{i=1}^{M}\sum_{j=1}^{N} |c^y_{i,j}|^2 + k} \tag{2}$$

Where c* denotes the complex conjugate of c, and k is a small positive constant.

4 The Proposed Algorithm

In the proposed algorithm the object is localized in a bounding box around the boundary of the object in the reference frame, either manually or by using an automated detector. The feature vector (the coefficients at level j) of the image inside the bounding box is calculated in terms of the Daubechies complex wavelet coefficients. The position of the object in the next frame is predicted by motion prediction algorithm as proposed by the authors in [8]. Candidate subframes of the object are generated by shifting the predicted position by ±p points. Iteratively comparing the feature vectors using structural similarity, the subframe with maximal similarity is found.

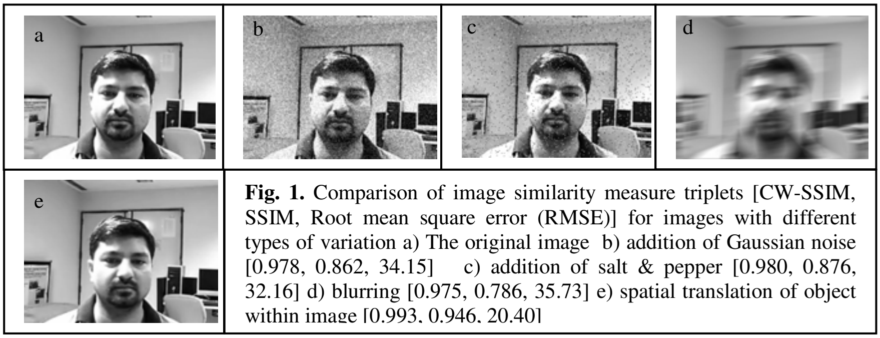


Fig. 1. Comparison of image similarity measure triplets [CW-SSIM, SSIM, Root mean square error (RMSE)] for images with different types of variation a) The original image b) addition of Gaussian noise [0.978, 0.862, 34.15] c) addition of salt & pepper [0.980, 0.876, 32.16] d) blurring [0.975, 0.786, 35.73] e) spatial translation of object within image [0.993, 0.946, 20.40]

It can be seen from fig. 1 that despite the images having different kind of variations and noise, a high value of similarity using CW-SSIM index (min. 0.975, max. 0.993) is obtained compared to SSIM (min. 0.786, max. 0.946) and RMSE(min 20.40, max. 35.73). This demonstrates the effectiveness of CW-SSIM index as an image similarity measure.

Table 1. The proposed algorithm

```
Initialisation
 1. Select object of interest as reference object ($O_{ref}$)
        $O_{ref} = \{CxWT_{i,j}\} i = 1...l, j = 1...w$
Prediction Step
 2. Predict the position of the object in next frame

Target Localization
 3. Considering the predicted position, define a search
    space having size ±p points larger than $O_{ref}$
 4. Compute the feature vector and similarity to get the
    target position as,
        $C^{*}_{a,b} = \arg\max[S_{a,b}(O_{ref}, C_{a,b})]_{a,b=-p}^{a,b=+p}$
    Where
        • $C_{a,b}$ represents candidates in search space.
        • $S_{a,b}$ represents the CW-SSIM in complex wavelet
          domain as defined in eq.(2).
 5. Extract the next frame and Go to step 2.
```

5 Experimental Results and Discussion

To evaluate the performance of the proposed object tracking method, we have used two videos, where we found significant variations in the object's pose, illumination and background. The first video is recorded in our lab and the second video has been taken from the dataset S1 (camera 3) from IEEE PETS 2006 workshop.

5.1 Qualitative Evaluation

For a qualitative comparison of the results of the proposed complex wavelet method (CWT Tracker) with one of the popular algorithms in literature, the Mean Shift method (MS Tracker) [2], we present here some snapshots of the results of both methods on the video sequences (see fig. 2, 3). The results show that CWT Tracker provides good performance as compared to MS Tracker. Fig. 2 shows that the MS Tracker drifts from the actual location of object due to lighting variation in the background (fig 2(b,d)), due to camera motion and change in pose (fig 2(e,f)). MS Tracker has poor performance particularly in the second video, where it leaves the target after a few frames only as shown in fig. 3(c,d), due to same appearance of the background and the object. Our method drifts from the ground truth data as shown in fig. 3(e), only when there is drastic pose and illumination change.

5.2 Quantitative Evaluation

To evaluate the tracking accuracy quantitatively, we compare our results with the manually labeled "ground truth" data and compute the Euclidian distance between the centroid positions of the tracked object and the corresponding ground truth data. For

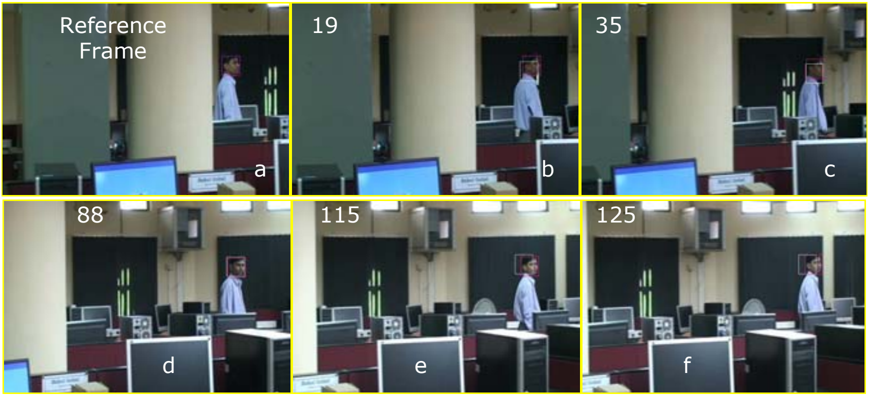


Fig. 2. A comparison of the proposed CWT tracker (indicated with magenta box) with MS tracker (represented by white box) on the video recorded in the lab

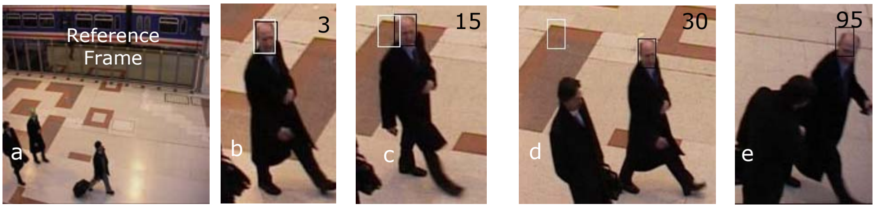


Fig. 3. A comparison of the proposed CWT tracker (indicated with black box) with MS tracker (represented by white box) on PETS 2006 video dataset S1 (camera 3)

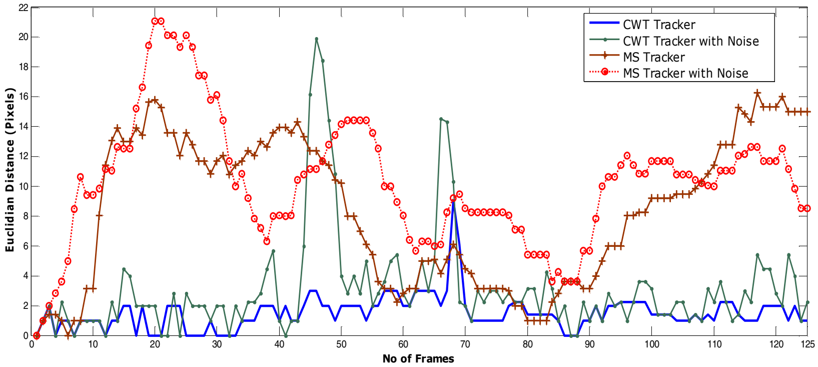


Fig. 4. The error the of object location using the Euclidian distance measure (in number of pixels) for CWT and MS Tracker without and with noise (Gaussian Noise with mean=0, variance=.1)

each frame of the first video, the error is plotted in fig. 4. For CWT Tracker the tracking position has very small deviation from the ground truth data, the average error in Euclidian distance is 1.52 pixels, while in the case of MS Tracker it is 8.44 pixels. From fig. 4 it is evident that the proposed method gives more accurate results even in the presence of noise (Gaussian Noise with mean=0, variance=.1).

The position, velocity and acceleration information from previous three frames are used to predict the location of the object in the current frame, which gives fairly accurate predictions even for complex motion.

6 Conclusion

Due to its approximate shift invariant nature Daubechies complex wavelet transform based method effectively tracks the object in varying real-life environment and even in the presence of noise. In this paper a new robust algorithm based on features and structural similarity is proposed for object tracking. The structural similarity in complex wavelet domain is exploited as similarity measure to make the tracker robust towards small luminance and contrast change and spatial translation. The experiments show that the average localization error for CWT Tracker is 1.5 pixels only. We can extend this work for multiple objects by introducing occlusion handling scheme.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object Tracking: A Survey. ACM Journal of Computing Surveys 38(4) (2006)
2. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based Object Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(5), 564–575 (2003)
3. Khansari, M., Rabiee, H.R., Asadi, M., Ghanbari, M.: Object Tracking in Crowded Video Scenes based on the Undecimated Wavelet Features and Texture Analysis. EURASIP Journal on Advances in Signal Processing, Article ID 243534 (2008)
4. Selsnick, I.W., Baraniuk, R.G., Kingsbury, N.: The Dual-Tree Complex Wavelet Transform. IEEE Signal Processing Magazine, 123–151 (November 2005)
5. Khare, A., Tiwary, U.S.: Symmetric Daubechies Complex Wavelet Transform and its Application to Denoising and Deblurring. WSEAS Transactions on Signal Processing 2(5), 738–745 (2006)
6. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image Quality Assessment: from Error Visibility to Structural Similarity. IEEE Trans. on Image Processing 13(4), 600–612 (2004)
7. Wang, Z., Simoncelli, E.P.: Translation Insensitive Image Similarity in Complex Wavelet Domain. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Philadelphia, PA, vol. 2, pp. 573–576 (2005)
8. Jalal, A.S., Tiwary, U.S.: A Robust Object Tracking Method for Noisy Video using Rough Entropy in Wavelet Domain. In: Proceedings of the International Conference Intelligent Human Computer Interaction, pp. 113–121. Springer, India (2009)

A Novel Multimodality Image Fusion Method Using Region Consistency Rule

Tanish Zaveri[1] and Mukesh Zaveri[2]

[1] EC Department, Nirma University, Ahmedabad, India
ztanish@nirmauni.ac.in
[2] Computer Engineering Department,
Sardar Vallabhbhai National Institute of Technology Surat, India
mazaveri@coed.svnit.ac.in

Abstract. This paper proposes an efficient region based image fusion scheme using discrete wavelet transform. This paper also proposes two new fusion rules namely mean, max and standard deviation (MMS) and region consistency rule. The proposed algorithm identifies the given images are multisensor or multifocus automatically. It allows best suitable algorithm for segmenting the input source images. Proposed method is applied on large number of registered images of various categories of multifocus and multimodality images and results are compared using standard reference based and nonreference based image fusion parameters. It is evident from simulation results of our proposed algorithm that it preserves more information compared to earlier reported pixel based and region based methods.

1 Introduction

In recent years, image fusion algorithms are used as effective tools in medical image processing, remote sensing, industrial automation, surveillance and defense applications. Due to these broad areas of applications, image fusion has emerged as a promising and important research area in recent years. The information coming from different sensors like optical cameras, millimeter wave cameras, Infrared cameras, x-ray cameras and radar cameras are required to fuse to increase amount of information in final fused image.

There are various techniques for image fusion at pixel level is available in literature [1] [2] [3]. In the recent literature [4], simulation results of region based image fusion method show better performance than pixel based image fusion method. The region based algorithm has many advantages over pixel base algorithm like it is less sensitive to noise, better contrast, less affected by misregistration but at the cost of complexity [3]. Recently researchers also recognized that it is more meaningful to combine objects or regions rather than pixels. Piella [3] also proposed a multiresolution region based fusion scheme using link pyramid approach. The proposed method provides automatic and powerful framework for region based image fusion method which produces good quality fused image for different categories of images.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 321–326, 2009.

2 Proposed Algorithm

Most image fusion methods are static and semiautomatic as they do not change and adapt different fusion rules automatically as kind of input source image changes so it is difficult to generate good quality fused image using single algorithm designed for one kind of images. The block diagram of proposed method is shown in Fig. 1. Image identification (Id) block shown in block diagram are used to identify the category of source images which may be multifocus and multimodality source images.

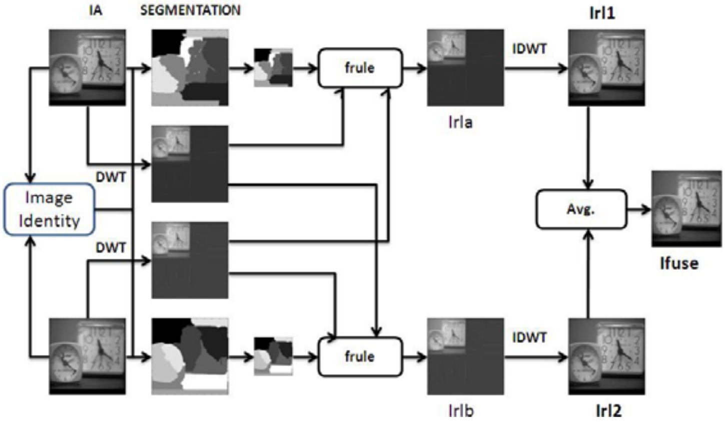


Fig. 1. Block Diagram of Proposed Method

Multifocus images are taken from the same camera but with the different focus points so it is expected that pixel intensity values between two source images do not change significantly. While in multisensor images are taken from the different sensor cameras so more difference in pixel intensity value will be expected between two source images which leads to more difference in value of MMS between two source images. The MMS value is significantly higher in case of multisensor images compared to multifocus images. The image identification (Id) is computed as described in equation (1).

$$Id = |Me1 - Me2| + |Std1 - Std2| + |MMS1 - MMS2| \tag{1}$$

Where Me1, Me2, Std1, Std2, MMS1 and MMS2 are mean, standard deviation and mean max and standard deviation parameter of source image IA and IB respectively. MMS parameter is described in detail later in this section and it is computed using eq. (3). If Id $\geq$ threshold than input source image is considered as multisensor source image. The details of threshold described in next section.

In the proposed algorithm DWT is used to decompose the input source images and inverse DWT is used to generate final fused image. The decomposed images arise from separable applications of vertical and horizontal filter. The resultant first level four image includes LL_1 sub band image corresponds to coarse level approximation image and other three image includes (LH_1, HL_1, HH_1) sub band images corresponds to finest scale wavelet coefficient detail images. Most image fusion method [1] based on DWT is applying max or average fusion rule on DWT decomposed approximate and detailed images to generate final fused image. This fusion rule based on DWT introduces undesired information in final fusion image. To remove this undesired information region consistency rule is used which described later in this section. The following the steps are used to generate final fused image after image identification step.

Step 1. The DWT is applied on image IA which gives first level decomposed image of one approximate image LL_{1A} and three detail images $(LH_{1A}, HL_{1A}, HH_{1A})$.

Step 2. If Id $\leq t_h$ than normalized cut segmentation algorithm is applied on image IA otherwise k-means algorithm is used as segmentation algorithm. Segmented image is then down sampled to match the size of DWT decomposed image.

Step 3. Then n numbers of segmented regions are extracted from approximate component of image IA and IB using segmented image. We have used two different fusion rules to compare extracted regions from different kind of source images. The SF is widely used in many literatures to measure the overall clarity of an image or region. The spatial frequency of that region is calculated using Row frequency (RF) and Column frequency (CF) as described in [4]. First fusion rule is region based spatial frequency (SF) rule as described in [4] is used to identify more informative region extracted from multifocus source images and image I_{rla1} is generated. SF of nth region of Image IA and IB is defined as SFAn and SFBn respectively.

$$I_{rla1} = \begin{cases} RA_{An} & \text{if } SF_{An} \geq SF_{Bn} \\ RA_{Bn} & \text{if } SF_{An} \leq SF_{Bn} \end{cases} \tag{2}$$

Here n is a number of regions and it varies from 1 to i. where $n = 1, 2, ..., i$. Regions extracted after applying normalized cut set segmentation algorithm on approximate image (LL_{1A}) are represented as RA_{An} and RA_{Bn} respectively. $I_{rla1}$is resultant fused image after applying fusion rule 1 as described in (3). This rule or any other existing fusion parameter is not enough to capture desired region so new mean max and standard deviation (MMS) rule is proposed in our algorithm. MMS is an effective fusion rule to capture desired information from multimodality images. This proposed fusion rule exploits standard deviation, max and mean value of images or regions. The MMS is described as

$$MMS_{An} = ME_{An}/SD_{An} * R_{Anmax} \tag{3}$$

Where ME, SD and R_{max} are mean, standard deviation and maximum intensity value of nth region of source image respectively. The MMS is computationally efficient and effective. From our study, it is analyzed that with visual images, SD is high and ME is low where in images captured using sensors like MMW and IR have ME value high and SD is low so in our algorithm we have used both SD and ME with maximum intensity value to derive new parameter MMS. From the experiments, it is observed that the low value of MMS is desired to capture critical regions especially man in this multisensor images. The fusion rule 2 is described as below

$$I_{ral1} = \begin{cases} RA_A n & \text{if } MMS_{An} \leq MMS_{Bn} \\ RA_B n & \text{if } MMS_{An} \geq MMS_{Bn} \end{cases} \tag{4}$$

Intermediate fused image I_{ral1} is generated by fusion rule 2 which is applied for multimodality images and first fusion rule is applied for multifocus images. After taking approximate component by above method, region consistency rule is applied to select detail component from both decomposed images, which is described in (5). Region consistency rule states that analyze the results of fusion rule 1 or 2 of approximate component and select corresponding nth region of detail component as described below.

$$I_{rlv1} = \begin{cases} RD_{An} & \text{if } I_{rla1} = RA_{An} \\ RD_{Bn} & \text{if } I_{rla1} = RA_{Bn} \end{cases} \tag{5}$$

Where I_{rlv1} is first region of vertical component of image HL_{1A}. Similarly I_{rlh1} and I_{rld1} is computed from image LH_{1A} and HH_{1A} respectively. After applying fusion rule 1, approximate component of nth region is selected from image IA than corresponding detail component nth regions also selected from the same image IA to generate final fused image.

Step 4. Then IDWT is performed to generate Irl1.

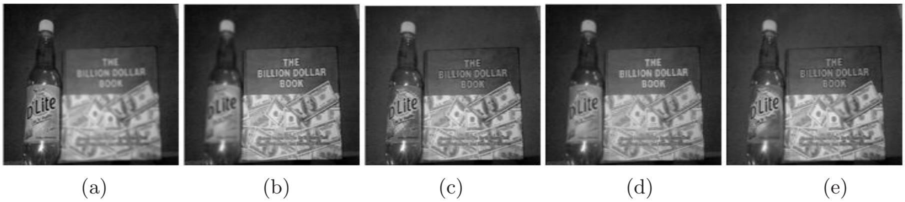
Step 5. Repeat the step 1 to 4 for image IB and generate intermediate fused image Irl2.

Step 6. Both Irl1 and Irl2 are averaged to improve the resultant fused image IFUSE.

These new frame work is an efficient way to improve the consistency in final fused image and it avoids distortion due to unwanted information added without using region consistency rule. In the next section image fusion evaluation criteria is described in brief.

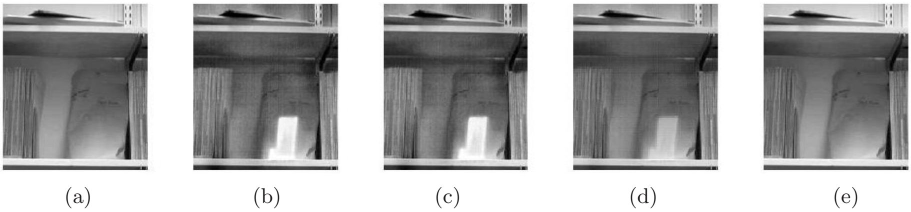
3 Simulation Results

Image fusion performance evaluation parameters are divided into two categories reference based and non reference based which is described in [7].The proposed algorithm has been implemented using Matlab 7. The proposed algorithm applied on large number of dataset images which contain broad range of multifocus and multimodality images of various categories like multifocus with only object, object plus text, only text images and multisensor IR (Infrared) images. The simulation results are shown in Fig. 2 and 3. Threshold value of image identification (Id) of source image is considered as 10. This value is used to differentiate between category of multimodality or multisensor image. This value is decided



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

Fig. 2. Fusion Results of Multifocus Book Image(a),(b) Source Images (c)Proposed Method(d) Li's Method [4] (e) DWT based method [1]



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

Fig. 3. Fusion Results of Multisensor Gun Image (a) Visual Image (b) IR image(c) Proposed Method (d) Li's Method [4] (e) DWT based method [1]

Table 1. Image Fusion Parameters for Multifocus Images

Image	Fusion Methods	Fusion Parameters			
		SF	MIr	RMSE	SSIM
Text Image	DWT Based [1]	8.1956	1.5819	6.3682	0.9259
	Li's Method [4]	10.405	1.4713	5.2671	0.97499
	Proposed Method	10.736	1.9997	2.8313	0.9904
Book Image	DWT Based [1]	12.865	3.809	7.599	0.927
	Li's Method [4]	16.416	3.637	5.084	0.978
	Proposed Method	17.415	6.517	1.612	0.997

Table 2. Image Fusion Parameters for Multisensor Images

Image	Fusion Methods	Entropy	MI
IR Image	DWT Based [1]	6.654	1.329
	Li's Method [4]	6.740	4.782
	Proposed Method	6.781	4.812
Book Image	DWT Based [1]	7.412	4.569
	Li's Method [4]	7.372	7.152
	Proposed Method	7.535	7.176

after many experiments on different category of images. Proposed algorithm is applied on various categories of images for different segmentation levels and after analyzing those results, we have considered nine segmentation levels for all our experiments which improve visual quality of final fused image. The performance of proposed algorithm evaluated using standard reference based and nonreference based image fusion evaluation parameters which are depicted in Table 1 and Table 2. The visual quality of the resultant fused image of proposed algorithm is better than the fused image obtained by other compared methods.

4 Conclusion

In this paper, new automatic DWT and region based image fusion method using region consistency rule is implemented. The proposed algorithm identifies the type of images that is a given set of images for fusion is multisensor or multifocus images. The proposed algorithm is applied on large number of dataset of various categories of multifocus and multisensor images. It has been observed that visual quality of proposed algorithm is better compared to other earlier reported pixel and region based image fusion method. The novel MMS fusion rule is introduced to select desired regions from multimodality images. Proposed algorithm also compared with standard reference based and nonreference based image fusion parameters and from simulation and results, it is found that visual quality and assessment parameters are better than other earlier reported methods.

References

[1] Anna, W., Jaijining, S., Yueyang, G.: The application to wavelet transform to multimodality medical image fusion. In: IEEE International Conference on Networking, Sensing and Control, pp. 270–274 (2006)

[2] Miao, Q., Wang, B.: A novel image fusion method using contourlet transform. In: International Conference on Communications, Circuits and Systems Proceedings, vol. 1, pp. 548–552 (2006)

[3] Piella, G.: A general framework for multiresolution image fusion: from pixels to regions. Journal of Information Fusion 4(4), 259–280 (2003)

[4] Shutao, L., Bin, Y.: Multifocus image fusion using region segmentation and spatial frequency. Image and Vision Computing 26, 971–979 (2008)

[5] Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)

[6] Zheng, L., Robert, L.: On the use of phase congruency to evaluate image similarity. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, vol. 2, pp. 937–940 (2006)

[7] Tanish, Z., Mukesh, Z.: Region Based Image Fusion for detection of Ewing Sarcoma. In: Seventh International conference on Advances in Pattern Recognition, pp. 240–242 (2009)

Deriving Sparse Coefficients of Wavelet Pyramid Taking Clues from Hough Transform

Jignesh K. Bhavsar and Suman K. Mitra

Dhirubhai Ambani Institute of Information and Communication Technology,
Gandhinagar, India
{jignesh_bhavsar,suman_mitra}@daiict.ac.in

Abstract. Many applications like image compression requires sparse representation of image. To represent the image by sparse coefficients of transform is an old age technique. Still research is going on for better sparse representation of an image. A very recent technique is based on learning the basis for getting sparse coefficients. But learned basis are not guaranteed to span l_2 space, which is required for reconstruction. In this paper we are presenting a new technique to choose steerable basis of wavelet pyramid which gives sparse coefficients and better reconstruction. Here selection of steerable basis is based on clues from Hough transform.

1 Introduction

A sparse signal has a large portion of its energy contained in a small number of coefficients. One obvious reason why we desire to have sparse representations of signals is, more sparsely one can represent a signal, more efficiently one can compress the signal. Moreover, sparse representations allow us to do much more than just compression. For example, sparse representations allow us to violate the Nyquist sampling theorem without loss of fidelity [1]. Sparse representations leads to solutions to previously insolvable blind source separation problems [2]. It has applications including speech processing, biomedical signal processing, financial time-series analysis, and communications [3]. Fourier Transform,Discrete Cosine Transform,Wavelet Transform are a few examples of choosing suitable basis for sparse transformation. This motivates researchers to explore new technique for finding suitable basis that leads to maximum sparsity as well as information.

Recently, machine learning based technique for finding suitable sparse basis coefficients is gaining attention [4]. Here one takes Cauchy distribution of the coefficients as a prior, then learn image basis using MAP with minimum residual error as constraint. But learned basis through this mechanism do not guarantee about spanning of l_2 space. Note that, without spanning of l_2 space, the original input signal could not be reconstructed from the basis. Reconstruction drawback of learned basis motives the current research. We are proposing new basis which are tight frames and derived without any learning. In the proposal, steerable pyramid wavelet transform is used. It is noted that energy distribution among

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 327–332, 2009.

the basis is depending on features present in the image. One of these features is orientation of edges. Hough transform provides us inference about orientation of edges. Using this inference we choose suitable steerable basis such that image energy is mostly confined in one of the basis. This leads to coefficients of basis close to zero. The major contribution of the present submission includes the following,

1. A new set of steerable basis leads to sparse representation of the image.
2. The basis are constructed based on edge orientation present in the image and hence it is adaptive in the sense of image requirement.
3. The reconstruction of input image is guaranteed.
4. The basis selection is cost effective as it requires no learning.

Note that for rest of the manuscript the steerable basis when implemented in the images will be termed as steerable band pass filters.

The organization of the article is as follows. In Section 2, we have discussed learning based method [4]. Section 3 deals with the current proposal followed by implementation and results in Section 4.

2 Learning Based Method for Sparse Representation

The detail description of learning based sparse representation is available in [4]. Only the salient features of the same is discussed here to make the readers appreciate the current proposal which does not require any learning. With a particular choice of basis coefficients a, one can express image $I(\boldsymbol{x})$ with an assumption that the noise present is Guassian,white and additive.

$$I\left(\boldsymbol{x}\right) = \sum_{i} a_i \phi_i\left(\boldsymbol{x}\right) + v\left(\boldsymbol{x}\right) \tag{1}$$

Where $\boldsymbol{x}$ denotes a discrete spatial position, ϕ_i are basis functions, v is noise or uncertainty. Hence for given set of a_i's, we need to find ϕ_i so that the noise present is as less as possible. ϕ_i's are learned from the model given the data.

Image I as presented in equation (1) can be represented canonically as,

$$I = Ga + \nu \tag{2}$$

where vector a is coefficient for all scales, positions and levels, G is basis functions for coefficient vector a. Probability of generating Image I, given coefficients a is assumed to be Gaussian,

Under this set up, and assuming a Cauchy prior coefficients of a given image are determined by MAP estimate, The learning time is often high and this leads to a disadvantage of this mechanism . Another disadvantage may come from the fact that these basis are statistically independent, however not guaranteed to span l_2 which is required for reconstruction.

All learned basis are directional band pass filters which lead to a new direction. Taking clue from this, we propose a new approach to construct basis inferred from Hough Transform. This approach removes learning part of the basis.

3 Proposed Methodology Inferred from Hough Transform

Here we are proposing a new approach to get sparse distribution of coefficients by suitably selecting basis. We have used pyramid wavelet transform with basis as steerable functions. As discussed in [5], each steerable basis functions are rotated copy of each other. We can design basis of any orientation direction by interpolating these basis functions. Here we have used steerable filters or basis and wavelet pyramid structure .

General equation for pyramid structure as described in [6] is,

$$\hat{X}\left(\overrightarrow{w}\right) = \left\{ H_0\left(\overrightarrow{w}\right)|^2 + |L_0\left(\overrightarrow{w}\right)|^2 \left(|L_1\left(\overrightarrow{w}\right)|^2 + \sum^{n} |B_k\left(\overrightarrow{w}\right)|^2 \right) \right\} X\left(\overrightarrow{w}\right)$$
$$+ \text{ aliasing terms} \qquad (3)$$

Where X is input signal, $\hat{X}$ is output of the same when treated with filters, H_0 is first level high pass filter and L_0 is first level low pass filter. After that, low pass filter L_1 and steerable band pass filters $B_i's$ are repeated recursively on 2 by 2 decimation. The reconstruction is possible if the following constraints such as unitary response, recursion relation and aliasing cancellation. There is one more constraint that B_i's need to satisfy is angular constraint.

The spectrum of wavelet pyramid structure which is divided into three types output of High Pass,Low Pass and Band Pass filters. Note that the band pass filters are directional oriented and are called basis. These filters are steerable. The main emphasis of the current work to select proper band pass filter which generate coefficients as sparse as possible. The direction orientations are chosen from the Hough Space [7] generated out of Hough Transformation. Hough transform is a mapping of image into Hough space. As described in [7], Hough space is made up of two parameters, distance from origin (ρ) and and angle with x-axis (θ) (positive in anticlockwise direction). Any point in image plane (x, y) can be represented as following in the Hough Space ,

$$\rho = x\cos(\theta) + y\sin(\theta) \qquad (4)$$

Note that infinite numbers of lines can pass through a single point of image plane. All lines passing through the point (x, y) should satisfy equation (4). Again for each point (x, y) in the image plane, a sinusoidal is generated in Hough space by varying ρ and θ. The lines passing through this point (x, y) in image plane increases co-occurrence of sinusoidal waves in Hough space. So, each line generates peak in hough space at a particular (ρ, θ). In other words each line in image can be represented as unique couple (ρ, θ) for $\theta \epsilon [0, \pi]$ and $\rho \epsilon R$ in Hough space. Finally, we conclude that hough transform is line to point conversion. Natural images have structures of edges and lines. If band pass filters (steerable basis) are getting tuned with lines, one can get sparse distribution for coefficients of band pass filters. We propose to tune orientation of steerable basis taking clue from Hough transform. Peaks in Hough space correspond to the lines or edges in image. Note that peaks at 0 and ±90 degrees are not considered as these angles do not bring any information. First, find the highest peak in a particular angle in Hough space. This angle indicates orientation direction of the most of the edges or the lines in the image. Hence this angle could be used to select the direction of steerable band pass filters. Finally, if required, from the Hough space some angles (few top peaks) that indicates orientation direction of edges present in the image are selected and used to choose the direction of steerable band pass filters.

In steerable pyramid structure, we choose first steerable basis in tune with the highest peak angle. This implies first steerable basis which is direction derivative function, has direction perpendicular to the orientation of highest peak angle in Hough space. First basis(band pass filter), thus contains maximum energy. So, other basis functions will have comparatively less energy leading toward sparse distribution of coefficients.

4 Exprimetal Results

Experiments are conducted for a set of natural images taken from [8]. Note that same set of images are used for learning the basis as presented in [4].

Figure 2 shows trained basis as described in Section 2, These are learned for image shown in Figure 1. Starting with real random basis, tunning has been carried out through training process (Section 2) for one and half hour in a pentium-IV,1.7 GHz CPU machine. Note that tunning of basis may improve if

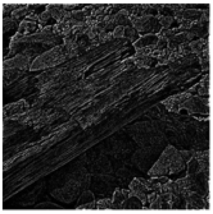


Fig. 1. Original Natural Image

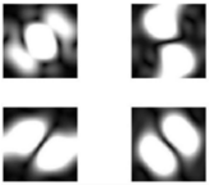


Fig. 2. Trained basis filters starting from real random basis. Training is carried out for one and half hour.

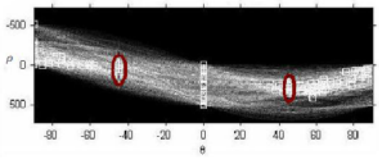


Fig. 3. Hough transform of the image shown in Figure 1. The circle indicates the maxima at Hough space.

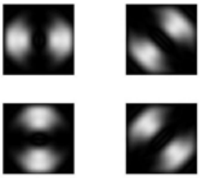


Fig. 4. Directional band pass filter whose orientation direction is inferred from Hough transform (Figure 3).

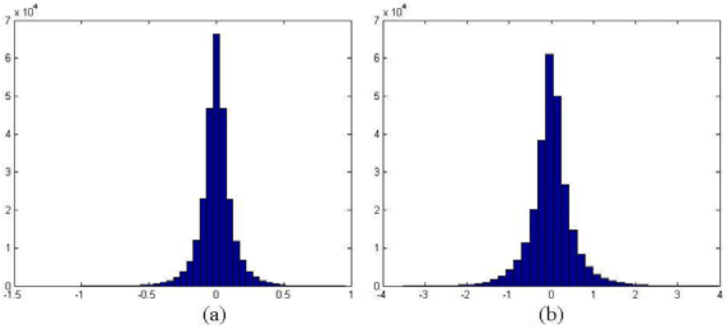


Fig. 5. Histogram of one level coefficients: (a) tunned band pass filters, (b) trained filters inferred from Hough transform

Table 1. Comparison of Filter Coefficients Distribution

	Trained Filter	Inferred Band Pass Filter
Mean	2.4835e-006	-1.3159e-018
Variance	0.2568	0.0147

Table 2. Comparison of quality measures for reconstruction with steerable basis (Inferred Band Pass Filter) and trained basis (filter)

	MSE	SNR	PSNR
Inferred Band Pass Filter	0.0546	2.5706	23.6916
Trained Filter	0.2895	-4.6715	16.4495

one extends the time of learning. However we felt that would be too costly as far as time is concern. So, we stop after one and half hour of learning. Trained basis (Figure2) are like directional derivatives where directions are perpendicular to orientation corresponding to peak angles of hough space as shown in Figure 3. This justifies our claim that orientation direction of steerable basis could be inferred from Hough Transform. At the same time these basis will generate sparse distribution of coefficients. Proposed directional band pass filters obtain from Hough are shown in Figure 4. The sparsity of proposed basis filters is compared with that of trained basis filters. The histogram of one level coefficients of the proposed band pass filter inferred from Hough transform is presented in Figure 5 (a) , where as same for trained basis is presented in Figure 5 (b) . The sparsity of the former seems to be efficient for compression. The mean and variance values of both histograms are presented in Table 1. Variance of coefficients of the proposed filters is significantly less that of trained filters. In Table 2, we have compared Quality Measures of reconstructed images with steerable basis and learned basis. This shows steerable basis has good reconstruction capability with sparse coefficients.

5 Conclusion

This article poses a new approach to make wavelet pyramid steerable basis coefficients as sparse as possible. The orientation directions required to select steerable band pass filters have been identified from the Hough transform and thus required no learning, yet it is appeared to be efficient for sparsity and reconstruction. Unlike same basis applied to a set of images, the basis proposed here seems to be tailor made for a given image. Though the current sparse basis that is learned [4], yet it is worth mentioning that current work of finding orientation direction for band pass filters has been assured by the learning based technique. Instead of learning the orientation directions, the same is inferred from Hough space generated out of Hough transform. Reconstruction power that is guaranteed by steerable pyramid structure, accuracy in sparsity and the cost effectiveness are major contribution of the present work.

References

1. Tropp, J.A., Laska, J.N., Duarte, M.F., Romberg, J.K., Baraniuk, R.G.: Beyond nyquist: Efficient sampling of sparse bandlimited signals. CoRR (2009)
2. Li, Y., Cichocki, A., ichi Amari, S., Shishkin, S., Cao, J., Gu, F., Cao, J., Gu, F.: Sparse representation and its applications in blind source separation. In: Seventeenth Annual Conference on Neural Information Processing Systems, NIPS 2003 (2003)
3. http://sparse.ucd.ie
4. Sallee, P., Olshausen, B.A., Lewicki, M.S.: Learning sparse image codes using a wavelet pyramid architecture 13, 887–893 (2001)
5. Freeman, W.T., Edward, H.A.Y.: The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 891–906 (1991)
6. Karasaridis, A., Simoncelli, E.: A filter design technique for steerable pyramid image transforms, pp. 2387–2390 (1996)
7. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. Commun. ACM 15, 11–15 (1972)
8. https://redwood.berkeley.edu/bruno/sparsepyr

Improvised Filter Design for Depth Estimation from Single Monocular Images

Aniruddha Das, Vikas Ramnani, Jignesh Bhavsar, and Suman K. Mitra

Dhirubhai Ambani Institute of Information and Communication Technology,
Gandhinagar, India
{aniruddha_das,vikas_ramnani,jignesh_bhavsar,suman_mitra}@daiict.ac.in

Abstract. Depth Estimation poses various challenges and has wide range applications. Techniques for depth prediction from static images include binocular vision, focus-defocus, stereo vision and single monocular images unfortunately not much attention has been paid on depth estimation from single image except [1]. We have proposed a method for depth estimation from single monocular images which is based on filters that are used to extract key image features. The filters used have been applied at multiple scales to take into account local and global features. Here an attempt is made to reduce the dimension of feature vector as proposed in [1]. In this paper we have optimized the filters used for texture gradient extraction. This paper also introduces a prediction algorithm whose parameters are learned by repeated correction. The new methodology proposed provides an equivalent quality of result as in [1].

1 Introduction

Emerging interest in self driven cars, boats and ships [7] has made it necessary for continuous and fast real time depth perception algorithms. Depth Estimation has very wide applications in robotics [3], scene understanding and 3-D reconstruction [9].

Collaborative Ocular Melanoma Study (COMS) [8] has defined depth perception as the power to discern three-dimensional objects and their relative position in space. Visual perception can be classified into Monocular Vision and Binocular Vision. Binocular Vision uses a pair of images from different angle of the same object to generate depth maps implemented in [6]. This paper deals with depth estimation in monocular vision, that uses a single image.

There are two methods for retrieving depths from monocular digital images. The first method includes using a sequence of closely related stream of images with prior information of the scenes or video for depth estimation. Second method is an algorithm for depth prediction from single independent monocular images. A novel algorithm for depth estimation from images using focus and defocus has been presented by Das and Ahuja [5].

Depth estimation from single monocular images can be implemented using stereo features of images and monocular features of images. Stereo Vision has also been used to generate depth maps from images [10]. It has shown excellent

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 333–338, 2009.

results on the Middlebury stereo evaluation test bed. Saxena et al. [1] presents a discriminately trained MRF for depth estimation from single monocular images. The algorithm presented hereby uses features extracted from single monocular images correlated with depth map to train a machine for depth estimation. The algorithm identifies the key steps involved in estimating depth from discreet and uncorrelated images. Saxena et. al [1] implements a set of 17 filters for computing the feature vector. The feature vector is defined on sub images which are called patches. The feature vector is then used in the formation of a machine learning model to predict depth. They have effectively reduced blocking effect by taking into consideration neighboring patches.

In this paper we have introduced a new learning methodology, after optimizing the feature selection criteria. Section 2 describes feature selection, extraction, learning approach and benefits of new methodology used for depth estimation. Section 3 depicts the comparison of new methodology and Saxena et al. [1].

2 Proposed Methodology

Here we are proposing a procedure for depth estimation of outdoor scenes. Note that exactly similar procedure may not be applicable to indoor scenes. The system of depth estimation from single monocular images consists of three basic steps. Initially, a set of images and there corresponding depth maps are gathered. Then a set of suitable features are extracted from the images. Based on the features and the ground truth depth maps a prediction model is learnt. The depth of new images are predicted from the learnt model.

It has been observed that the changes in depths along the row of any image compared to the same along the column is comparatively less. Along the row the depth of the scene observed by camera eye would depend on the angle of camera. On other hand the depth along the column till infinity since the outdoor scene is unbounded due to the presence of sky. Moreover, depths are estimated for small patches within the image. Hence, the feature selection has been done for each patch of the image. On the other hand learning of depth map has been done for a row of patches instead of a single patch. Note that similar argument has been used in [1]. Next, the feature selection and learning methodology are discussed. A set of filters are designed to extract features. In particular these filters are expected to extract texture energy, texture gradient and haze. These cues are used because texture energy of objects varies according to distances from the viewers, texture gradient captures the distribution of the direction of edges and haze is used to detect atmospheric light scattering [1]. These three cues help to indicate depths.

Figure 1 indicates a total of 16 such filters used by Saxena et al. [1]. The filters used in [1] for texture gradient are mainly edge detection filters. The edge detection filters are primarily directional oriented high pass filters. In this paper, a single Omni- directional high pass filter, in place, of the six directional texture gradient filters [1] is used to compute the texture gradient information. The use of an Omni- directional high pass filter reduces the net size of the feature vector.

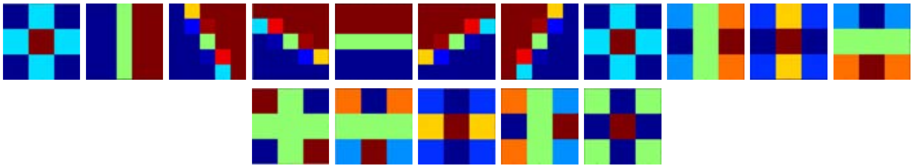


Fig. 1. A set of 16 filters as implemented by Saxena et al.[1]. (The first low pass filter is used twice on two different color channels results in a total of 17 features.)

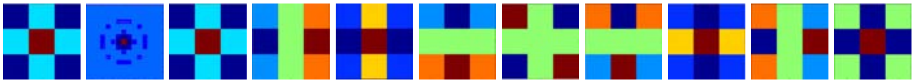


Fig. 2. The set of new filter used for the current experiment. Filter 2 is the new omni-directional high pass filter changed. (The first low pass filter is used twice on two different color channels which gives a total of 12 features.)

The Omni-directional high pass filter has been shown in filter 2 of Figure 2. Other filters used are same as in [1].

In the prior work of Saxena et al.[1] the prediction model assumed to be linear with error distributed as Gaussian and Laplacian. However, no justification of the same has been reasoned out. In our experimentation it has been observed that the relation between the ground truth depth and its prediction could be non linear and hence we are proposing a linear model for depth estimation with non linear error. The model used is

$$d = \hat{d} + \eta \tag{1}$$

$$\hat{d} = \underline{W}^T \ \underline{f} \tag{2}$$

where $\hat{d}$ denotes the predicted depth, d is the ground truth depth, $\underline{W}^T$ denotes the weight estimated to compute the predicted depth and $\underline{f}$ denotes the feature vector obtained. Here η is the error of the model. Our task is to learn the weight vector $\underline{W}$ from the given data. The weight vector $\underline{W}$ is learned for all patches in a row taking information about the depth of same patches of row coming from different images. The weight vector $\underline{W}$ mentioned in the model is learned in an iterative process such that the error converges to zero.($\eta \rightarrow 0$)

Let us assume that there are l patches in a row and there are n number of images in the data base. So we are getting $N = nl$ number of ground truth depth maps for learning $\underline{W}$. Using the data in the model we get :

$$\underline{\hat{d}} = \underline{W}^T \ \mathbf{F}, \quad \text{where} \qquad \underline{\hat{d}} = \left(\hat{d}_1 \ \ \hat{d}_2 \ \ \hat{d}_3 \ldots \hat{d}_N\right)^T$$

$$\text{and} \quad \mathbf{F} = \left(\underline{f}_1 \ \ \underline{f}_2 \ \ \underline{f}_3 \cdots \underline{f}_N\right)$$

We start with an initial choice of $\underline{W}^T$ as $\underline{W}_0^T = \underline{d} \ \mathbf{F}^{-1}$, Where $\underline{d}$ is the ground truth depth vector form for all patches in a row and from all images. If $\mathbf{F}$ is not a square matrix than we can use generalized inverse of $\mathbf{F}$.

Taking this initial choice of $\underline{W}_0^T$ and using features of each patch in a row of each image we estimate the depth by equation(2). Let it be $\hat{d}_i$, i=1,2 …N. These estimates of depth are then subjected to correction to reduce the error. A non linear function, in particular, a p^{th} degree polynomial, is fitted using the least square technique on the estimated depth $\hat{d}_i, i = 1, 2 \dots N$ to get corrected estimate $\hat{\hat{d}}_i, i = 1, 2 \dots N$. Hence, the new choice of $W^T$is obtained as:

$$\underline{W}_{new}^T = \hat{\hat{\underline{d}}}\, F^{-1}$$

where, $\hat{\hat{\underline{d}}} = \left(\hat{\hat{d}}_1 \;\; \hat{\hat{d}}_2 \;\; \hat{\hat{d}}_3 \dots \hat{\hat{d}}_N\right)^T$ and F is as defined earlier

This completes one iteration. In the next iteration, we will start with a choice of $\underline{W}^T$ as $\underline{W}_{new}^T$. We stop modification $\underline{W}^T$ if there is no significant.

The same learning procedure is adapted for predicting the absolute depth as well as the relative depth and hence the final depth is obtained. Here relative depth is the predicted difference of depths from the neighbors for a given patch.

3 Implementation and Results

The above mentioned methodology has been implemented on a large data set and the result obtained is comparable with the result presented in [1]. A large data base of images along with their ground truth depth is available in [4]. For our experiments we have used 100 images of size 2272X1704 of outdoor scenes and there depth maps of size 51x78. 80 images from the above set were used for training and the rest 20 images were used for testing. The three monocular cues haze, texture edges and texture energy are extracted using the 17 Laws' masks [1] applied at three scales (one, one-third and one-ninth of the image). The feature vector for each patch is calculated for L1 and L2 norm using its 4 neighbors at 3 scales and the column feature vector for 4 patches. This results in a feature vector of size 646 ((((5x3+4) x17=323) x2) =646).

The texture gradient filters used in Saxena et al. [1] are high pass filters oriented at 0, 30, 60, 90, 120 and 150 degrees. In the new methodology, an Omni directional high pass filter is used for extracting texture gradient. Resulting in a reduction of the size of the image feature vector from 646 to 456((((5x3+4) x12=238) x2) =456). The rest of the filters are same as used in Saxena et al. [1].

The resultant feature vector and the corresponding depth maps in the database were used to find the parameters of the model and the nonlinear function as described in Section 2. For the current implementation a polynomial of degree four is used which is obtained experimentally. So far we do not have any mathematical justification for the same. Note that the four iteration, as described in Section 2 appeared to be sufficient to estimate the model parameters.

The model of predicting relative depth is similar as of absolute depth. In absolute learning part learning was done by absolute value of depths for the patch and its neighboring patches. The difference of histograms of feature vectors are used as parameters to predict the difference of depth with neighbors of a

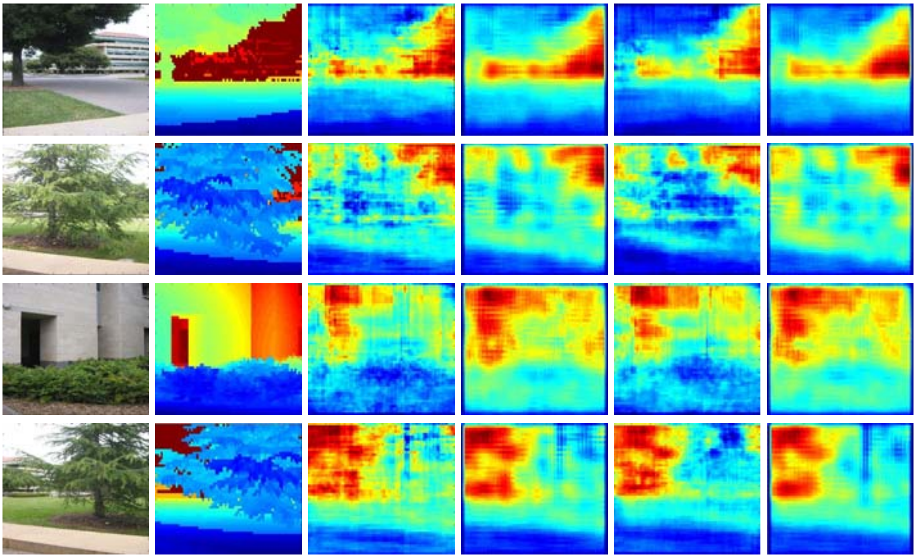


Fig. 3. A few test images along with their depth map predicted. Column 1 Depicts actual images. Column 2 Depicts actual ground truth depth maps. Column 3 Depicts absolute depth maps resulted from Saxena et al.[1]. Column 4 Depicts relative depth maps resulted from Saxena et al.[1]. Column 5 Depicts absolute depth maps predicted using new methodology. Column 6 Depicts relative depth maps predicted using new methodology.

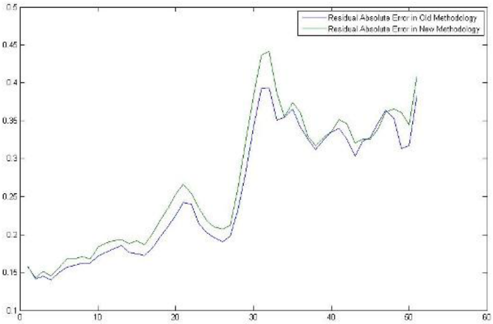


Fig. 4. Comparison of the absolute error in Saxena et al. [1] and after implementation of the new methodology.

given patch. Depth predicted for some of the test images are shown in Figure 3. Column 1 and 2 in Figure 3 shows the actual image and the ground truth depth maps respectively. The predicted depth maps generated by Saxena et al. [1] (column 3& 4) and by the proposed methodology (column 5 & 6) are also shown. Column 3 and 5 in Figure 3 corresponds to absolute depth maps and

Column 4 and 6 in Figure 3 shows the depth maps after taking into account relative features.

The error comparison of the absolute depth prediction of Saxena et. al [1] and the absolute depth prediction after the implementation of the final filter is shown in the graph Figure 4. The change in edge detection filters reduces the sharp peaks in the relative difference prediction.

4 Discussions and Conclusion

Omni Directional Filters proposed here appear to be a promising filters. The result obtained is comparable with same presented in [1]. Moreover, the computational cost is reduced drastically by reducing the feature dimension. The 6 edge detection filters in [1] are replaced by a single Omni-directional high pass filter which reduce the size of feature vector, without significant change in the performance output. The time complexity of the model is evaluated to be $O(k.n)$, where n is the number of features in the feature vector and k is the number of iterations to adjust the parameters. The space complexity of the algorithm also reduces as the number of features in the feature vector decrease. The space complexity is evaluated to be $O(k.n)$, where k is the number of iterations performed to determine the parameters of the learning model and n the number of features in the feature vector.

References

1. Saxena, A., Chung, S.H., Ng, A.Y.: Learning Depth from Single Monocular Images. In: NIPS (2005)
2. Saxena, A., Schulte, J., Ng, A.Y.: Learning Depth from Images using Monocular cues and stereo vision. In: ICJAI (2007)
3. Michels, J., Saxena, A., Ng, A.Y.: High speed obstacle avoidance using monocular vision and reinforcement learning. In: ICML (2005)
4. Saxena, A.: Stanford Range Image Data, http://ai.stanford.edu/~asaxena/learningdepth/data.html (last viewed on January 20, 2009)
5. Das, S., Ahuja, N.: Performance analysis of stereo, vergence, and focus as depth cues for active vision. IEEE Trans. Pattern Analysis & Machine Intelligence 17, 1213–1219 (1995)
6. Lungeanu, D., Popa, C., Hotca, S., Macovievici, G.: Modeling biological depth perception in binocular vision: The local disparity estimation. Informatics for Health and Social Care 23(2), 131–143 (1998)
7. Van Amerongen, J., Ten Udink, C.A.J.: Model Reference Adaptive Autopilots for Ships. Automatica 11, 441–449
8. http://www.jhu.edu/wctb/coms/booklet/book5.htm (Visited on 19/4/2009 at 7:31 PM)
9. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int'l. Journal of Computer Vision 47, 7–42 (2002)
10. Andreas, K., Mario, S., Konrad, K.: Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In: ICPR (2006)

Color Image Segmentation Based on Vectorial Multiscale Diffusion with Inter-scale Linking

V.B. Surya Prasath

Department of Mathematics, Indian Institute of Technology Madras, India
surya.iit@gmail.com

Abstract. We propose a segmentation scheme for digital color images using vectorial multiscale anisotropic diffusion. By integrating the edge information, diffusion based schemes can remove noise effectively and create fine to coarse set of images known as scale-space. Segmentation is performed by effectively tracking edges in an inter-scale manner across this scale space family of images. The regions are connected according to color coherency, and scale relation along time axis of the family is taken into account for the final segmentation result. Fast total variation diffusion and anisotropic diffusion facilitate denoising and create homogenous regions separated by strong edges. They provide a roadmap for further segmentation with persistent edges and flat regions. The scheme is general in the sense that other anisotropic diffusion schemes can be incorporated depending upon the requirement. Numerical simulations show the advantage of the proposed scheme on noisy color images.

Keywords: Image Segmentation, Multiscale diffusion, Nonlinear scale space, Color images.

1 Introduction

Digital image segmentation is the task of separating and identifying important objects in a scene. Segmentation techniques can be broadly classified into two categories, namely, edge based and region based approaches. Edge based segmentation methods [5,12] make use of the probable edges found as a roadmap in driving the segmenting contours. Variational minimization methods [9] and partial differential equations (PDEs) [1] based segmentation schemes are very popular in this regard. Digital images usually consists of objects of different shapes and scales. Capturing the correct local scale is an important problem in edge based image processing schemes. Thus if the objects are not differentiated with respect to their corresponding scale then the segmentation can give false results. The random noise makes it difficult to calculate the exact scale of objects, and this, in turn, gives over-segmentation or under-segmentation. In this paper we propose to tackle these problems by considering the scale space approach based on an edge preserving multiscale diffusion scheme.

Starting from the linear scale space studied in image denoising [7,8] nonlinear diffusion PDEs [10,1,11] are widely used in early computer vision problems. By

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 339–344, 2009.

using such scale space generated by them and pruning it according to inter-scale relations with region linking gives an effective segmentation of the image at the coarse scale. The edges are preserved through the scale space due to the use of multiscale diffusion and depending on the choice of scale, meaningful segmentations can be obtained. The proposed scheme is a general one with possible modifications can be made at the scale space generation step. The rest of the paper is organized as follows. Section 2 introduces the PDE based segmentation scheme. Section 3 presents numerical results which illustrate advantages of the proposed scheme on noisy images. Finally, section 4 concludes the paper.

2 Diffusion Based Segmentation

2.1 Anisotropic Diffusion

To denoise a noisy image u_0 Perona and Malik proposed the following anisotropic diffusion scheme (ADS) [10]

$$\frac{\partial u(x,t)}{\partial t} = div\left(g(|\nabla u(x,t)|)\nabla u(x,t)\right) \quad \text{with} \quad u(x,0) = u_0 \tag{1}$$

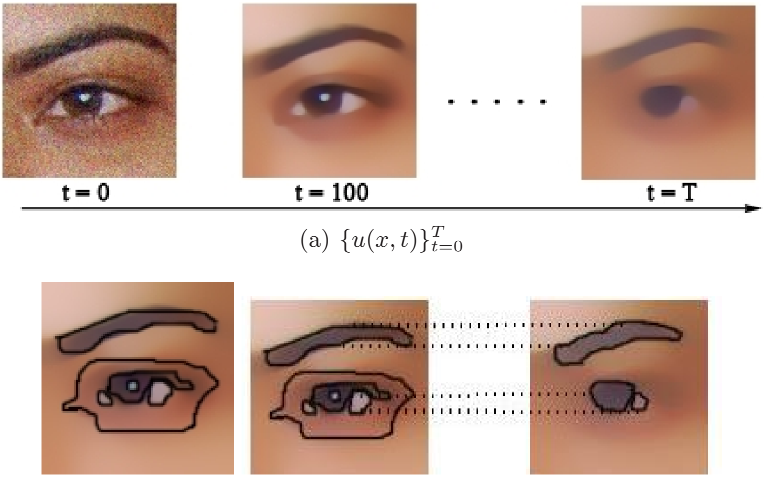
where g is a decreasing function, for example $g(s) = \exp -(s^2/K)$, $K > 0$ is contrast parameter. The main idea here is to control the diffusion near edges and smooth the flat regions, thereby integrating edge preserving nature into the diffusion process. The ADS (1) creates the scale space of images $\{u(x,t)\}_{t=0}^{T}$ with T as the final scale of choice, see Figure 1. Total variation (TV) function $g(s) = s^{-1}$ based diffusion PDE is proven to be more effective [2] than the original formulation (1) and it preserves edges across multiple scales better. TV based PDE is given by

$$\frac{\partial u(x,t)}{\partial t} = div\left(\frac{\nabla u(x,t)}{|\nabla u(x,t)|}\right) \quad \text{with} \quad u(x,0) = u_0 \tag{2}$$

As a first step in our segmentation algorithm we run TV scheme (2) to get rid of of the noise present and create a family of images based on the scale parameter t, see Figure 1(a).

2.2 Segmentation Scheme

Motivated by studies on scale space [10,7,8] and region extraction methods we prune the scale space created from the nonlinear diffusion (1) using color distances and coherent linking of objects along the scale axis. Since CIE-La*b* color space offers more perceptually uniform color distances with L^2 distance metric, $\|u - v\|^2 := (u_L - v_L)^2 + (u_{a*} - v_{a*})^2 + (u_{b*} - v_{b*})^2$, we make use of it in our algorithm. Our segmentation scheme is based on a key observation: as one traverse across the compact scale space $\{u(x,t)\}_{t=0}^{T}$, the local and coarse scale information is connected to a fine scale information progressively.



(a) $\{u(x,t)\}_{t=0}^{T}$

(b) $\{\mathcal{C}_{300}^{r}\}_{r=0}^{6}$ (c)

Fig. 1. (a) Scale space, $T = 500$ (b) Intra-scale splitting, shown at $t = 300$ (c) Inter-scale linking, shown for $\mathcal{C}_{300}^{r} \to \mathcal{C}_{400}^{k}$

(1) **Intra-scale splitting**: We partition the image domain $\mathcal{D}_t$ of $u(x,t)$ by connected components $\{\mathcal{C}_t^r\}_{r=0}^{P_t}$ where P_t the total number of distinct regions at scale t, in the following way:

$$\mathcal{C}_t^r := \{x \in D_t : u(x,t) = const\} \quad \& \quad \mathcal{C}_t^r \bigcap \mathcal{C}_t^l = \emptyset, \quad \mathcal{D}_t = \bigcup_{r=0}^{P_t} \mathcal{C}_t^r$$

Note that $\{\mathcal{C}_t^r\}$ are the level lines of the image $u(x,t)$ (see Figure 2(a) and (b)) and by [4] we know that the geometry of a color image is entirely described by them. Hence this splitting respects image boundaries through diffused images.

(2) **Inter-scale linking step**: Next we link two subsequent scale regions (say t and $t+1$) using the minimum color distance and overlapping as follows:

$$\mathcal{L}(\mathcal{C}_t^r) = \left\{ \min_{\mathcal{C}_{t+1}^r} \left\| u(\mathcal{C}_{t+1}^k, t+1) - u(\mathcal{C}_t^r, t) \right\| : \mathcal{C}_{t+1}^k \bigcap \mathcal{C}_t^r \neq \emptyset \right\}$$

and we denote it by $\mathcal{C}_t^r \to \mathcal{C}_{t+1}^k$. This can be iterated among scales and in general one can obtain $\mathcal{C}_{t_1}^r \to \mathcal{C}_{t_2}^k$ for $t_1 < t_2$ (see Figure 1(c)).

(3) **Final merging step**: All connected regions $\{\mathcal{C}_0^r\}_{t=0}^{P_0}$ of the initial image $u(x,0)$ that are linked to the same region at the end scale T are coming from one object. Thus the final segmentation can be obtained at scale level T as

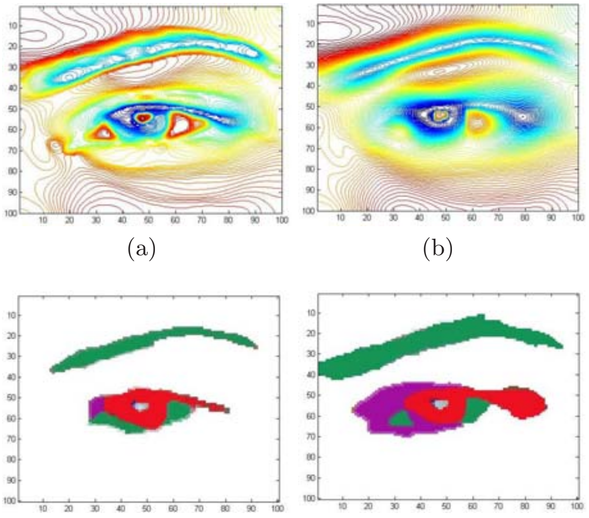
$$\mathcal{S}_T^k = \bigcup_{r=0}^{P_0} \mathcal{C}_0^r \quad \& \quad \mathcal{C}_0^r \to \mathcal{C}_T^k$$

where $\{\mathcal{S}_T^k\}_{r=0}^{P_T}$ are the segmented objects and P_T is the number of segmented objects.

The above methodology effectively combines the edge preserving property of nonlinear scale space based on multiscale anisotropic diffusion (1) with hierarchical segmentation and scale tracking.

3 Numerical Results

All the images in the computations are normalized to $[0, 1]$ and MATLAB®7.4 is used for visualization purposes. To compute color distances in the inter-scale linking step we convert the image from RGB to CIE-La*b* using the D65 white point reference. Recently, Bresson and Chan [2] formulated a fast algorithm for vectorial version of TV (2) using duality arguments. In our experiments we make use of it for its computational efficiency and the edge preserving property of TV diffusion. The scheme takes about 80 secs for an image of size 256 × 256 to run on a Pentium IV 2.40 GHz PC with 20 secs for diffusion (100 iterations) and the remaining time for the segmentation scheme. An additive Gaussian noise of $\sigma = 20$ is added to the original test images in all the experiments. Figure 2 compares the TV based (2) and ADS scheme (1) based segmentation results. As can be seen by comparing the corresponding contour maps



(a)

(b)

(c)

(d)

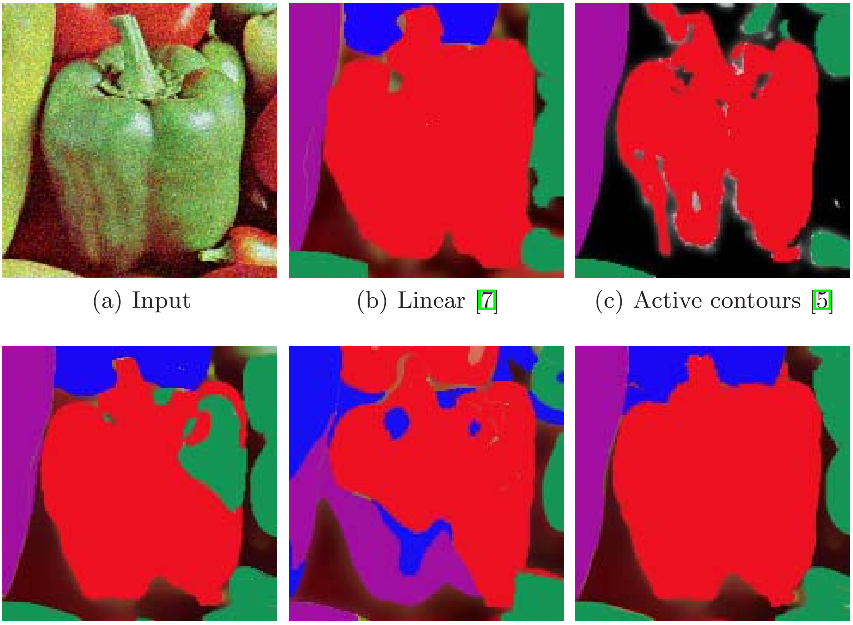
Fig. 2. (a) Level lines of the L channel for u^{100} using the TV flow (2) (b) Level lines of the L channel for u^{100} using the ADS with $(g(s) = \exp -(s^2/20))$in (1) (c) Segmented image based on TV flow and (d) Segmented image based on ADS

(Figure 2(a) and (b)), TV scheme segmentation (Figure 2(c)) yields a better segmentation in terms of the pupil and the eyebrow than ADS scheme segmentation (Figure 2(d)).

3.1 Comparison Results

We compare our scale space based segmentation scheme with the following schemes: linear scale space based scheme [7] (setting $g \equiv 1$ in (1)), geodesic active contours [5], Level set based scheme [12] (based on Mumford-Shah functional), Graph based scheme [6] (bottom-up clustering). Figure 3 shows the visual comparison results on a noisy *Peppers* color image. We use the same colors for segments in all the images to show the effect. As can be seen our scheme using TV scheme (2) (Figure 3(f)) results in a better segmentation of the middle pepper without any artifacts.

We used the CIE-La*b* space and L^2 metric for inter-scale linking step, other choices are also possible, for example HSV and chrominance-brightness space. Also adaptive selection of stopping time T [11] (also known as scale selection [3,8] in scale space literature) for the diffusion scheme (1) involved, determines the level of segmentation, and it remains to be explored. Comparing with other region based segmentation approaches quantitatively and the use of other anisotropic diffusion PDE based schemes [1] (for example, chromatic diffusion) defines our future work in this direction.



(a) Input (b) Linear [7] (c) Active contours [5]

(d) Level set [12] (e) Graph [6] (f) TV-Pruning

Fig. 3. Comparison of segmentation results on *Peppers* image

4 Conclusions

We used the nonlinear scale space concept to segment color images. The continuous family of images generated from nonlinear diffusion PDE are arranged according to their increasing scale and pruned using a track and merge scheme. Multiscale diffusion captures the edges without spurious oscillations and aides in scale capturing at the final scale of choice. By utilizing the fine to coarse set of images the segmentation is done iteratively. Preliminary numerical examples on images and comparison with related schemes show that the proposed approach gives better segmentation result on noisy and real images.

References

1. Aubert, G., Kornprobst, P.: Mathematical problems in image processing: Partial differential equations and the calculus of variations. Applied Mathematical Sciences, vol. 147. Springer, New York (2006)
2. Bresson, X., Chan, T.F.: Fast dual minimization of the vectorial total variation norm and applications to color image processing. Inverse Problems and Imaging 2(4), 455–484 (2008)
3. Brox, T., Weickert, J.: A TV flow based local scale measure for texture discrimination. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3022, pp. 578–590. Springer, Heidelberg (2004)
4. Caselles, V., Coll, B., Morel, J.-M.: Geometry and Color in Natural Images. J. Math. Imag. Vis. 16, 89–105 (2002)
5. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. 22(1), 61–79 (1997)
6. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. Int. J. Comput. Vis. 59(2), 167–181 (2004)
7. Florack, L.M., ter Haar Romeny, B.M., Koenderink, J.J., Viergever, A.M.: Linear scale-space. J. Math. Imag. Vis. 4(4), 325–351 (1994)
8. Lindeberg, T.: Edge detection and ridge detection with automatic scale selection. Int. J. Comput. Vis. 30(2), 117–156 (1998)
9. Morel, J.-M., Solimini, S.: Variational Methods in Image Processing. Birkhauser, Boston (1994)
10. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Machine Intell. 12(7), 629–639 (1990)
11. Vanhamel, I., Mihai, C., Sahli, H., Katartzis, A., Pratikakis, I.: Scale selection for compact scale-space representation of vector-valued images. Int. J. Comput. Vis. 84(2), 194–204 (2009)
12. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation using the Mumford Shah model. Int. J. Computer Vis. 50(3), 271–293 (2002)

Effective and Efficient Tracking and Ego-Motion Recovery for Mobile Cameras

Huiyu Zhou[1] and Gerald Schaefer[2]

[1] Institute of Electronics, Communications and Information Technology
Queen's University Belfast, Belfast, U.K.
[2] Department of Computer Science
Loughborough University, Loughborough, United Kingdom

Abstract. Estimating 3-D structure and camera motion from 2-D image sequences is an important problem in computer vision. In this paper we present an effective approach to tracking and recovery of ego-motion from an image sequence acquired by a single camera attached to a pedestrian. Our approach consists of two stages. In the first phase, human gait analysis is performed and human gait parameters are estimated by frame-by-frame analysis utilising a generalised least squares technique. In the second phase, the gait model is employed within a "predict-correct" framework using a maximum a posteriori expectation maximisation strategy to recover ego-motion and scene structure, while continuously refining the gait model. Experiments on synthetic and real image sequences confirm that the use of the gait model allows for effective tracking while also reducing the computational complexity.

1 Introduction

Estimating 3-D structure and camera motion from 2-D images is one of the fundamental problems in computer vision. One strategy for recovering structure from passive image data is structure from motion (SFM) which also allows recovery of the ego-motion of the sensor with respect to a world co-ordinate system, performing 3-D scene reconstruction and navigation [1]. The majority of existing methods for feature tracking use frame-to-frame prediction models based for example on Kalman filters [2], particle filtering [3], or optimisation based approaches [4].

Although working well in certain applications, these approaches do not take into account the long-term history of the camera's motion. In contrast, in this paper we track features and recover ego-motion from a temporal image sequence acquired by a single camera mounted on a moving pedestrian. We show that the use of an explicit longer term, non-linear human gait model is more efficient in this case since fewer features are lost and the processing time per frame is reduced as either the search window or the frame rate can be reduced. Our work is inspired by the study of Molton *et al.* [5] who used a robotic system to measure the gait of a walking person, while a digital compass and an inclinometer were used to record rotation. They employed an iterated and extended Kalman filter to initialise wavelet parameter estimates, then running across the whole period of activity. In our work, motion is computed directly from video data, and the emphasis is on the use of a longer term model to increase algorithmic efficiency.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 345–350, 2009.

2 Tracking and Ego-Motion Recovery

In the following, we use the term “ego-motion" to refer to both frame-to-frame and longer-term periodic motion while “camera transformation" refers to the ego-motion of the camera between any two frames, and “gait model" to the longer-term ego-motion of the camera over many frames. Our proposed approach consists of two phases: initialisation and continuous tracking. We assume that m frames are included in the initialisation phase and n frames ($n > m$) in the whole sequence, and that the intrinsic camera parameters are known.

2.1 Phase 1: Initialisation

The purpose of the first phase is to acquire the long-term gait model of the pedestrian. The general workflow of the initialisation phase is outlined in Algorithm 1. Of the well-tested feature tracking algorithms, reviewed by Lepetit and Fua [6], the Shi-Tomase-Kanade (STK) tracker [7] is chosen because of its accuracy and reliability. The STK tracker has two stages: feature selection and feature tracking. The feature selection process computes the eigenvalues of a gradient function for each pixel, comparing the result with a fixed threshold. In the published algorithm, image features with higher eigenvalues are considered as good features to track. However, in our approach we use the SUSAN (Smallest Univalue Segment Assimilating Nucleus) [8] operator instead to select an initial feature set in the first frame, due to its known robustness to noise. This is not used subsequently, unless the number of tracked features falls below a pre-defined threshold in which case we re-initialise with new features.

Algorithm 1. Pseudo-code of Phase 1 of the algorithm

```
Select $C$ corner features in frame 1
for $i$ = 1 to $m$-1 do
    Match features across frames $i$ and ($i$+1)
    Estimate fundamental matrix
    Refine list of matched features
    Recover camera transformation and scene geometry
end for
Fit periodic gait model to camera transformation data from frames 1 to $m$
```

As we track features in successive frames, we minimise the residual error using an affine transformation, as in the STK algorithm, assuming the displacement is small. However, we use a generalised least squares (GLS) algorithm [9] to recover the frame-to-frame camera transformation and scene geometry. The application of GLS provides a more robust estimation technique that can derive good results also when the error distribution is not normally distributed [10]. Typically, m is chosen large enough to recover about two complete strides, e.g. $m = 50$ for a 25Hz sampling rate. The number of features C is a user parameter, typically set to 150 in our experiments. This leads to the recovery of a temporary motion model with 6 degrees of freedom, i.e. the 3 displacements and 3 Euler angles, which can be stored in a gait library using a truncated Fourier series representation.

2.2 Phase 2: Continuous Tracking

In Phase 2, outlined in Algorithm 2, we use and update the gait model to improve the prediction of the location of features in each new frame. Since this prediction is based on a longer history, not every frame needs to be considered and we include a parameter k that describes the gap between frames. The coarse-to-fine matching parameters, e.g. window size, are selected according to the variance of the Euclidean distance between the predicted and measured feature positions. A maximum a posteriori expectation maximisation (MAP-EM) algorithm is used to find the best feature match and determine the camera transformation parameters between two frames. The gait model is examined and updated if necessary on a periodic basis using iteratively re-weighted least squares (IRLS) [11]. To maintain its validity, the m most recent frames are used, applying Spearman correlation to compare the recent gait parameters individually with the stored values. If the correlation coefficients are lower than a pre-defined threshold (typically 0.9), then the current gait model is updated. Accumulation of errors in the motion tracking can deteriorate the performance of the proposed tracker. In order to reduce this accumulated error we can re-estimate the overall parameters by collecting very similar (or the same) images that were just used. For example, multiple estimates of the position of a known 3-D point should be identical. This strategy works effectively if the iteration runs several times.

Algorithm 2. Pseudo-code of Phase 2 of the algorithm

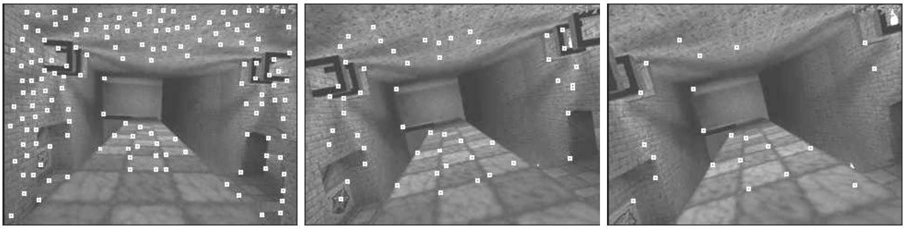
```
for $i = m$ to $n$ step $k$ do
    Predict the feature positions in the ($i$+$k$)-th frame using the gait model
    Apply a coarse-to-fine strategy to match the features
    Compute frame-to-frame transformation and scene geometry on every $m$-th frame
    Update the periodic gait model using iteratively re-weighted least squares
end for
```

3 Experimental Results

To demonstrate improved performance, we compare the proposed gait-based ego-motion tracking with the original STK algorithm which uses a short-term displacement model. Experiments were obtained based on several synthetic and real image sequences, however due to space limitation we present here only the results of one synthetic and one real sequence.

The synthetic test data is obtained from a computer game simulation[1] for comparison of the algorithms with known periodic gait parameters. We can compare the recovered parameters against the ground truth as the gait parameters can be fully programmed using the game engine. Here we show the results of a demonstration video (“rolling sequence") in which the player wanders in a castle in changing illumination, which together with the more complex 3-D environment can cause missed feature

[1] The Quake Engine available at
http://www.codeproject.com/managedcpp/quake2.asp



(a) Frame 1 (b) Frame 5 (c) Frame 10

Fig. 1. The “rolling sequence" with the feature points superimposed: (a) frame 1 (150 feature points); (b) frame 5; (c) frame 10 (22 feature points)

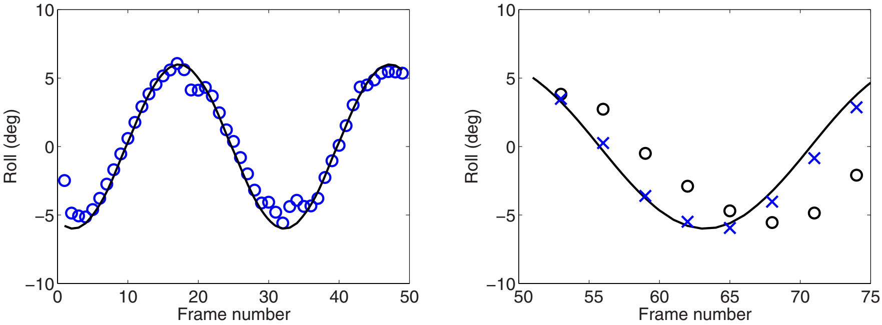
correspondences. In the sequences, we defined a forward translation, and the roll angles of the “observer" were altered according to the expression:

$$\text{roll angles (deg)} = -6 \times \sin(2\pi \times t_f/30 + 1.1),$$

where t_f is the image frame index.

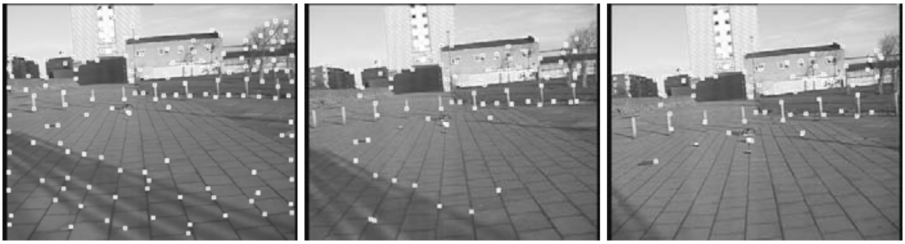
In Phase 1, feature tracking shown in Fig. 1 leads to the history of the roll angles shown in Fig. 2(a). This model is used for subsequent tracking with an interval of 3 frames. Fig. 2(b) shows the 25 frames immediately succeeding the learning phase, in which our gait-based strategy clearly outperforms the STK scheme in terms of measurement accuracy. The longer interval between the neighbouring frames leads to the violation of the linearisation assumption in the STK scheme, resulting in large errors in the estimated motion parameters.

In further experiments, we also employ real data from a camera mounted on a pedestrian. As we do not have independent extraction of gait in this case, we compare texture mapped images using extracted parameters with the real image data, which gives a



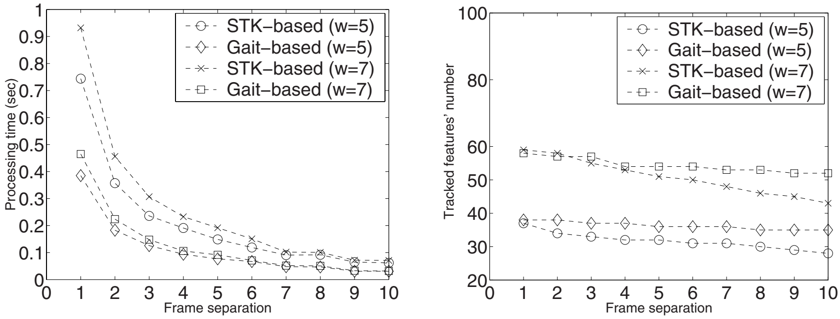
(a) Roll of frames 1-50 (b) Roll of frames 51-75

Fig. 2. Performance comparisons of the STK and the gait-based tracker: the line denotes the ground truth, the circles the result of the STK tracker, and the crosses that of our gait-based tracker



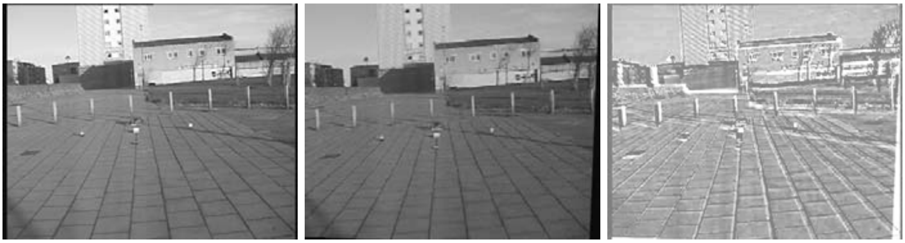
(a) 1st frame (b) 20th frame (c) 50th frame

Fig. 3. Real walking sequence with feature points superimposed



(a) Processing time (b) Retention of features

Fig. 4. Comparison between the STK-based and gait-based framework in tracking real walking sequence (w = width of search window)



(a) (b) (c)

Fig. 5. Comparison of real frames and their texture-maps for real walking sequence: (a) The actual frame 55. (b) The texture map of frame 55 created by rendering. (c) Subtraction of (a) and (b). For better visibility, the brightness/contrast of (c) has been increased.

strong subjective comparison. Fig. 3 gives some example frames from an image sequence collected by the camera, where the detected and tracked feature points are superimposed. After a learning period of 50 image frames, we obtain a motion pattern. We then predict the motion parameters in successive frames, followed by immediate correction using the gait-based scheme.

Fig. 4 shows a comparison between the STK- and gait-based approaches in tracking the sequence of Fig. 3 in terms of processing time (using a Pentium II-300 MMX PC), and retention of features. Compared to the STK-based approach using a fixed number of pyramid levels, our gait-based method localises the search better and hence reduces the time taken to find a match. The gait-based scheme also loses fewer feature points compared to the STK-based strategy.

In Fig. 5 we show the real and the texture mapped scenes for a frame using the estimated motion parameters. The subtraction image between the two demonstrates the accuracy of the approach.

4 Conclusions and Future Work

In this paper we have presented an approach for effective, efficient and robust ego-motion tracking using a single mobile camera without using any other indicators of position, speed or inclination. An initial gait model is extracted from a fixed training period of two strides, using feature correspondences to estimate the ego-motion parameters, represented as a truncated Fourier series. Experiments on synthetic and real data have shown that the proposed strategy has more accurate and efficient ego-motion estimates and structural recovery than a comparable method that does not incorporate long-term motion estimates.

References

1. Oliensis, J.: A critique of structure-from-motion algorithms. Computer Vision and Image Understanding 80, 172–214 (2000)
2. Nava, F., Martel, A.: Wavelet modeling of contour deformations in Sobolev spaces for fitting and tracking applications. Pattern Recognition 36, 1119–1130 (2003)
3. Ristic, B., Arulampam, S., Gordon, N.: Beyond the Kalman Filter: Particle Filters for Tracking Applications. Artech House (2004)
4. Vidal, R., Hartley, R.: Three-view multibody structure from motion. IEEE Trans. Pattern Anal. Mach. Intell. 30, 214–227 (2008)
5. Molton, N., Brady, J.: Modelling the motion of a sensor attached to a walking person. Robotics and Autonomous Systems 34, 203–221 (2001)
6. Lepetit, V., Fua, P.: Monocular model-based 3D tracking of rigid objects: a survey. Foundations and Trends in Computer Graphics and Vision 1, 1–89 (2005)
7. Shi, J., Tomasi, C.: Good features to track. In: International Conference on Computer Vision and Pattern Recognition, pp. 593–600 (1994)
8. Smith, S., Brady, J.: SUSAN: A new approach to low-level image-processing. Int. J. of Comput. Vis. 23, 45–78 (1997)
9. Zhou, H., Green, P., Wallace, A.: Fundamental matrix estimation using generalized least squares. In: Proc. of International Conference on Visualisation, Imaging, and Image Processing, pp. 79–85 (2004)
10. Zhou, H.: Efficient motion tracking and obstacle detection using gait analysis. PhD thesis, Heriot-Watt University, Edinburgh, UK (2005)
11. Beaton, A., Tukey, J.: The fitting of power series, meaning polynormials, illustrated on band-spectroscopic data. Technometrics 16, 147–185 (1974)

Class Specific Threshold Selection in Face Space Using Set Estimation Technique for RGB Color Components

Madhura Datta[1] and C.A. Murthy[2]

[1] UGC-Academic Staff College, University of Calcutta, Kolkata 700 009, India
madhuradatta@gmail.com

[2] Machine Intelligence Unit, Indian Statistical Institute, Kolkata 700 108, India
murthy@isical.ac.in

Abstract. In conventional face recognition techniques, any query image is always classified to one of the face classes irrespective of whether the query image is a face or not. Most of the recognition algorithms are dissimilarity based, and one needs to put a proper threshold on the dissimilarity value for the sake of classification. In this paper, we have introduced a novel thresholding technique for classification, where the query image is not necessarily classified to one of the classes in the training set. The theoretical formulation of the thresholding technique and its utility are demonstrated on color face and non face datasets with RGB color components as features in the subspace. The proposed threshold selection is based on statistical method of set estimation and is guided by minimal spanning tree. Experiment shows that the proposed class specific threshold based technique performs better than the non threshold based systems using subspace algorithms.

Keywords: Feature extraction, Color components, Minimal spanning tree, Intra-class threshold, Set estimation.

1 Introduction

Among the popular algorithms for face recognition, subspace methods [1,2]are generally based on dissimilarity, where the query image is put in the class for which the dissimilarity is minimal. This is a classical approach of identification and known as closed test identification where the test face always exists in the gallery dataset. However, in a real life scenario the identification system may face a situation where the query face image may not be present in the database, i.e. often referred by the biometric researchers as the open test identification [3]. A way of achieving such a task is to put a threshold on the dissimilarity value at the identification stage. On the basis of the decision threshold, a biometric recognition system should be in a position to accept the query image as client or reject it as imposter. The selection of proper threshold of a given class in a dataset is an open question.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 351–356, 2009.

As found in literature Mansfield [4]et. al proposed equal error rate (EER),the point on ROC curve where FAR = FRR, to be selected as the operating threshold. Martin [5] proposed the use of detection error trade-off (DET) curve which is a non-linear transformation of ROC curve. In reality, the common practice is to use the global threshold for a system rather than using user dependent versions of ROC. In the current problem we are interested in finding the threshold values using set estimation method for color face identification system. The set estimation method in literature has been mainly used to find the pattern class and its multi-valued boundary from its sample points in $\Re^2$ and higher dimensional spaces. Some investigations on estimation of α- shapes for point sets in $\Re^3$ had been proposed by Edelsbrunner [6]. Later Mandal et al. [7] developed and extended their method to $\Re^m$ and found it very useful in developing a multi-valued recognition system. As one can get the shape or boundary of a given set then that procedure of set estimation also generates the intuition for determination of the class thresholds of the set. As a tool of set estimation, minimal spanning tree (MST) is proposed to calculate threshold value. For the purpose of recognition, we proposed two types of threshold values (i) global (ii) local (or user specific) for each face classes. Here two types of decisions can be made by the system. Local threshold of a class is used to conclude whether the given query image belongs to the given class. Global threshold is a general threshold which can be used to conclude whether the query image belongs to the given collection of face classes. These two types of thresholds will satisfy a biometric system in both the closed and open test identification cases.

In this paper we assume the frontal color views are available. Color information is commonly used for detection of faces but its discriminating power is still in question. Role of color in face verification needs to be explored since, generally, it is considered to be a low level feature. This feature is surprisingly effective in race classification. The common approach in most of the face recognition systems is to convert the color images to monochromatic form using simple transformations. There was also another approach where the RGB space has been compared with the YUV and HSV color models by L. Torres [8] for PCA based face recognition. Several studies have been made to evaluate color spaces. They have concluded that RGB space should be suitable for any application on face identification among the other color spaces. In this paper, we considered a few linear combinations of R, G, and B for dimensionality reduction and consequently for classification. To reduce the dimensionality, we need to extract features by using any one of the sub space methods. Two types of feature extraction methods of the images in the face database are explored. These are (i) principal component analysis (PCA) [1,2,3] (ii) kernel PCA [1,2,3].

2 Mathematical Preliminaries for Threshold Selection

The system assumed to have n images corresponding to a particular expression of a particular person P. If we represent an image of an expression of P by a vector x_0, then the set corresponding to the small variations in the same expression may be assumed to be a disc of radius $\epsilon > 0$ around x_0. The set corresponding to an expression of the same person P may be taken as $\bigcup_{i=1}^{n} x \in \Re^m : d(x_i, x) \leq \epsilon$ where x_1,x_2,$x_n$are the n vectors corresponding to the given n images. The set corresponding to the union of all possible expressions of a person may also be taken as a connected set since for two different images of the same expression, P must be able to provide the intermediary images of that expression (a path connecting the two points is completely contained in the set.). The face class of a person is nothing but the set of all possible expressions of that person. A general formulation of the face class, probably, would have the radius value depending on the center of the disc. The radius value is taken to be independent of the the center of the disc.In the above formulation, as the number of face images of the same person increases, we shall be obtaining more information regarding the face class, and hence the radius value needs to be decreased. Thus the radius value needs to be a function of the number of images.

Usually one may want to estimate a set on the basis of the given finitely many points.Set estimation problem and its utility, methodology, conditions on sets are well documented in the literature [6,9,10]. A formulation with only the relevant part used for the class specific threshold selection is presented in this regard. Let $\alpha \subseteq \Re^m$ be the unknown path connected set in m dimensional space and ϵ_n be a sequence of positive numbers such that $\{\epsilon_n\} \rightarrow 0$ and $n\epsilon_n^m \rightarrow \infty$ as $n \rightarrow \infty$.

$$\alpha_n = \bigcup_{i=1}^{n} \{x \in \Re^m : d(x, X_i) \leq \epsilon_n\}$$

where d denotes the Euclidean distance. Then α_n be a consistent estimate of α and ϵ_n's are chosen in different ways for different problem domain.

3 Threshold Incorporated Face Classification Using Set Estimation Method

There are several ways in which we can make the estimated set connected. We shall describe a generic way of making the estimated set connected, where only finite union of disks is considered.

Method: (a) Find minimal spanning tree (MST) of $S = \{X_1, X_2, \ldots X_n\}$ where the edge weight is taken to be the Euclidean distance between two points.

(b) Take ϵ_n as Maximum of the (n-1) edge weights of MST.

(c) Take the estimate α_n as

$$\alpha_n = \bigcup_{x \in S} \{y : d(x, y) \leq \epsilon_n\}$$

Note that ϵ_n is the threshold for the set α_n since no point outside the ϵ_n disks is considered to be a member of the set. If we represent an image by a vector $\underline{x}$, then we are considering all possible such vectors corresponding to a human being. Let us represent such a set by α . This set denotes the face class of that human being. It can be noted that we don't know α completely. Only a few points of α like the different expressions of a face are known to us. The proposed set estimation method can be utilized in two ways in face identification problem as described in the next subsection.

3.1 Local Threshold Based Recognition

We assume that we have M classes, each class denoting a human being. Each class consists of N vectors of m dimensions. Here for each class we calculate MST of the respective N vectors and find its maximal edge weight. Let us denote the maximal edge weight of the MST of the i_{th} class by ξ_i for any m dimensional vector $\underline{x}$ in the following way. The total number of given vectors is MN. For each class i, find the minimum distance of $\underline{x}$with all the N points in the class. Let the minimal distance be ρ_i

(a) If there exists an i such that $\rho_i < \xi_i$ then put x in the i_{th} class.

(b) If there does not exist any i such that$\rho_i < \xi_i$ then there the given image does not fall in any one of the given face classes. If the given image does not fall into any of the given face classes, it may still lie within the given face space. In order to analyze this possibility, the global threshold is formulated.

3.2 Global Threshold Based Recognition

In this recognition system, the number of images is large. For determining the global threshold, we have the following algorithm.

(a) Find MST of MN points and find half of its maximal edge weight. Let it be ξ

(b) Find

$$\rho = Min_{i=1,2,...n}\rho_i \tag{1}$$

(c) If $\rho < \xi$ then the image is a face image and will form a new face class in the face space. If $\rho > \xi$ then the given vector x does not belong to the given face space. Here the image is either a non face image or a face image not belonging to the given face space. Note that $\xi_i \leq \xi, \forall i$ is true generally but not for any set of MN points.

4 Experimental Design and Result Analysis on Color Components

The proposed method of threshold based classification has been used for open and closed test face identification and tested over the well known color AR face database [11] and one object (i.e., non face) database namely, the COIL-100 [12]. For AR database we have taken the first 10 images from each of 68 classes and the face portions are manually cropped to dimension 128X128. To form the face space from the image space we have used two well known feature selection methods (i) principal component analysis (PCA), (ii) Kernel PCA The selection of color components on the color dataset have been done in several ways before applying PCA and KPCA on the datasets.In the first procedure we applied the scheme proposed by L. Torres et al. [8] the PCA projection was computed on color components separately and the global distance is measured between test and training images. As procedures 2, 3, and 4, we have used the R, G and B color channels are separately used. In procedure 5, PCA and kpca are applied on (R+G+B)/3, where as in procedure 6, it is applied on (3R+4G+3B)/10. These procedures are denoted by P1, P2,, P6 in the table. The number of color features is 340 for each of R, G, B, and thus the total number of components in feature vector is 340x3. For each of the other procedures, the number of features is 340.

4.1 Open and Closed Test Identification in Color Space

After formation of the face space,the proposed MST based method (discussed in section 3.1) of class specific threshold selection is applied on the each training class of color face points for each procedure. The number of training images for each class of the AR dataset is taken to be 5 chosen randomly. Remaining images for each class are considered for test set. Nearest neighbor (NN) classifier is used to get the non threshold based identification rate.It is observed from Table 1 that for each color components proposed method performs better than the non threshold based systems.

For Open test identification AR used as gallery face points and COIL-100 non face dataset containing 7200 color images of 100 non face objects of size 128X128 are used as probe set. The global threshold formulated in section 3.2 is applied in this case.It is apparent from table 1 that no object image in the COIL-100 dataset is classified to none of the face classes of the training set. Both the results satisfy the theoretical aspects of the proposed local and global threshold selection method.

Table 1. Open and close test identification

Procedure number	closed test identification		open test identification	
	NN classifier	local thresholds	non face points inside global threshold	outside global threshold
P1	96.5(PCA)	98.75	1	99
	91(KPCA)	93.5	0	100
P2	96(PCA)	96.75	0	100
	94(KPCA)	94.5	0	100
P3	96.5(PCA)	98	0	100
	97(KPCA)	98	0	100
P4	92(PCA)	94	0	100
	92(KPCA)	93	0	100
P5	95(PCA)	97	0	100
	92(KPCA)	92	0	100
P6	96(PCA)	98	0	100
	91(KPCA)	94	0	100

5 Conclusions

A method is proposed for threshold selection for subspace based face identification scheme. The method has been applied on the color components of RGB space of the AR databases and it is found to give better results than those methods without threshold. We have also tested the utility of the method, where the training and the test datasets are face and non face datasets using global threshold. One has to take care of several inherent issues regarding the subspace based methods. In all our experiments on the face datasets, we have considered only five points for each class in the training set. We have probably taken the lowest possible such value for the number of training points of each class. It is expected that the results can improve if the size of the training set is larger.

References

1. Solar, J.R.D., Navarrete, P.: Eigenspace-based face recognition: A comparative study of different approaches. IEEE Transactions on Systems, Man and Cybernetics, Part C 35(3), 315–325 (2005)
2. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face recognition: A literature survey. ACM Computing Surveys, 399–458 (2003)
3. Shakhnarovich, G., Moghaddam, B.: Face Recognition in Subspaces. In: Li, S.Z., Jain, A.K. (eds.) Handbook of Face Recognition. Springer, Heidelberg (2004)
4. Mansfield, A.J., Wayman, J.L.: Best practices in testing and reporting performance of biometric devices, version 2.01. Centre for Mathematics and Scientific Computing, National Physical Laboratory,middlesex (August 2002)
5. Martin, A., Doddington, G., Kamm, T., Ordowski, M.: The det curve in assessment of detection task performance. In: Proc. of Eurospeech 1997, vol. 4, pp. 1895–1898 (1997)
6. Edelsbrunner, H., Kirkpatrick, D., Seidel, R.: On the shape of a set of points in a plane. IEEE Trans. on Inform. Theory IT-29, 551–559 (1983)
7. Mandal, D.P., Murthy, C.A., Pal, S.K.: Determining the shape of a pattern class from sampled points: Extension to rn. Int. J. of General Systems 26(4), 293–320 (1997)
8. Torres, L., Reuttter, J.Y., Lorente, L.: The importance of the color information in face recognition. In: Proceedings of IEEE Int. Conf. on Image Processing Cobe, Japan, pp. 25–29 (1999)
9. Murthy, C.A.: On consistent estimation of classes in the context of cluster analysis (1988)
10. Grenander, U.: Abstract inference. John Wiley, New York (1981)
11. Martinez, A., Benavente, R.: The ar face database. CVC Technical report 24 (June 1998)
12. Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library (coil-100). Technical report cucs-006-96 (February 1996)

Human Action Recognition Based on Spatio-temporal Features

Nikhil Sawant and K.K. Biswas

Dept. of CSE, IIT Delhi-110016, India
{mcs072899|kkb}@cse.iitd.ac.in
http://cse.iitd.ac.in/~mcs072899

Abstract. This paper studies the technique of human action recognition using spatio-temporal features. We concentrate on the motion and the shape patterns produced by different actions for action recognition. The motion patterns generated by the actions are captured by the optical flows. The Shape information is obtained by Viola-Jones features. Spatial features comprises of motion and shape information from a single frame. Spatio-temporal descriptor patterns are formed to improve the accuracy over spatial features. Adaboost learns and classifies the descriptor patterns. We report the accuracy of our system on a standard Weizmann dataset.

1 Introduction

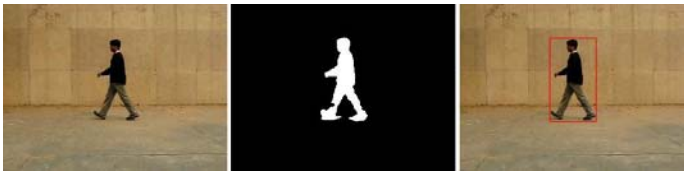
Human action recognition is becoming increasingly important for automation of video analysis. With the growing need of surveillance related applications the research in the field of action recognition has been fueled in past few years. A detailed survey of action recognition techniques has been presented by Gavrila [1]. The researchers have used space-time features to identify the specific action [2,3]. Computer vision scientists have tried motion based techniques [4,5,6] as any action is associated with some motion. Niu et. al. [4] have used both motion and eigen shape features to carry out view invariant activity recognition. Bag of words [5] is recently has been used for the task of action recognition.

We make use of both shape and motion patterns as well as space-time pattern features. Adaboost learns and detects the patterns of different actions. Patterns are spatio-temporal features made up of motion and shape information. The paper is organized as follows: Section 2 explains about target localization. In sections 3 and 4, we discuss the motion and shape descriptors respectively followed by discussion on spatio-temporal features in section 5. The learning process is explained in section 6 and we present our results and conclusions in section 7.

2 Target Localization

Target localization helps reducing the search space. Background subtraction is used to generate a silhouette of the target as shown in Figure 1(b). We assume

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 357–362, 2009.



(a) (b) (c)

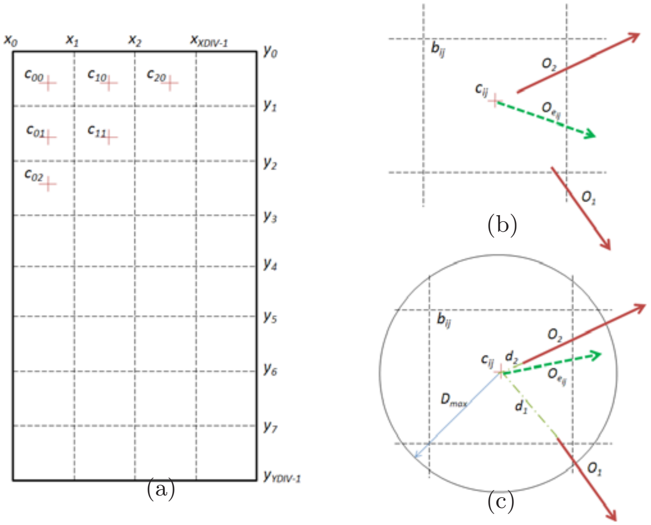
Fig. 1. Target localization. (a) shows a video frame, (b) is the extracted silhouette where the background is stable, (c) is the original frame (a) with *ROI* marked with the help of silhouette in (b).

that the action is being performed in front of a stable background. With the help of silhouette information the region of Interest (*ROI*) can be determined as shown in Figure 1(c). Once the *ROI* is marked we can concentrate only on the area inside *ROI*.

3 Motion Descriptor

It has been shown that different activities produce different motion patterns. We make use of Lucas - Kanade method [7] to generate the optical flows in the *ROI* to capture the motion patterns. The advantage of this method is that it comparatively yields robust and dense optical flow fields.

Organizing optical flows using averaging. After computation of optical flow our job is to arrange it in some fashion so that a descriptor can be formed.



(a) (b) (c)

Fig. 2. (a) shows a grid overlaid over the *ROI*, (b) Simple averaging: $O_{e_{ij}}$ formed by averaging of all the vectors within the box b_{ij} i.e O_1 and O_2, (c) Weighted average: O_2 has more influence on $O_{e_{ij}}$ compare to O_1

A fixed size grid is overlaid over the ROI as shown in Figure 2(a). This grid divides the silhouette into boxes $\{b_{ij}, ...\}$ with centres at $\{c_{ij}, ...\}$ respectively. An intuitive way is to average out the optical flow from each box to form the effective optical flow $O_{e_{ij}}$ which is formulated in Eq 1.

$$O_{e_{ij}} = \frac{\sum_{k=1}^{m'} O_k}{\sum_{k=1}^{m'} 1} \quad \text{where } (x_i \leq x_{O_k} < x_{i+1}) \ \& \ (y_j \leq y_{O_k} < y_{j+1}) \quad \forall k \tag{1}$$

Here m' is the set of optical flows within the box b_{ij}. $O_{e_{ij}}$ is the effective optical flow for box b_{ij}. x_{O_k} and y_{O_k} are the x and y co-ordinates of the k^{th} optical flow O_k respectively. In Figure 2(b), $O_{e_{ij}}$ has equal contribution from all the optical flows present within the box.

A possible drawback in the simple averaging method is that all the optical flows have same weight irrespective of their distance from the centre of the box. Thus the net optical flow may be swayed by an optical flow in different direction sitting at the boundary of the box. Thus we present a new weighted method to compute the net optical flow for each box.

Organizing optical flows using weighted average. As shown in the Figure 2(c) we sum up the contribution of various optical flow vectors at c_{ij} the centre of each grid cell. We assume that only the flow vectors lying within distance D_{max} from the grid center would be allowed to influence the computation of the effective optical flow. The contribution of each flow is weighted inversely by the distance from the center. The net flow is computed by the following Eq 2.

$$O_{e_{ij}} = \frac{\sum_{k=1}^{m''} ((D_{max} - d_k) O_k)}{\sum_{k=1}^{m''} (D_{max} - d_k)} \quad \text{where } D_{max} \geq d_k \quad \forall k$$
$$\text{and } d_k = \sqrt{(x_{O_k} - x_{c_{ij}})^2 + (y_{O_k} - y_{c_{ij}})^2} \tag{2}$$

Here m'' is the set of optical flows within range of D_{max} from the centre c_{ij}. $x_{c_{ij}}$ and $y_{c_{ij}}$ is nothing but the x and y co-ordinates of the centre c_{ij} of the box b_{ij} respectively.

Figure 3(c) shows optical flow after applying weighted average on each box b_{ij} of the overlaid grid. In order to form the descriptor the effective optical flow $O_{e_{ij}}$ is split into $O_{ex_{ij}}$ and $O_{ey_{ij}}$ in the respective direction. For a grid size of $(M\text{x}N)$, we have $2MN$ values representing the motion descriptor.

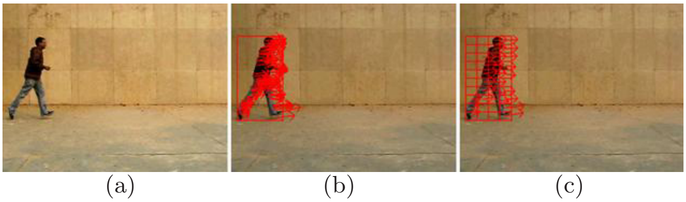


Fig. 3. Organized optical flows. (a) original frame, (b) unorganized optical flows, (c) organized optical flows.

4 Shape Descriptor

The shape of a silhouette is also a characteristic of the action being performed. Shape information is helpful when motion information is noisy or not sufficient. Niu et. al. [4] used shape information by adding the value of each and every pixel to the descriptor, but this requires resizing the silhouette.

Differential shape information. We propose use of rectangular features introduced by Viola-Jones [8] for face detection. These features are simple and have already proven their usefulness. Figure 4(b)(c) show $two - rectangle$ and $four - rectangle$ features used by us. Rectangular features are applied at box level and not at pixel as done by Viola-Jones. We overlaid the grid over the ROI as shown in Figure 4(a). Each box is assigned the percentage of foreground pixels within.

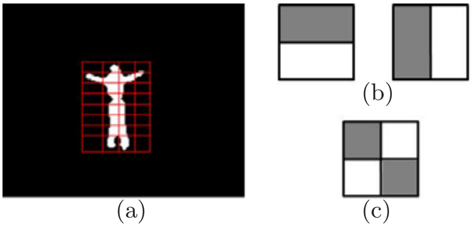


Fig. 4. Shape Information. (a) silhouette with the grid overlaied within the ROI, (b) $two - rectangle$ features, (c) $four - rectangle$ features.

5 Spatio-temporal Descriptors

The motion and the shape descriptor described in previous two sections can be obtained from each individual frame of the video sequence. In order to improve the accuracy of action recognition we make use of spatio-temporal features. Spatio-temporal features not only carry information about current frame but also about neighboring frames. We fix the number of frames for all the videos under consideration, and call it $TLEN$. To reduce the computational overhead in considering all the frames, we select the frames with a fixed offset called $TSPAN$, for example if we choose $TSPAN$ as 3, we pick up every third frame from the video clip and stack all these together to extract the spatio-temporal features. *TSPAN* helps in reducing the descriptor length without much change in the accuracy.

6 Learning with Adaboost

We use standard Adaboost learning algorithm [9]. Adaboost is the state of art boosting algorithm. Linear decision stumps are used as the weak hypothesis. We train our system for patterns of all the chosen actions. Once trained the System can recognize the patterns produced by different actions. Our training and testing data is mutually exclusive. Also we have used different subjects for training and testing sequences.

7 Results and Conclusion

We conducted our initial experiments on a small dataset built by us, infront of stable background. The dataset has 7 actions performed by 5 to 8 actors. Our dataset contains around 10 videos of each action. Figure 5(a) shows the snapshot of our dataset, various actions covered are walking, running, waving1, waving2, bending, sit-down, stand-up (left to right, top to bottom). We also tested our method for standard Weizmann dataset [3] . Snapshot of the Weizmann dataset is shown in Figure 5(b), various actions covered are bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2 (left to right, top to bottom), each action has 9 videos performed by 9 separate actors. The fixed grid of 4 x 8 is overlaid on the *ROI*. For spatio temporal features, we experimented with *TLEN* as 5 and *TSPAN* as 5. Figure 6(a) shows the confusion matrix for our dataset. As we see there is 0% error rate for walking, running, waving2, sit-down. Overall error rate of recognition is 4:28%. For standard Weizmann dataset, results are shown in Figure 6(b). There is slight error in run and wave1 actions, rest all the actions are performed with 0% error. Our error is rate 2:17% which is better than 16:3% reported recently by T. Goodhart [5].

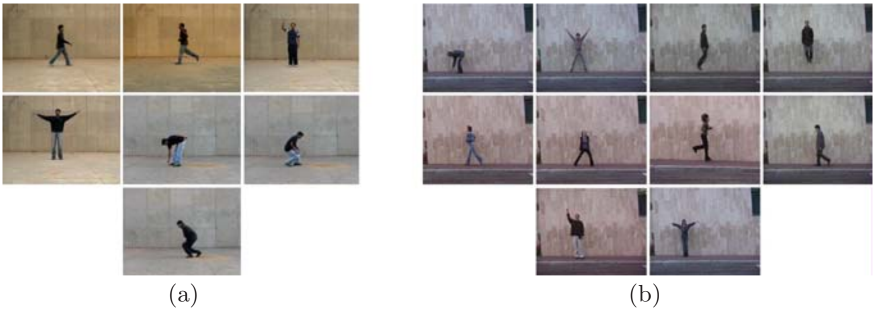
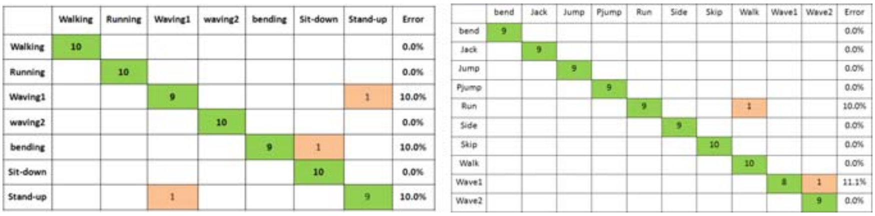


Fig. 5. Dataset used. (a) Our own dataset consisting of 7 actions, (b) standard Weizmann Dataset containing 10 actions.



	Walking	Running	Waving1	waving2	bending	Sit-down	Stand-up	Error
Walking	10							0.0%
Running		10						0.0%
Waving1			9				1	10.0%
waving2				10				0.0%
bending					9	1		10.0%
Sit-down						10		0.0%
Stand-up			1				9	10.0%

(a)

	bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2	Error
bend	9										0.0%
Jack		9									0.0%
Jump			9								0.0%
Pjump				9							0.0%
Run					9			1			10.0%
Side						9					0.0%
Skip							10				0.0%
Walk								10			0.0%
Wave1									8	1	11.1%
Wave2										9	0.0%

(b)

Fig. 6. Confusion matrix. (a) Results on our dataset, (b) Results on standard Weizmann Dataset.

Conclusion. We propose a method for action recognition technique which uses motion and shape features. Spatio-temporal patterns generated by different actions clearly highlight the differences between them. Results of our technique are better than some of the recently reported results [5]. We have successfully shown that spatio-temporal features consisting of motion and shape patterns can be used for action recognition with stable background.

References

1. Gavrila, D.M.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. 73(1), 82–98 (1999)
2. Sullivan, J., Carlsson, S.: Recognizing and tracking human action. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 629–644. Springer, Heidelberg (2002)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV 2005: Proceedings of the Tenth IEEE International Conference on Computer Vision, Washington, DC, USA, pp. 1395–1402. IEEE Computer Society, Los Alamitos (2005)
4. Niu, F., Abdel-Mottaleb, M.: View-invariant human activity recognition based on shape and motion features, pp. 546–556 (December 2004)
5. Goodhart, T., Yan, P., Shah, M.: Action recognition using spatio-temporal regularity based features, 745–748 (31 2008 - April 4 2008)
6. Danafar, S., Gheissari, N.: Action recognition for surveillance applications using optic flow and SVM. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 457–466. Springer, Heidelberg (2007)
7. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision (1981)
8. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vision 57(2), 137–154 (2004)
9. Freund, Y., Schapire, R.: A short introduction to boosting. J. Japan. Soc. for Artif. Intel. 14(5), 771–780 (1999)

A Novel Hybrid Pansharpening Method Using Contourlet Transform

Tanish Zaveri[1] and Mukesh Zaveri[2]

[1] EC Department, Nirma University, Ahmedabad, India
ztanish@nirmauni.ac.in
[2] Computer Engineering Department,
Sardar Vallabhbhai National Institute of Technology Surat, India
mazaveri@coed.svnit.ac.in

Abstract. In this paper hybrid multispectral image fusion method using contourlet transform is proposed which provides novel tradeoff solution to preserve both spectral and spatial information simultaneously in resultant fused image. Standard Pan-sharpening methods do not allow control of the spatial and spectral quality of the fused image. The color distortion is also most significant problem in standard pan-sharpening methods. Proposed method is applied on number of priori geometrically registered Panchromatic and Multispectral images and simulation results are compared with standard image fusion parameters. The proposed method simulation results also compared six different standard and recently proposed Pan sharpening methods. It has been observed that simulation results of our proposed algorithm is consistent and preserves more detailed spatial and spectral information and better visual quality compared to earlier reported methods.

1 Introduction

In recent years image fusion algorithms are used as effective tools in medical, remote sensing, industrial automation, surveillance, and defense applications. According to Piella [1], fusion process is nothing but a combination of salient information in order to synthesize an image with more information than individual image and synthesized image is more suitable for visual perception. In this paper, we focus on multispectral image fusion application which is an important research area in the field of remote sensing. The synthesis of multispectral (MS) images to the higher spatial resolution of the Panchromatic (Pan) image is called as Pan sharpening method. Various pan sharpening methods have been reported earlier [1][2][3][4][5]; the comprehensive review of most published image fusion techniques described by Pohl and Van Genderen [2]. The natural images contain many intrinsic geometrical structures. In such images, if we apply wavelet transform in 2D than it isolates edges in the images with discontinuities at edge points and smoothness of the boundaries of object will loss. To overcome these limitations, in this paper a novel hybrid multispectral image fusion algorithm based on contourlet transform is proposed. The proposed method is combination of new modified IHS method and Contourlet transform.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 363–368, 2009.

2 Proposed Method

The proposed method is a novel framework which provides novel tradeoff solution to get better spectral and spatial quality Pan sharpened image. The block diagram of proposed method is shown in Fig. 1(a) & (b). Both MS and Pan image are considered as input source image. The IHS color space is used to apply proposed algorithm because of it has less computational complexity and more practical applications. In the proposed method, to increase spectral component while preserving spatial details both input intensity images are considered as source images. The block diagram to generate modified intensity image I_new is shown in Fig. 1. In the block diagram Histogram Equalization (HE), average (avg) and match measure operations are performed. The block diagram of proposed method is shown in Fig. 2. In proposed method both pixel based and region based hybrid image fusion rule is used so this method is called as hybrid method in this paper. It is applied to preserve more details in final Pan sharped image and also it carries advantages of both types of fusion methods. The Pansharped image can be generated by following steps as described below.

Step 1. The contourlet transform (CT) [11] applied to the I_new and Im images which decomposes approximation and detail components from each source images which are represented as $I_{ACT,j}$ and $I_{DCT,j}$ respectively. Where j represents decomposion level of CT.

Step 2. Hybrid fusion rule is divided into three categories; pixel based, block processing based and region based. The hybrid fusion rule is applied on CT based decomposed detail image $I_{nDCT,j}$ and $I_{mDCT,j}$ of both source image I_new and Im respectively. The four fusion rules are designed based on four important features parameters contrast, energy, standard deviation and average gradient as explained in [15]. Among these four fusion rules one fusion rule is pixel and region based each and other two fusion rules are block processing based. All the four fusion rules are applied on $I_{nDCT,j}$ and $I_{mDCT,j}$.

The fusion rule 2 and 3 are energy and standard deviation feature extraction parameter based respectively. One region based image fusion rule 4 with average

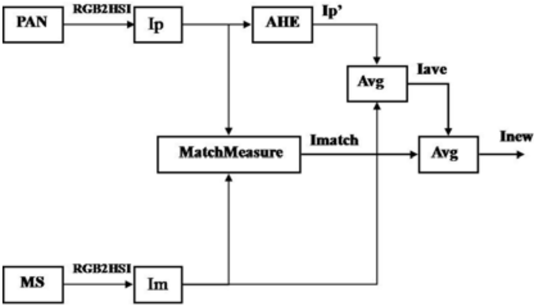


Fig. 1. Block Diagram of Proposed Modified IHS method

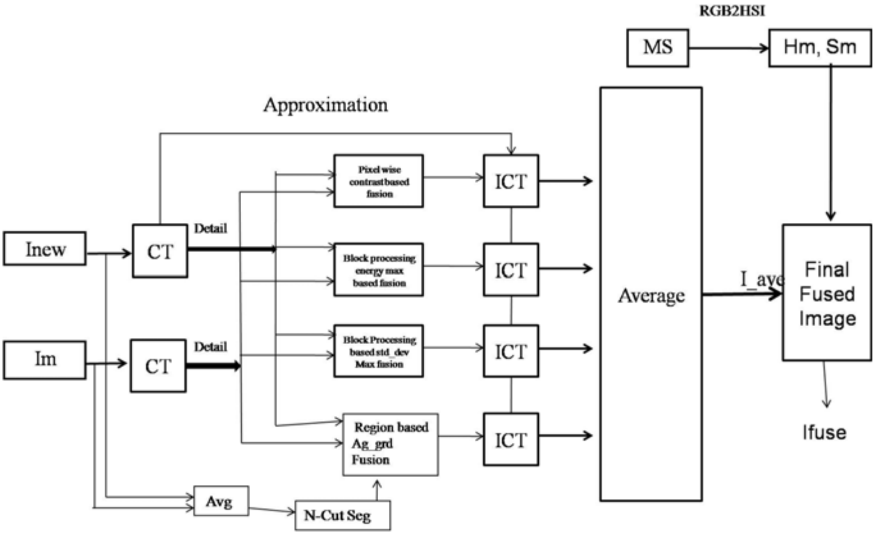

RGB2HSI
MS
Hm, Sm
Approximation
Pixel wise contrast based fusion
Block processing energy max based fusion
Block Processing based std_dev Max fusion
Region based Ag_grd Fusion
ICT
ICT
ICT
ICT
Average
I_avg
Final Fused Image
Ifuse
Inew
CT
Detail
Im
CT
Detail
Avg
N-Cut Seg

Fig. 2. Block Diagram of Proposed method

gradient parameter is applied. All four fusion rules are described in (1)(2)(3)(4). Fusion rule 1 is pixel based fusion rule based on contrast parameter. The contrast gives information about differences of objects in the image and in this contrast max fusion rule high directive contrast coefficients are preserved in final fused image. First fusion rule is based on contrast parameter first which is defined as

$$I^d_{FC,j} = \begin{cases} I^d_{DCT,j} & \text{if } C^d_{n,j} \geq C^d_{m,j} \\ I^d_{DCT,j} & \text{if } C^d_{n,j,k} < C^d_{m,j} \end{cases} \tag{1}$$

The fusion rule 2 is block processing based energy max rule for window size (M x N). It can also call as directive energy fusion rules. The block of pixels whose directive energy is maximum are saved to be the pixels of the fused image.

$$I^d_{FE,j} = \begin{cases} I^d_{nDCT,j} & \text{if } E^d_{n,j} \geq E^d_{m,j} \\ I^d_{mDCT,j} & \text{if } E^d_{n,j} < E^d_{m,j} \end{cases} \tag{2}$$

The directive SD based image fusion rule 3 is defined as

$$I^d_{FS,j} = \begin{cases} I^d_{nDCT,j} & \text{if } S^d_{n,j} \geq S^d_{m,j} \\ I^d_{mDCT,j} & \text{if } S^d_{n,j} < S^d_{m,j} \end{cases} \tag{3}$$

The fusion rule 4 is region based image fusion rule based on average gradient which is used to compare spatial resolution or clarity of information from the images or regions. The average mean gradient is computed as described in equation (5). The fusion rule 4 is defined as

$$I^d_{g,i,j} = \begin{cases} Rn_i & \text{if } G^d_{n,i,j} \geq G^d_{m,i,j} \\ Rm_i & \text{if } G^d_{n,i,j} < G^d_{m,i,j} \end{cases} \tag{4}$$

Here i is a number of regions and it varies from 1 to n. where $i = 1, 2, ..., n$. Normalized cut set segmentation algorithm is applied on I_new. Segmented regions extracted from image I_new and Im are represented as Rn_i and Rm_i respectively which is extracted using segmentation results of I_new. After applying fusion rule 4 to all n regions; merge all the regions to generate resultant fused image $I^d_{g,i,j}$.

Step 3. Repeat first two steps for all the CT decomposed detail images.

Step 4. Four detail images produced after applying four fusion rules and for all the four type of detail image; $I^d_{nACT,j}$ is considered as common approximation image to apply inverse CT as shown in Fig. 2(a).

Step 5. After applying inverse CT resultant four fused images are averaged to produce final fused intensity image called as $I^d_{F,j}$.

Step 6. Finally, the H and S components of MS image is combined with the $I^d_{F,j}$ intensity image to obtain the final fused RGB image.

All four important activity level measurement feature parameters are considered to compare the details from both source images. Hybrid fusion rule is proposed in the paper to take an advantage of both pixel based and region based image fusion rules. The next section describes evaluation criteria to compare results of different Pan sharpening methods in brief.

3 Simulation Results and Assessment

The test dataset images are downloaded from [9]. The IKONOS-2 images covering an area of the city of Sherbrooke, QC, Canada, also considered as input source images as shown in Fig. 3 (a) and (b) are Pan and MS images respectively [4]. The proposed algorithm has been implemented using Matlab 7. The raw multispectral image taken from the site has been resampled to the same size of the panchromatic image in order to perform registration. Our experiment results show that CT decomposition level 3 provides better visual quality.

Nine segmentation regions are considered to apply region based image fusion rule 4. This value is considered after analyzing different results of different segmentation levels. The most widely used three standard Pan sharpening methods IHS method [2], Brovey Method [7], PCA based method [12] in remote sensing area described in [2] and two recently proposed multiresolution based wavelet transform (WT) [8] and CT based [9] substitution methods are used to compare proposed method.

Average value of each quality assessment parameters of all three band R,G and B of source images are depicted in Table 1. for six comparison methods. All the spectral based fusion parameters are better for proposed method. Spatial quality parameters are better for PCA based method. The average correlation

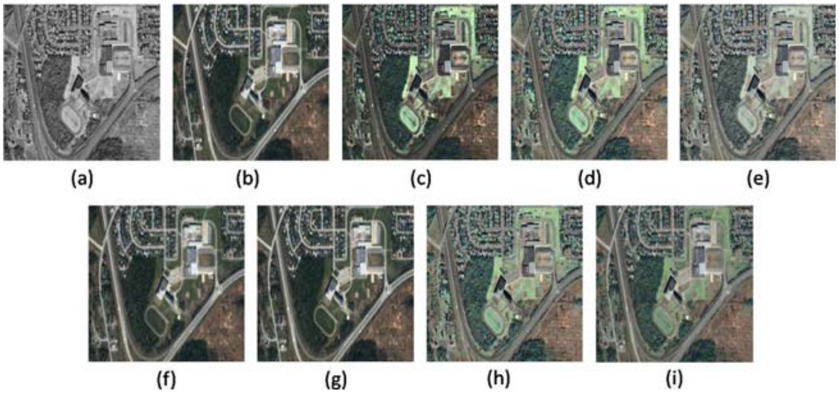


Fig. 3. Fusion Results of IKONOS2 image (a) 1m Panchromatic image (b) 4m Multispectral image(c) IHS Method (d) Modified IHS Method (e) PCA method (f) WT method (g) CT Method (h) Brovey Transform Method (i) Proposed Method

Table 1. Image Fusion Quality Assessment Parameters for Comp Images

Comp	Spectral				Spatial			Common				
	SNR	CC	DE	SAM	SNR	CC	DE	Avg.CC	AG	SD	RCE	MCE
I.H.S. [2]	52.807	0.297	40.196	0.547	56.668	0.926	33.204	0.611	56.374	45.783	0.137	0.102
MI-I.H.S.[7]	52.693	0.353	45.012	0.475	60.932	0.973	11.183	0.663	53.403	43.129	0.128	0.101
PCA [12]	53.716	0.358	48.907	0.460	65.319	0.990	5.359	0.674	61.143	42.999	0.189	0.175
Sub. WT [8]	56.735	0.889	15.355	0.218	54.911	0.519	45.026	0.704	59.357	47.545	0.200	0.197
CT [9]	57.624	0.928	11.986	0.177	54.710	0.452	47.037	0.690	52.335	47.890	0.219	0.202
Brovey [2]	53.151	0.447	41.343	0.434	60.876	0.964	9.695	0.705	50.289	39.822	0.147	0.113
Proposed	**56.028**	0.836	19.629	0.001	55.589	0.693	40.167	**0.765**	51.295	44.584	0.237	0.195

coefficient and average SNR are significantly better for proposed method which shows that proposed method preserves both spatial and spectral parameters better than other reported methods. Proposed method provides novel tradeoff solution. Multiresolution based WT and CT based methods have less color distortion but spatial resolution is affected. The color distortion also very less in proposed method while color distortion is highest in IHS and PCA based method. Resultant Pan sharped image of all seven method are shown in Fig. 3 (c) to (i).

4 Conclusion

There are number of applications in remote sensing that require images with both spatial resolution with less color distortion. The fusion of multispectral and panchromatic images provides a solution by combining the clear geometric features of the panchromatic image and color information of the multispectral image. The proposed algorithm is novel method which uses contourlet transform and hybrid fusion rule framework. Due to this framework, the visual quality and fusion quality assessment parameters of resultant image are significantly better than earlier reported method. The SNR and average CC are remarkably higher than other compared standard and recent methods. The algorithm can be extended by applying artificial intelligent method for more robust fusion. The computational time is only limitation of the algorithm.

References

[1] Piella, G.: A general framework for multiresolution image fusion: from pixels to regions. Journal of Information Fusion 4(4), 259–280 (2003)

[2] Pohl, C., Van Genderen, J.: Multisensor image fusion in remote sensing: concepts, methods, and applications. International Journal of Remote Sensing 19(5), 823–854 (1998)

[3] Schowengerdt, R.A.: Remote Sensing: Models and Methods for Image Processing, 2nd edn. Academic, Orlando (1997)

[4] Wang, Z., Ziou, D., Armenakis, C., Li, D., Li, Q.Q.: A Comparative Analysis of Image Fusion Methods. IEEE Trans. Geosci. Remote Sens. 43(6), 1391–1402 (2005)

[5] Zhang, Y.: Understanding Image Fusion. PCI Geomatics 24 (2008)

[6] Zhou, J., Civco, D.L., Silander, J.A.: A wavelet transform method to merge Landsat TM and SPOT panchromatic data. International Journal of Remote Sensing 19(4), 743–757 (1998)

[7] Tu, T., Su, S., Shyn, H., Huang, P.: A new look at IHS like image fusion methods. Information Fusion 2, 177–186 (2001)

[8] Lez-Audi, G., Cana, M., Saleta, J.L., Catala, N.R.G., Garcia, R.: Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Transactions on Geoscience and Remote Sensing 42, 1291–1299 (2004)

[9] Aboubaker, M., Jaily, A.L.E., Ibrahim, A., El Rube Mohab, A., Mangoud: Fusion of Remote Sensing Images Using Contourlet Transform. In: Proceeding Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, pp. 213–218. Springer, Heidelberg (2008)

[10] Behnia, P.: Comparison Between Four Methods for Data Fusion of ETM+ Multispectral and Pan Images. Journal of Geo-spatial Information 8(2) (2005)

[11] Miao, Q., Wang, B.: A novel image fusion method using contourlet transform. In: International Conference on Communications, Circuits and Systems Proceedings, vol. 1, pp. 548–552 (2006)

[12] Tania, S.: Image Fusion: Algorithms and Applications, 1st edn. Elsevier, Amsterdam (2008)

[13] Gonzales, R., Richard, W.: Digital Image Processing, 2nd edn. Pearson Education (2006)

[14] Shutao, L., Bin, Y.: Multifocus image fusion using region segmentation and spatial frequency. Image and Vision Computing 26, 971–979 (2008)

[15] Zaveri, T., Zaveri, M., Makwana, I.: A Novel Hybrid Multispectral Image Fusion Method using Contourlet Transform. In: IEEE International conference TENCON 2009, Singapore, November 23-26 (2009)

[16] http://studio.gge.unb.ca/UNB/zoomview/examples.html

Text Line Segmentation for Unconstrained Handwritten Document Images Using Neighborhood Connected Component Analysis

Abhishek Khandelwal[1], Pritha Choudhury[1], Ram Sarkar[2], Subhadip Basu[2], Mita Nasipuri[2], and Nibaran Das[2]

[1] CSE Department, Sikkim Manipal Institute of Technology, Sikkim, India
[2] CSE Department, Jadavpur University, Kolkata, India

Abstract. Text line extraction is the first and one of the most critical steps in optical character recognition (OCR) of unconstrained handwritten documents. The present work reports a new methodology based on comparison of neighborhood connected components to determine whether they belong to the same text line. Components which are very small or very large compared to the average component height are ignored in the preprocessing step. During post-processing, such components are reconsidered and allocated to the lines to which they most suitably belong. The performance of the developed technique is evaluated on the benchmark training dataset for the ICDAR 2009 handwriting segmentation contest. The dataset consists of English, French, German and Greek handwritten texts. The overall text line identification accuracy on the mentioned dataset is observed to be around 93.35%.

Keywords: Text line identification, handwritten script, neighborhood connected component analysis.

1 Introduction

Identification of text lines is the first and most important step in the process of optical character recognition (OCR) of handwritten document images. If line identification is not accurate, (for example, if two or more lines are merged) then none of the words and consequently none of the characters in the constituent lines can be identified correctly. The same problem occurs if a line is erroneously split into two or more parts. Such scenarios are unacceptable for large-scale recognition of handwritten documents.

Previous techniques using Hough transform consider a set of points of initial image as input while the lines that fit best to these points are calculated. These methods vary in the set of points considered for voting procedure, viz. gravity centers [1] or minima points [2] of the connected components. A recent block based Hough Transform method [3] takes into account gravity centers of parts of CCs.

Smearing methods include the Fuzzy RLSA [4], in which the value of each pixel is the sum of all pixels in the original image within a specified horizontal distance. Adaptive RLSA [5] evolves from Classical RLSA [6] and uses additional smoothing

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 369–374, 2009.

constraints in regard to the geometrical properties of neighboring CCs. Another technique that uses morphological operations and RLSA to segment text lines from unconstrained handwritten document images is described in [7].

An MST based clustering technique [8] with distance metric learning has been suggested for text line identification in Chinese handwritten documents. A technique described in [9] makes use of Mumford-Shah (MS) model; line segmentation in this technique is achieved by minimization of MS energy. A technique described in [10] uses density estimation and level set methods for text line extraction. A novel technique using hypothetical water flows at specific angles from both sides of the document image for text line extraction is described in [11].

In this paper, we present an effective alternative technique for identifying text lines in handwritten documents. As a first step, we have implemented an eight-way connected-component-labeling (CCL) algorithm to identify the most basic elements in the text document as unique objects. During preprocessing, we ignore components identified as noise. Some very large components might be formed as a result of overlap of two or more elements belonging to adjacent lines. Such components are split. Thus, a final set of components is derived for text line identification. Post-processing steps include reconsidering the components previously ignored as noise. Some of these might have been actually been small handwritten parts of text and such components are allocated to suitable lines.

2 The Present Work

The scanned handwritten document image is first binarized, with each foreground data pixel is represented by label ‘1’ and each background pixel is represented by label ‘0’. All the following steps are then implemented on these binarized data file.

2.1 Preprocessing of Scanned Document Images

Preprocessing of the text document in order to phase out noise and outlier data is very important for identification of text lines correctly. During scanning of handwritten documents, many small dot-like noisy elements may appear on the scanned image. It has been observed that these elements are usually of a size much smaller than the smallest components in the handwritten documents. So, two threshold values, T_1 subjective to the height of the document image, and T_2 subjective to the width of the document image have been set. A component with a height less than T_1 and width less than T_2 is ignored, while creating the set of uniquely labeled components to be considered for text line identification. In this process, some very small components belonging to the actual handwritten text might also get ignored erroneously, but they shall be reconsidered and merged into suitable text lines during refinement of the identified text lines at a later stage of the algorithm.

Another type of noise that may appear on the document image while scanning, is long lines at the edges of the documents. These lines shall interfere in the implementation of the current work severely, and hence must be removed while preprocessing the document image. Components having a height greater than 1/4th the page height or width greater than 1/3rd the page width are identified as such noise elements and are ignored for subsequent processing.

2.2 Analysis of Identified Components

An eight-way Connected Component Labeling (CCL) algorithm is implemented to uniquely label and build the set of components to be considered for text line identification. For analysis of identified components, first the average component height H_{Cavg} and average component width W_{Cavg} is computed as follows:

Let N be the total number of components, then the component height takes values from the finite data set x_1, x_2, …, x_N. Let the arithmetic mean of the heights of the components be μ. Then,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$

Only components with height ranging from μ-σ to $\mu+\sigma$ are considered for calculating the final average component height H_{Cavg}, which is the arithmetic mean of all such values. The average component width W_{Cavg} is computed in a similar fashion.

It has been observed in some cases of handwritten text that a character belonging to a particular line may touch another character belonging to an upper or lower line. These two characters are identified as a single component by CCL algorithm, and have a height much larger than the average height H_{Cavg} which is computed using the above mentioned method. Hence it becomes necessary that this component be split into its constituent characters so that two or more lines are not merged in the step of line identification. In the present work, components having a height greater than $\delta*H_{Cavg}$ have been split horizontally into two equal halves; where δ is a heuristically chosen tuning parameter. Here it has been assumed that not more than two consecutive text lines touch each other. Thus a final set $\{S\}$ of components is constructed for formation of lines from the scanned document image.

2.3 Neighborhood Formation and Text Line Identification

For each component C_i, a general neighborhood N_{Ci} is defined, and a component under consideration looks for candidate components in its neighborhood to determine whether any such component belongs to the same line as the current component. The neighborhood N_{Ci} is defined with respect to the average component width W_{Cavg} and average component height H_{Cavg} computed previously. It is a rectangular area of width $\alpha*W_{Cavg}$ and height $\beta*H_{Cavg}$ with the dimensional center of the component C_i as the centre of the rectangle; where α and β are experimentally chosen tuning parameters.

Each component C_i within the set $\{S\}$ is considered in sequence. Let $\{C_{iN}\}$ be the set of components in the neighborhood N_{Ci} of C_i. The component C_i looks for neighborhood components which satisfy at least one of the following conditions:

i) C_i spans at least a fraction η_1 of the height of C_{ij}; j ϵ N, height wise or vice versa.

ii) The height-wise midpoints of C_i and C_{ij}; j ϵN, have a vertical distance less than a threshold η_2 of the height of C_i or C_{ij}.

Here η_1 and η_2 are experimentally chosen tuning parameters. A component C_{ij}, found to satisfy either or both of these conditions, is allocated to the line to which C_i belongs. The bounds of the line to which C_i belongs, are initialized with the corresponding bounds of C_i and the area of the line keeps increasing as more components are allocated to the same line. A component *Cij*, which has already been allocated to a line, shall not be considered for future comparisons with any other component for identification of text lines.

While forming the set $\{S\}$ of components for text line identification we have ignored some small components which were considered to be noisy data. Such components may be ascendants/descendants or very small characters belonging to text lines thus identified. We thus consider all such components one by one and try to allocate them to the lines to which they most suitably belong.

For a particular text line, we consider all such components which lie partially or completely within its bounding box. We then label these components with the same label as that of the text line. This step fairly allocates small components into the text lines of which they should be a part.

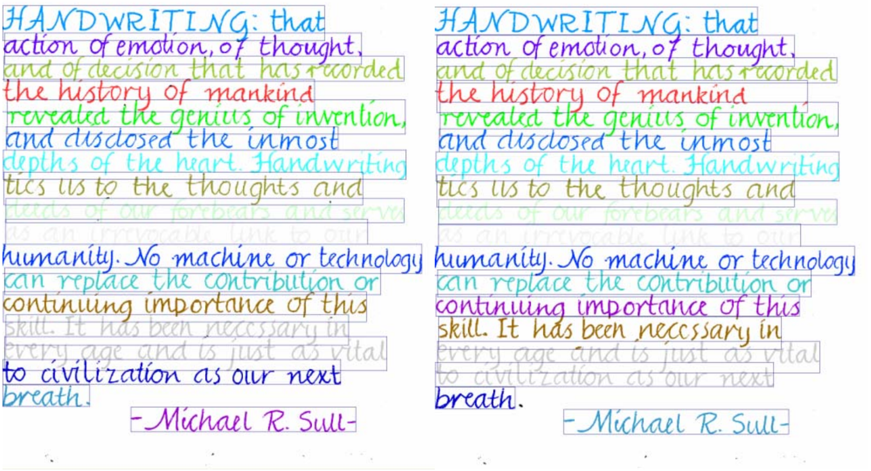
3 Experimental Results

We have adopted a means of performance evaluation which has already been used in ICDAR 2003, 2005 and 2007 page segmentation contests, and ICDAR 2007 and 2009 handwriting segmentation contests [3, 5].

In this method, the segmentation result is compared to an already annotated ground truth. It is based on counting the number of matches between the generated segmentation results and the existing ground truth. The performance of our present work was evaluated on the training set of handwritten documents used in ICDAR handwriting segmentation contest 2009. The test set consisted of 100 pages of handwritten texts in English, French, German, and Greek. These documents were either historical documents or modern texts written by different individuals, not containing any images, logos, etc. The text in most historical documents was highly fragmented. Most of these handwritten documents had skewed/curvy text lines. For all these document images, we had the ground truth data files for comparison with our segmentation results. The documents were scanned at 300dpi and had a typical resolution of 2500x2000 pixels.

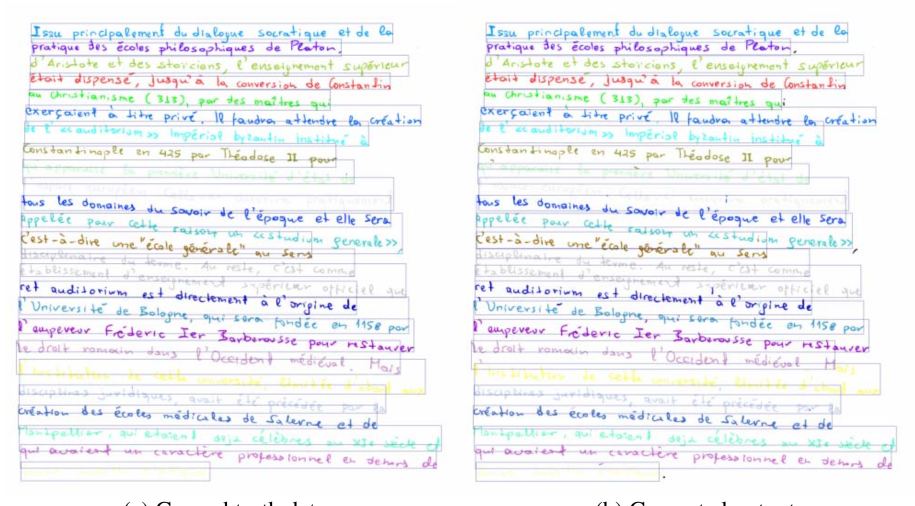
The experimental values for the tuning parameters, mentioned in our technique, were set to: $\delta = 2.7$, $\alpha = 7$, $\beta = 5$, $\eta_1 = 0.6$, $\eta_2 = 0.6$. It is worth noting that our proposed method gives a *Detection Rate* (DR) of 94.0%, *Recognition Accuracy* (RA) of 92.7%, and hence *F-Measure* (FM) of 93.35% which outperforms most of the orthodox approaches for text line identification in handwritten documents.

Figs. 1-2 show results of the developed technique on sample test pages, along with the ground truth data. The bounding boxes of the text lines have been highlighted for ease of identification of the same.


HANDWRITING: that
action of emotion, of thought,
and of decision that has recorded
the history of mankind
revealed the genius of invention,
and disclosed the inmost
depths of the heart. Handwriting
ties us to the thoughts and
deeds of our forebears and serves
as an irrevocable link to our
humanity. No machine or technology
can replace the contribution or
continuing importance of this
skill. It has been neccssary in
every age and is just as vital
to civilization as our next
breath.
-Michael R. Sull-

(a) Ground truth data (b) Generated output

Fig. 1. Identification results of the developed technique on a page with very close text lines



(a) Ground truth data (b) Generated output

Fig. 2. Identification results of the developed technique on a page with skew and extraneous ascendants/descendants

4 Conclusion

In this paper, we have developed an effective technique for identification of text lines in handwritten documents. The method uses a novel approach of height-specific dimensional comparison of neighborhood components to classify them into text lines. A post-processing step based on a Euclidean distance metric efficiently classifies small components into text lines to which they most suitably belong. Although the performance of the developed methodology was evaluated on English, French, Greek and German, it can also be applied to various other scripts.

Future work may involve segmentation of text lines into constituent words. It also concerns improvement of the technique used for splitting components into constituent words/characters belonging to different lines. Inclusion of a skew-correction step just before the actual identification of text lines would improve the performance of the key modules by a significant amount. Small components ignored during preprocessing can be allocated to suitable lines using an improved technique. In a nutshell, the work presented here outlines the technicalities of an effective method for identification of text lines from unconstrained handwritten document pages.

Acknowledgement

Authors are thankful to the CMATER, SRUVM, CSE Department, Jadavpur University, for providing infrastructural facilities during progress of the work.

References

1. Likforman, L., et al.: A Hough based algorithm for extracting text lines in handwritten documents. In: Proc. of the Third ICDAR, Montreal, Canada, pp. 774–777 (1995)
2. Pu, Y., et al.: A natural learning algorithm based on Hough transform for text lines extraction in handwritten documents. In: Proc. of the 6th IWFHR, pp. 637–646 (1998)
3. Louloudis, G., et al.: A block-based Hough transform mapping for text line detection in handwritten documents. In: The 10th IWFHR, France, October 2006, pp. 515–520 (2006)
4. Shi, Z., et al.: Line separation for complex document images using fuzzy run-length. In: First International Workshop on Document Image Analysis for Libraries, p. 306 (2004)
5. Gatos, B., et al.: ICDAR2007 Handwriting Segmentation Contest. In: the Ninth ICDAR, Curitiba, Brazil, September 2007, pp. 1284–1288 (2007)
6. Wahl, F.M., et al.: Block segmentation and text extraction in mixed text/image documents. Computer Graphics and Image Processing 20, 375–390 (1982)
7. Roy, P.P., et al.: Morphology Based Handwritten Line Segmentation Using Foreground and Background Information. In: Proc. of ICFHR, Canada, pp. 241–246 (2008)
8. Yin, F., et al.: Handwritten Text Line Segmentation by Clustering with Distance Metric Learning. In: Proc. of ICFHR, Canada, August 91-21, pp. 229–234 (2008)
9. Du, X., et al.: Text Line Segmentation in Handwritten Documents Using Mumford-Shah Model. In: Proc. of ICFHR, Canada, August 91-21, pp. 253–258 (2008)
10. Li, Y., et al.: Script-Independent Text Line Segmentation in Freestyle Handwritten Documents. IEEE Transactions on PAMI 30(8), 1313–1329 (2008)
11. Basu, S., et al.: Text line extraction from multi-skewed handwritten documents. Pattern Recognition 40(6), 1825–1839 (2007)

Model-Guided Segmentation and Layout Labelling of Document Images Using a Hierarchical Conditional Random Field

Santanu Chaudhury, Megha Jindal, and Sumantra Dutta Roy

Dept of Electrical Engg, IIT Delhi, Haux Khas, New Delhi - 110 016, India
{schaudhury,megha1jindal}@gmail.com, sumantra@cse.iitd.ac.in

Abstract. We present a model-guided segmentation and document layout extraction scheme based on hierarchical Conditional Random Fields (CRFs, hereafter). Common methods to classify a pixel of a document image into classes - text, background and image - are often noisy, and error-prone, often requiring post-processing through heuristic methods. The input to the system is a pixel-wise classification based on the output of a Fisher classifier based on the output of a set of Globally Matched Wavelet (GMW) Filters. The system extracts features which encode contextual information and spatial configurations of a given document image, and learns relations between these layout entities using hierarchical CRFs. The hierarchical CRF enables learning at various levels - 1. local features for text, background and image areas; 2. contextual features for further classifying region blocks - title, author block, heading, paragraph, etc.; and 3. probabilistic layout model for encoding global relations between the above blocks for a particular class of documents. Although the work has been motivated for an automated layout analyser and machine translator for technical papers, it can also be used for other applications such as search, indexing and information retrieval.

1 Introduction

Automatic segmentation and layout analysis of documents can be used for interpretation and machine translation of technical documents, search and information retrieval, in general. Common approaches have often been heuristic in nature, for instance [1]. A learning-based approach is more general than using assumptions about document layouts. Further, this can make it tunable for a particular class of documents. An earlier work [2] presents a learning-based scheme for extraction of text, image and background pixels using a learning-based approach - globally matched wavelets (GMWs, hereafter), and a Markov Random Field (MRF, hereafter) model for smoothing the results. This works at a pixel level, and does not consider the problem of layout analysis. Knowledge of the layout itself can remove page segmentation errors. Shafait et al. [3] present a statistical learning-based mechanism for layout analysis. They overcome the exponential computational cost of optimal geometric parsing of methods relying on probabilistic grammars [4], [5], [6]. They model a page as a mixture of layout

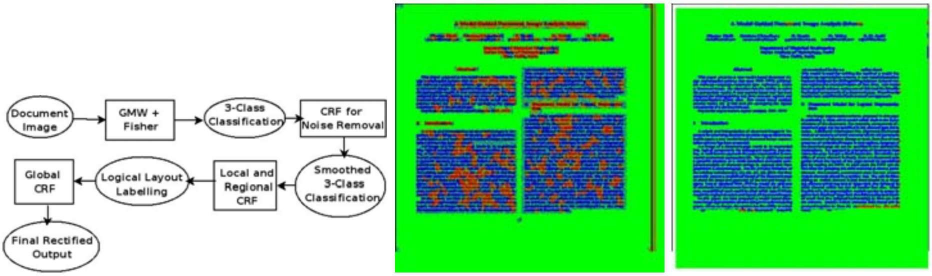
S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 375–380, 2009.

structures, and use a probabilistic matching algorithm for find the most probable layout. Our work here uses a much more general structure - conditional random fields (CRFs, hereafter), which avoid the limitations of generative models such as MRFs [7]. MRF-based layout modelling [8] and generative zone models [9] typically need large amounts of labelled training data [10].

Xuming He et al. [7] propose the use of multi-scale CRFs for image labelling. CRFs have been used for document segmentation and labelling e.g., Shetty et al. [11]. The authors use a CRF with simple features such as height of a patch, component width, density, etc. In contrast, our work uses a unified CRF learning-based approach to document layout segmentation at three different levels:

1. A CRF framework for filtering a Fisher classifier output on GMW filters applied on document images.
2. Contextual features for classifying text blocks, and
3. A CRF-based probabilistic method for encoding global relations between the above blocks, for different document classes.

Fig. 1(a) gives a graphical overview of the ideas in this paper. To the best of our knowledge, no other work encompasses a general learning-based procedure at all levels of a segmentation and layout understanding task. Sec. 2 gives an overview of the GMW-based pixel-wise classification of the image. In the next section, we enumerate the first level of application of CRFs - to the output of the Fisher classifier of Sec. 2, for smoothing/rectification. Secs. 3 and 4 work on the idea of CRFs for regional and global features. We then show representative results of extensive experimentation, and list some of our planned future extensions.



(a) (b)

Fig. 1. (a) An overview of our multi-level CRF learning-based document segmentation and layout analysis method, and (b) GMW-based pixel-level Fisher classification of a document page, and its CRF-filtered output: Sec. 2.1

2 Pixel Classification Using Globally Matched Wavelets

We use the outputs of globally matched wavelet (GMW) filters which learn parameters to classify a pixel as being from a text, background or image region - at different scales (See our earlier work [2] for details.) The GMW filter responses are features, which we pass through three Fisher classifier, optimised for a two-class problem (Text-Image, Image-Background, and Text-Background classification). Each classifier gives a confidence value. We combine the outputs of all above classifiers to make the final decision. Independently at each pixel, the Fisher classifier produces a distribution over different class variable given filter outputs which is often noisy due to the overlap between neighboring classes. The next section describes a novel CRF based learning procedure for removing this noise.

2.1 CRF-Based Rectification of Fisher Classified GMW Responses

The input to this module is the output of the Fisher classifier of the previous section (Sec 2). We divide an image into a series of overlapping regions. We define features of size 3 × 3 which encode these error conditions and contextual information. The local content of the image at each pixel is encoded by representing binary values of gray level, the average gray level of neighbouring pixels, the gradient in horizontal and vertical direction and the output of the Fisher classifier. These extracted features are given to CRF for learning. Fig. 1(b) shows a representative example of a document page with pixel-level classification (text in blue, image in red, and background in green), and its corresponding CRF-filtered version. The CRF formulation for this stage is similar to formulation of CRF model based on local and regional image features. (The next section, Sec. 3 discusses this in detail.) *It is important to note that the present CRF-based approach scores over our original MRF-based approach [2] - in encoding contextual information better than a generative model such as an MRF, which typically encodes continuity information. Further, an MRF typically needs a large amount of training labelled data [10].*

3 Hierarchical CRFs for Regional and Global Document Features: Regional Features

We use a hierarchical CRF model to capture both regional and global features [7]. *Two image patches can be indistinguishable at a local scale, but layout relationship can provide the context for correct labelling.* We represent images as rectangular grids of patches by overlaying a grid of fixed size and then associate a hidden class label with each patch. (Fig. 2(a) illustrates the above idea, on a typical 480×640 image, with 8×8 patches. The size of a patch is often a compromise between processing complexity reduction, and labelling quality.) The CRF sequentially combines predictions or probability distributions corresponding to each image patch. (Fig. 2(b) shows templates for some regional features.) The local image content of each patch is encoded using the labels produced by local classifier f_l and position descriptors f_p. The Regional image content of each patch is encoded using regional feature f_r that encode certain patterns within a Image. We use the cell index of the location of a patch, as its position feature.

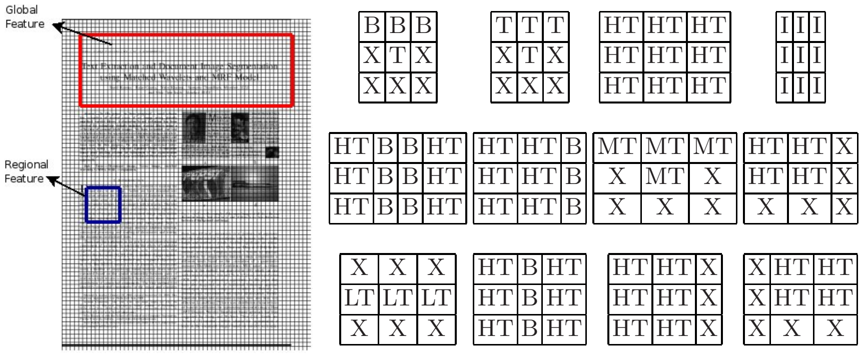


Fig. 2. (a) An example of regional and global features for a typical document, and (b) Templates for some regional features. B: background, T: Medium Text density, I: Image, LT: Low Text density, HT: High Text density, X: Don't care.

These regional feature patterns are matched at each patch in the image. If particular pattern of feature is matched at patch in a image, then value of feature at that patch is taken as 1, otherwise 0.

Each patch is thus coded by these binary vectors along with position descriptor and labels produced by the feature classifier. We define CRF observation functions as linear functions of these binary vectors. These modalities are modelled as being independent given the patch label. Now, we provide a formulation for conditional model for patch labels that incorporates both local patch level features and regional features aggregated over neighbourhood patches. Let $x_i \in \{1, ..., L\}$ denote the label of patch i, y_i denote the $D-$ dimensional concatenated binary indicator vector of its local patch content, position descriptor and regional features. $D = (f_l, f_r, f_p)$. The conditional probability of the label x_i is then modelled as

$$p(x_i = l|y_i) \propto exp\left(-\sum_{d=1}^{D} \alpha_{dl} y_{id}\right)$$

where α_{dl} is a $D \times L$ matrix of coefficients which need to be learnt using CRF training. Thus at each patch we get a distribution over the label variables. The output of this stage classifies and segments out the logical regions of document image but it is somewhat noisy and thus needs some post-processing. The following section describes our global features based CRF learning for the same.

4 CRF Based on Global Document Features

This stage takes the classification output of the previous stage and focus on removing the errors by using global features and probabilistic layout model. For our experiments, we define a hierarchical CRF model with seven output labels, and the parameters are learned on fully labelled images. The output labels are title block, author blocks, headings, background, paragraph, column-separator and figure. *Our models take the global image context into account by including feature functions based on probabilistic layout model of documents.This global CRF aims to remove the ambiguities that arise when patches are classified using local and contextual image features only.* For the domain of technical papers, one needs to learn the probabilistic layout of the region labels using a CRF. These global feature represent the relationship between different output labels e.g., an author block cannot come above the title block. In this way, we tried to remove the ambiguities that could arise. The Image is again considered to be divided into overlapping patches. These global feature are learnt using CRF learning. The posterior distribution over the label variables given the hidden variables, can be written as for regional features. The conditional probability of the label x_i using probabilistic layout model k is modelled as:

$$p(x_i = l|k) \propto exp\left(-\sum_{d=1}^{D} \beta_{dl} k_d\right)$$

where β_{dl} is $D \times L$ matrix of coefficients which we need to learn.

5 Experimental Results and Discussion

We have experimented with a large number of document images of technical papers, of different layouts. We show some representative results in this section. In Fig. 3, we show the three important outputs of our layout analysis procedure: the original image, the image prior to the use of the global CRF, and that after the application of the global CRF. The colour coding is as follows: Blue - Title block, Sky Blue - Author block, Green - Figure block, Red - Background, Pink - Heading, White - Paragraph, and Yellow - Column Separator. For numerous experiments with the system, the layout segmentation is almost always correct, with very few exceptions. We measure the performance of the labelling system

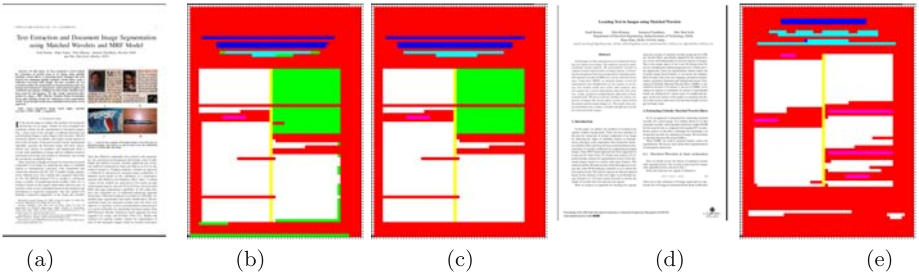


Fig. 3. (a) A document page (b) the corresponding labelled regions before, and (c) after the global probabilistic layout model application. (d) Another document page, and (e) the final segmentation output.

by considering manual ground truth (manually labelled images), and examining ROC parameters - the precision and recall rates. In most case, both these parameters are above 95%.

6 Discussion and Conclusions

This paper presents a novel integrated scheme for a general CRF-based learning framework for document image segmentation and layout analysis at many different logical levels. This starts from filtering and smoothing of the output of Fisher classified GMW outputs, and goes up to learning regional and global features, and their inter-relationships, which further aid in top-down layout analysis and document image segmentation. An interesting extension of our work will be to examine the use of Markov field aspect models [12] and CRFs [13] for a first-level pixel labelling which may be incorrect, or incomplete.

References

1. Gupta, G., Niranjan, S., Shrivastava, A., Sinha, R.M.K.: Document Layout Analysis & Classification and its Application in OCR. In: Proc. IEEE EDOCW (2006)
2. Kumar, S., Gupta, R., Khanna, N., Chaudhury, S., Joshi, S.D.: Text Extraction and Document Image Segmentation. IEEE Transactions on Image Processing 16(8), 2117–2128 (2007)
3. Shafait, F., van Beusekom, J., Keysers, D., Bruel, T.M.: Background Variability Modeling for Statistical Layout Analysis. In: Proc. ICPR (2008)
4. Kanungo, T., Mao, S.: Stochastic language models for style-directed layout analysis of document images. IEEE Transactions on Image Processing 12(5) (2003)
5. Shilman, M., Liang, P., Viola, P.: Learning Non-generative Grammatical Models for Document Analysis. In: Proc. IEEE ICCV, pp. 962–969 (2005)
6. Tokuyasu, T., Chou, P.A.: Turbo Recognition: A Statistical Approach to Layout Analysis. In: Proc. SPIE Document Recognition and Retrieval, pp. 123–129 (2001)
7. He, X., Zemel, R.S., Carreira-Perpinan, M.A.: Multiscale Conditional Random Fields for Image Labeling. In: Proc. IEEE CVPR, pp. II:695–II:702 (2004)
8. Liang, J., Haralick, R.M., Phillips, I.T.: A Statistically based, Highly Accurate Text-Line Segmentation Method. In: Proc. ICDAR, pp. 551–555 (1999)
9. Gao, D., Wang, Y., Hindi, H., Do, M.: Decompose Document Image using Integer Linear Programming. In: Proc. ICDAR, pp. 397–401 (2007)
10. Nicolas, S., Dardenne, J., Paquet, T., Heutte, L.: Document Image Segmentation using a 2D Conditional Random Field Model. In: Proc. ICDAR, pp. I:407–I:411 (2007)
11. Shetty, S., Srinivasan, H., Beal, M., Srihari, S.: Segmentation and labeling of documents using conditional random fields. In: Proc. SPIE Document Recognition and Retrieval (2007)
12. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: Proc. IEEE CVPR (2007)
13. Verbeek, J., Triggs, B.: Scene segmentation with conditional random fields learned from partially labeled images. In: Proc. NIPS (2008)

Recognition of Numeric Postal Codes from Multi-script Postal Address Blocks

Subhadip Basu[1,*], Nibaran Das[1], Ram Sarkar[1], Mahantapas Kundu[1],
Mita Nasipuri[1], and Dipak Kumar Basu[1,2]

[1] Computer Science & Engineering Department, Jadavpur University,
Kolkata-700032, India
[2] A.I.C.T.E. Emeritus Fellow
{subhadip,nibaran,ramsarkar,mkundu,mnasipuri,
dkbasu}@cse.jdvu.ac.in

Abstract. The objective of the current work is to recognize postal codes written in *Roman*, *Devanagari*, *Bangla* and *Arabic* scripts. In the first stage 25 unique digit patterns are identified from the handwritten numeral patterns of the said four scripts. A script independent unified pattern classifier is then designed to classify any digit pattern of these scripts into one of the 25 classes. In the next stage a rule-based script inference engine infers about the script of the numeric string, that invokes one of the four script specific classifiers. The average script-inference accuracy over a six digit numeric string is observed as 95.1% and the best recognition rates for the four script specific digit classifiers are obtained as 96.10%, 94.40%, 96.45 % and 95.60% respectively.

Keywords: OCR, script-identification, classification, postal automation.

1 Introduction

Postal documents are primarily sorted on the basis of a numeric string, popularly known as PIN *(Postal Identification Number)* code or ZIP *(Zone Improvement Plan)* code. For development of an automated mail sorting system, a key challenge is to interpret the handwritten/printed postal code written in different scripts. In a multilingual country like India with 22 official languages, the postal code is often written in different regional scripts along with the *Roman* script. In the present work, we have attempted to address the problem related to the interpretation of handwritten pin codes of aforementioned four scripts, viz., *Roman*, *Devanagari*, *Bangla* and *Arabic (RDBA).* We first identify the specific script in which the numeric postal code is written and then focus on recognition of that postal code.

Among the related works in this domain, Sinha *et al.* [1] and Roy *et al.* [2] developed similar techniques for word-wise identification of *Roman, Devanagari* and *Bangla* scripts in handwritten textual postal addresses using topological and structural features. However, they have not shown any result on identification of the scripts for the numeric postal codes. In another work, Zhou *et al.* [3] developed a connected

[*] Corresponding author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 381–386, 2009.

component profile analysis technique for separation of *Roman* and *Bangla* script based postal documents. Other works, reported in the literature, related to postal automations [4, 5] do not explicitly address the issue of multiple script identification.

Despite these research contributions, the true issue of multi-script address block interpretation still remains an unsolved problem. This is so because in all these works [2-5], the authors had either assumed that the address blocks, including the numeric postal codes, are written using the same script of the textual address block, or remained silent on the script of the postal codes.

Research contributions on recognition of handwritten numerals [4-10] mostly focus on feature based recognition of isolated handwritten digit samples of a given script using standard classifiers. In one of our earlier works [6], a two-pass feature based approach was designed for recognition of handwritten numerals of Bangla script. In another work [7], a classifier combination scheme was proposed to infer over the decisions taken by two different classifiers on each digit pattern. In one of the recent works, Pal *et al.* [8], used contour based directional features to recognize handwritten numerals of six popular Indian scripts, viz., *Devanagari, Bangla, Telegu, Oriya, Kannada* and *Tamil*. They used six different quadratic classifiers, each for the six different scripts, and obtained good recognition accuracy. In another recent work Wen *et al.* [9], developed a handwritten *Bangla* numeral recognition system for automatic sorting of mails for the Bangladesh Post. Using the principles of *Principal Component Analysis* and *Support Vector Machine*, they achieved high reliability in recognition of handwritten numerals of *Bangla* script, but remained silent over the script identification technique for the said pattern classes.

However, in any postal document the script of the numeric postal codes may vary from the script of the textual address part. More specifically, people often write postal address in two scripts, *i.e.*, the textual parts in regional scripts like *Devanagari*, *Bangla* or *Arabic* and the numeric part including the postal code in the *Roman* script. Therefore, inferring the script of the postal code on the address block may often mislead the script recognition process. This has been one of our key motivations behind the current work, discussed in this paper.

2 The Present Work

It is evident from the earlier discussions that limited research contributions have been reported so far on interpretation of multi-script postal documents. To address these issues we have developed a multi-stage approach for recognition of multi-script postal codes written in *Roman*, *Devanagari*, *Bangla* and *Arabic* scripts (as shown in Fig. 1). In the first stage, 25 unique digit patterns, as shown in Fig. 2, are identified from the handwritten numeral patterns of the said four scripts. A script independent unified pattern classifier is then designed to classify any digit pattern of the *RDBA* scripts into one of the 25 classes. In the next stage, a rule-based script inference engine is designed to infer about the script of the numeric string, based on the recognition decisions obtained in the first stage. This decision on the script of the postal code invokes one of the four script specific classifiers from the multi-script numeral recognition engine. Finally, the chosen pattern classifier is used to recognize normalized, binary digit patterns of the corresponding script.

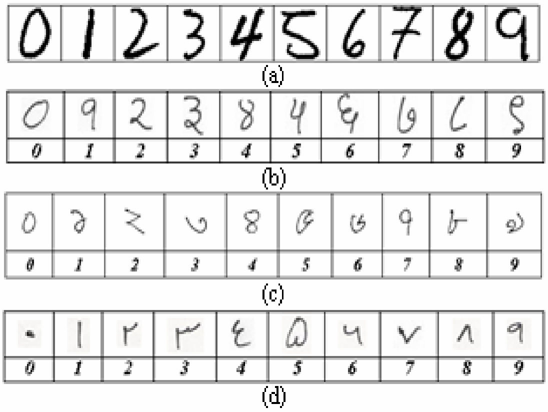


Fig. 1. (a). The decimal digit sets of *Roman* script. ***(b-d)***,The respective digit sets of *Devanagri, Bangla and Arabic* scripts with corresponding labels in Roman script.

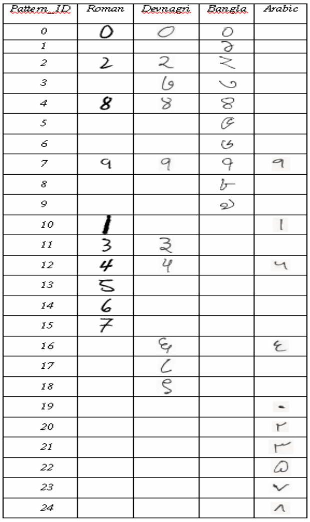

Pattern_ID
Roman
Devnagri
Bangla
Arabic
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Fig. 2. 25 unique digit patterns are identified from the four RDBA scripts

2.1 Design of the Feature Descriptor

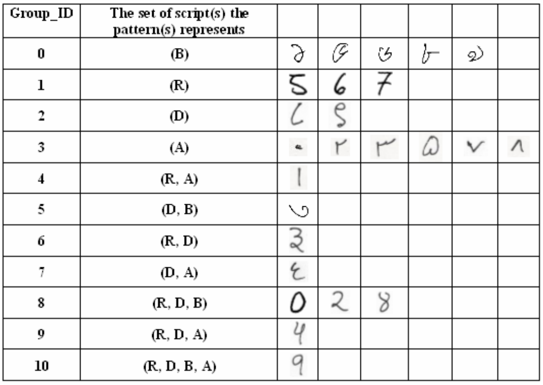
For extraction of the features for both the unified pattern classifier and the multi-script numeral recognition engine, quad-tree based longest-run features are used in the current work. Within a rectangular image region, longest run features are computed in *four directions*, *viz* row wise, column wise and along the directions of two major diagonals. The row wise longest run feature [11] is computed by considering the *sum* of the lengths of the longest bars that fit consecutive black pixels along each of all the rows of the region.

In the current work, we have used a novel modified version of quad tree structure to partition any digit pattern into multiple sub-images. Here, partitioning a digit pattern (or a subpart of it) into 4 regions is done by drawing a horizontal and a vertical line through the *Centre of Gravity* (*CG*) of black pixels in that region. In the current work, we have considered the depth of the quad-tree structure as 2. This generates 4^2, i.e., 16 sub-images at the leaf node positions, thereby resulting in 64 (16 x 4) longest-run features for any digit pattern.

2.2 Design of a Multi-script Pattern Classification Framework

As already mentioned, handwritten digit patterns of the *RDBA* scripts (as shown in Fig. 1) often bear significant similarities in shapes among themselves and we can identify 25 unique pattern shapes, as shown in Fig. 2, from the 40 digit patterns of four different scripts. A multi-layer perceptron based classifier is designed, with the aforementioned features, as a unified pattern classifier for recognizing these 25 different pattern classes.

These unique shapes may represent either a single numeral of any given script, or different numerals of different scripts. Considering these possibilities, 25 unique pattern classes are further classified into 11 groups. Each such group may be viewed as a triplet, {*Group_ID, (Set of unique pattern IDs constituting the group), (Set of identity of scripts the unique patterns represent)*}. Descriptions of the observed 11 groups of patterns are given below which are also illustrated in Fig. 3.



Group_ID	The set of script(s) the pattern(s) represents
0	(B)
1	(R)
2	(D)
3	(A)
4	(R, A)
5	(D, B)
6	(R, D)
7	(D, A)
8	(R, D, B)
9	(R, D, A)
10	(R, D, B, A)

Fig. 3. Compositions of the pattern groups designed for the rule-based inference engine

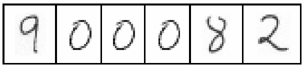


Fig. 4. An ambiguous string of numeric postal code is shown. Is it 900082 in *Roman* or 100042 in *Devanagari* or 700042 in *Bangla* ?

To make a final inference on the script of a numeric postal code, the developed rule based inference engine works in two phases. Firstly, inference about the script is made for each numeral, and secondly, cumulative inference about the script is done on a string of numerals (as required in a numeric postal code). It is apparent from the above discussion that it is impossible to predict the script of a single numeral, unless the pattern belongs to any of the aforementioned four groups i.e., 0, 1, 2 and 3.

In case the script of the digit pattern cannot be determined directly, multiple digit patterns are required to infer on the script of the string of numerals. For example, it may be observed from above that either Group_ID#4 or 8 alone is incapable to decide on the script of a single pattern. In the case of numeric pattern of Group_ID#4, there may be ambiguity among the *Roman/Arabic* scripts and for the Group_ID#8, the ambiguity may be among the *Roman/Devanagari/Bangla* scripts. But if two numeric patterns belonging to these two groups appear simultaneously in a numeric string, the script inference engine decides that the script of the numeric string is *Roman*. The justification behind this may also be observed from the set intersection of the aforesaid groups, i.e. (R, A) $\cap$ (R, D, B) $\Rightarrow$ (R).

It may however be noted that due to inherent ambiguities of handwritten numerals the script inference engine may lead to indecision on the script of the numerals in postal codes (illustrated in Fig. 4). In all other cases the script inference engine identifies the true script for any numeral string. This in turn invokes the numeral recognition engine for the corresponding script. Details of the training and test datasets, prepared for each such classifier, are discussed in the following section. Similar to unified pattern classifiers, 64 longest-run features are extracted from each of the pattern classes of any given script using a quad-tree structure. A multi-layer perceptron with back-propagation learning algorithm is again used for each script.

3 Experimental Results

To evaluate the performance of the present technique, isolated handwritten numeral datasets of RDBA scripts are prepared at the CMATER laboratory of Jadavpur University, Kolkata, India. One of these datasets is formed for the unified pattern classifier consisting of 25 unique shaped numerals of the *RDBA* scripts and one each for the numeral recognition engines of the *Roman, Devanagri, Bangla* and *Arabic* scripts. Details of the script specific digit datasets are given in www.cmaterju.org.

The dataset for these 25 unique shaped numerals is formed from 6000 randomly selected handwritten samples of *RDBA* scripts, with 1500 samples taken from the dataset of each of the four scripts. If any unique pattern appears in multiple scripts, the same pattern is considered multiple times from multiple scripts with the same label in the overall dataset. This is so because there may be minor variations of any unique shape across different scripts. Therefore, this dataset contains an unbalanced proportion of samples for each pattern.

For the present work an MLP [12] with one hidden layer is chosen. *Back Propagation* (BP) learning algorithm with learning rate (η) = 0.8 and momentum term (α) = 0.7 is used here for training of the MLP based classifier for different numbers of neurons in its hidden layer. As observed from Table 1, the best recognition rate achieved for the Unified Pattern Classifier with different numbers of neurons in the hidden layer is 88.8%. The decision on the label of the unique digit pattern, as obtained from the unified pattern classifier, is fed to the script inference engine, which subsequently re-groups the patterns into 11 categories.

To evaluate the performance of the script recognition engine on a string of numerals of any of the *RDBA* scripts, random strings of variable lengths are populated from the aforementioned numeral datasets. The average script-inference accuracy over a six digit numeric string is observed as 95.1%. Similar to the unified pattern classifier, for classification of the 10 digit patterns for each of the *RDBA* scripts, the 64 element feature set is again used. As observed from these experiments, the best recognition rates of the four classifiers for *Roman, Devanagari, Bangla* and *Arabic* scripts are obtained as 96.10%, 96.40%, 96.45 % and 95.60% respectively.

Table 1. Recognition performances of different MLP classifiers on the respective test samples with different numbers of neurons in the hidden layer of each

No of Hidden neurons	**Unified pattern classifier**	***Roman* digit classifier**	***Devanagari* digit classifier**	***Bangla* digit classifier**	***Arabic* digit classifier**
40	86.4	95.7	95.8	95.85	95.20
45	87.35	95.85	95.4	96.4	94.90
50	87.45	95.7	96.1	96.35	95.40
55	88.05	95.85	96.1	**96.45**	95.20
60	87.6	95.8	95.5	96.4	95.40
65	87.65	95.6	96	96.35	95.40
70	87.65	95.95	95.9	96.15	**95.60**
75	87.45	**96.1**	**96.4**	96.3	95.30
80	88.15	95.8	95.6	96.25	95.20
85	**88.8**	95.8	95.3	96.45	95.30
90	87.45	95.85	96.2	96.45	95.20

4 Conclusion

A novel multi-stage framework has been introduced here for automatic sorting of multi-script postal documents. The designed framework is novel in the sense that it addresses the need of a practical mail sorting system in a multi-script environment based on the analysis of numeric postal codes alone. The technique is also having potential applications in numeral based script identification schemes from multi-script document images. The designed framework may be extended to incorporate rest of the regional Indian scripts for potential applications in the Nation-wide postal automation system. One of the limitations of the designed system is its bottleneck in resolving inherent ambiguities in script identification in a string of handwritten numerals. In a random numeric string, if the rule-based inference engine fails to converge on a specific script, the numeric string remains ambiguous even through manual intervention. In such cases, scripts of the numeric string may be inferred from the script of the textual address parts, as far as practicable.

Acknowledgement

Authors are thankful to the CMATER, SRUVM, CSE Department, Jadavpur University, for providing infrastructural facilities during progress of the work. Prof. Dipak Kumar Basu is thankful to the A.I.C.T.E. (New Delhi, India) for awarding him an Emeritus Fellowship (F. No: 1-51/RID/EF (13)/2007-08).

References

1. Sinha, S., Pal, U., Chaudhuri, B.B.: Word–Wise Script Identification from Indian Documents. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 310–321. Springer, Heidelberg (2004)
2. Roy, K., et al.: A System for Wordwise Handwritten Script Identification for Indian Postal Automation. In: IEEE INDICON 2004, pp. 266–271 (2004)
3. Zhou, L., Lu, Y., Tan, C.-L.: Bangla/English Script Identification Based on Analysis of Connected Component Profiles. In: Bunke, H., Spitz, A.L. (eds.) DAS 2006. LNCS, vol. 3872, pp. 243–254. Springer, Heidelberg (2006)
4. Roy, K., et al.: A System towards Indian Postal Automation. In: Proc. of the 9th IWFHR, pp. 361–367 (2004)
5. Roy, K., et al.: A System for Indian Postal Automation. In: Proc. of the 8th ICDAR (2005)
6. Basu, S., Chaudhuri, C., Kundu, M., Nasipuri, M., Basu, D.K.: A Two-Pass Approach to Pattern Classification. In: Pal, N.R., Kasabov, N., Mudi, R.K., Pal, S., Parui, S.K. (eds.) ICONIP 2004. LNCS, vol. 3316, pp. 781–786. Springer, Heidelberg (2004)
7. Basu, S., Sarkar, R., Das, N., Kundu, M., Nasipuri, M., Basu, D.K.: Handwritten Bangla digit recognition using classifier combination through DS technique. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 236–241. Springer, Heidelberg (2005)
8. Pal, U., et al.: Handwritten Numeral Recognition of Six Popular Indian Scripts. In: ICDAR 2007, pp. 749–753 (2007)
9. Wen, Y., et al.: Handwritten Bangla numeral recognition system and its application to postal automation. Pattern Recognition 40(1), 99–107 (2007)
10. Basu, S., et al.: Recognition of Pincodes from Indian Postal Documents. Soft Computing, 239–245
11. Basu, S., et al.: A Hierarchical Approach to Recognition of Handwritten Bangla Characters. Pattern Recognition 42(7), 1467–1484 (2009)
12. Nilson, N.J.: Principles of Artifcial Intelligence, pp. 21–22. Springer, Heidelberg

Foreground Text Extraction in Color Document Images for Enhanced Readability

S. Nirmala and P. Nagabhushan

Dept of Studies in Computer Science, University of Mysore, Mysore-570 006, India
nir_shiv_2002@yahoo.co.in, pnagabhushan@compsci.uni-mysore.ac.in

Abstract. Quite often it is observed that text information in documents is printed on colorful complex background. Smooth reading of text content in such documents is difficult due to background patterns and mix up of foreground text color with background color. Further the character recognition rate when such documents are OCRed, is low. In this paper we are presenting a novel approach for extraction of text information in complex color document images. The proposed approach is a three stage process. In the first stage the edge map is obtained utilizing the Canny edge operator. The edge map is split into blocks of uniform size and image blocks are classified as text or non-text. In each text block the possible text regions are identified and enclosed in tight bounding boxes using x-y cut on edge pixels. Further the text regions that are immediate adjacent to each other in vertical direction in which the character(s) are split horizontally are merged so as to enclose the character(s) fully in one text region. In the second stage certain amount of false text regions are eliminated based on a property of printed text. In the last stage the foreground text in each text region is extracted by unsupervised thresholding using the data of refined text regions. We conducted exhaustive experiments on documents having variety of background complexities with printed foreground text in any color, font and tilt. The experimental evaluations show that on an average 98.03% of text is identified. The processed document images showed better performance when OCRed compared with the corresponding unprocessed source document images.

Keywords: Color document image, Complex background, Foreground Text extraction, Text region detection, Unsupervised thresholding, OCR.

1 Introduction

Often we find many documents that are designed deliberately with colorful and complex background for instance news paper articles, advertisements, magazine pages. Background patterns, high level variation of background color(s), combination of foreground text color and background color cause non smooth readability of the document contents. Further, automatic OCRing of such documents result in low recognition accuracy. In past many efforts were reported on separation of foreground from background of document images [1]-[5]. Thresholding is a simple and effective method of isolation of foreground from background of a

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 387–392, 2009.

document [1]. In [3] the performance of five popular local thresholding methods on four types of 'difficult' document images is evaluated and it is reported that no single algorithm works well for all types of images. Most of the thresholding methods are based on the apriori knowledge on foreground and background intensity. Practically it is not possible to know the polarities of foreground and background intensities which call for a specialized binarization technique. In [5], a specialized binarization method is proposed to extract the characters in color document images. Text-regions in a document image can be detected either by connected component analysis [2] or by texture analysis method [6].The connected component based methods detect text at faster rate but are not very robust for text localization. Also they result in false text regions for images with complex background. On the other hand texture based methods [6]are robust in detecting the text regions but they are very expensive.Most of the works discussed earlier on separation of foreground in color/gray document images show some serious shortcomings and impossible to apply on documents with complex color background with foreground text in any color and tilt. To overcome the above drawbacks, in this work we propose a three stage novel approach to extract the foreground text in complex color document images. The rest of the paper is organized as follows. Section 2 introduces the proposed approach. Experimental results are provided in section 3.Conclusions drawn from the current study are summarized in section 4.

2 Proposed Method

In this paper we propose a novel approach for extraction of printed foreground text in complex color documents which are of low resolution and scanner based. Fig.1 shows the block diagram of the proposed method. The proposed method is based on a property of printed characters that they form the edges against the background due to high intensity gradient.

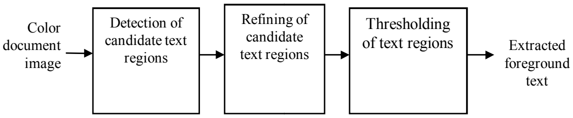
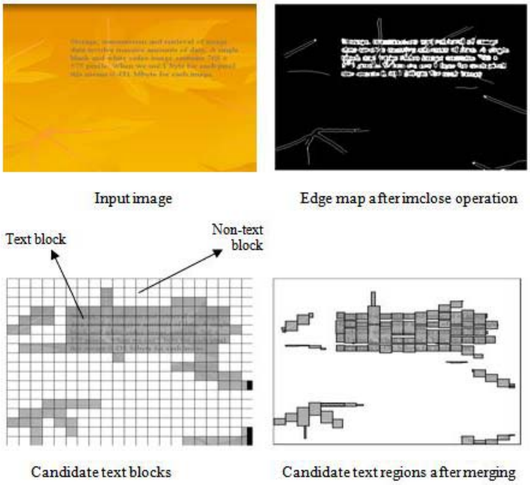

Color document image
Detection of candidate text regions
Refining of candidate text regions
Thresholding of text regions
Extracted foreground text

Fig. 1. Block diagram of proposed method

2.1 Detection of Candidate Text Regions

As Canny edge detector has low probability of missing an edge, we propose Canny edge operator for detection of text edge pixels. We have set the threshold values 0.3 and 0.4 for the hysteresis thresholding step of Canny edge detection. To avoid loss of text edges, edge detection is carried out in each color channel of RGB color model using Canny edge operator [4]. The final edge map is formed by assimilating the results of edge detection in all the three color components [4]. Suppose E_R, E_G and E_B are the edge images of red, green and blue components the final edge map 'E' is given by, E=$E_R \vee E_G \vee E_B$, where '$\vee$'represents logical 'OR' operator. As the resolution of the image is very low the broken edges are connected using 'imclose' operation with a structuring element in vertical direction. We conducted experiments to set the size of the structuring element. Structuring element size = 4 pixels result into maximum reduction of false text regions without loss of true text regions in stage-2 of the proposed method. With structuring element size < 4 pixels result in high reduction of false text regions but certain percent of true text regions are lost. Hence structuring element size is empirically set to 4 pixels. The modified edge map is split into blocks of uniform size. The blocks in the edge map that do not contain even a single edge pixel are classified as non-text blocks and these non-text blocks are ignored as they compose only the background pixels. The blocks that contain at least one edge pixel is considered as candidate text block. In each candidate text block the text regions are enclosed in tight bounding boxes by performing x-y cut on edge pixels. We identify the text regions which are originally placed in adjacent text blocks in vertical direction and merge those adjacent text regions in which the horizontal split of character strings appear. Fig 2. shows the output of sub processes of stage-1 in sequence for an initial block of size 25 × 25.


Input image

Edge map after imclose operation

Candidate text blocks

Candidate text regions after merging

Fig. 2. Output of sub processes in stage-1

2.2 Refining of Candidate Text Regions

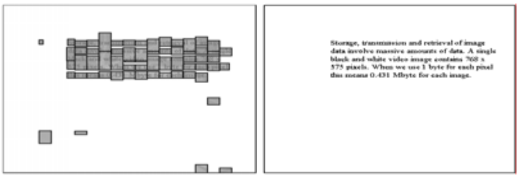
Due to high intensity of some background objects the edges of background objects might be detected by Canny edge operator. The candidate text regions that contain only the edges of background objects (false text region) are identified and eliminated in stage-2. To develop a criterion to identify the false text region the following feature of text is observed [6]. The fact we observed is that the number of text edge pixels is more in true text region compared to false text region. Suppose 'W' and 'H' are the width and height of the text region. If the text edge pixel count in a text region is greater than maximum(2 * W, 2 * H) we classified it as true text region else as false text region. The false text regions are removed in this stage and true text regions are considered for further processing which is described in the subsection that follows.

2.3 Thresholding of Text Regions

The true text regions which are obtained from stage-2 are thresholded locally to extract the foreground pixels and deposited in proper position on a uniform white background. We considered the gray scale equivalent of the corresponding text region for thresholding the text region. A specialized unsupervised thresholding is designed based on foreground pixel intensity and background pixel intensity and deposited foreground characters in black on uniform white background. The approximate background intensity is computed by averaging the intensity values of non edge pixels in the updated edge map. For each text region the approximate foreground intensity is computed by averaging the intensity values of edge pixels of updated edge map. Threshold value 'Th' for each text region is computed as follows:

If (abs(average foreground pixel intensity - average background pixel intensity) > 40)
Th=average foreground intensity
Else
Th= 0.5 * (average foreground pixel intensity + average background pixel intensity).

Fig. 3. shows the output images obtained from stage-2 and stage-3.


Storage, transmission and retrieval of image data involve massive amounts of data. A single black and white video image contains 768 x 575 pixels. When we use 1 byte for each pixel this means 0.431 Mbyte for each image.

(a) (b)

Fig. 3. (a) Output image from stage-2 (b) output image from stage-3

3 Experimental Results

Since no standard corpus of images is available for this work we created our own corpus of printed color document images by scanning the documents from various sources viz. magazines, story books, postal envelopes and newspapers. In addition we created another corpus of synthesized images. All the images in both the corpus are of low resolution. Irrespective of the foreground font color the output image is created by depositing black characters on white background. The performance of amount of text detection is evaluated in terms of Recall ((correct detects / (correct detects + missed detects)) and Precision (correct detects / (correct detects + false alarms)). Table 1. shows the average value of Precision and Recall in percentage for document images in the corpus.

Table 1. Performance evaluation of text detection

No. of samples	Total number of characters	Recall(%)	Precision(%)
160	31633	98.03	97.80

Table 2. OCR results

Initial Block size	OCR Recognition rate (before processing)	OCR Recognition rate (after processing)
25 × 25	56.25	66.84
50 × 50	56.25	62.97
75 × 75	56.25	58.29

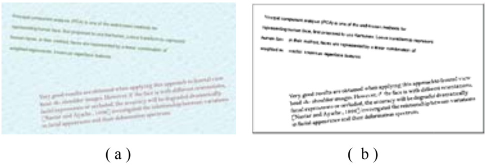


Fig. 4. Result of sample a document image with foreground text in different font, color and tilt : (a) Input , (b) Output

Reading of the extracted text is evaluated on Readiris 10.04 pro OCR. Readiris 10.04 pro OCR handles color /gray document images. In this work readability of the extracted foreground text is evaluated in terms of character recognition rate. Table 2. shows the average character recognition rate by OCR before and after applying the proposed approach. Although the average performance after processing appears to be around 66% it should be noted that in some specific difficult cases the recognition rate drastically improved to nearly 100% (after processing) from recognition rate of 0% (before processing).It is also observed that the better performance is with initial block of size 25 × 25 which is evident from table 2. Fig. 4. shows result of a sample document image with foreground text in different color, font and tilt.

4 Conclusion

In this paper a novel approach is presented for extraction of foreground text from complex background in color document images which are of low resolution and scanner based. The candidate text blocks are identified based on a characteristic property of the printed characters. Identified text regions in each block are enclosed in tight bounding boxes. By designing a criterion based on text pixel count in a text region the false text regions are filtered out. An unsupervised thresholding is devised to extract the foreground text in the refined text regions. The proposed approach detects 98.03% of text content in the source document. We achieved 66.84% of OCR character recognition accuracy. The proposed approach fails to extract the characters that are too thick. Devising an effective criterion for elimination of false text regions so as to improve the recall rate and designing a thresholding technique to extract the foreground characters to improve the character recognition accuracy by OCR are considered as future works of the current study.

References

1. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic imaging 13, 146–165 (2004)
2. Pietikäinen, M., Okun, O.: Text extraction from grey scale page images by simple edge detectors. In: Proceedings of the 12th Scandinavian Conference on Image Analysis, SCIA, Norway, pp. 628–635 (2001)
3. Leedham, G., Chen, Y., Takru, K., Tan, J.H.N., Mian, L.: Comparison of some thresholding algorithms for text/background segmentation in difficult document images. In: Proceedings of seventh International Conf. on Document Analysis and Recognition (ICDAR), pp. 859–864 (2003)
4. Shivananda, N., Nagabhushan, P.: Separation of Foreground Text from Complex Background in Color Document Images. In: Proceedings of Seventh international conference on advances in pattern recognition, ISI Kolkata, pp. 306–309 (2009)
5. Kasar, T., Kumar, J., Ramakrishnan, A.G.: Font and Background Color Independent Text Binarization. In: Proceedings of 2nd Intl. workshop on Camera Based Document Analysis and Recognition (workshop of CBDAR), pp. 3–9 (2007)
6. Liu, Y., Goto, S., Ikenaga, T.: A contour based robust algorithm for text detection in color images. IEICE Transactions on Information and Systems 89, 1221–1230 (2006)

A Novel Approach to Skeletonization for Multi-font OCR Applications

C. Vasantha Lakshmi, Sarika Singh, Ritu Jain, and C. Patvardhan

Dayalbagh Educational Institute
cvasantha@rediffmail.com, sarika_singh01@yahoo.com

Abstract. A novel approach to generate skeletons of binary patterns that has a wide variety of applications including multi-font OCR is proposed in this paper. The proposed algorithm ensures connectedness of the pattern and minimizes loss of information while capturing the essential shape characteristics. Computational tests on printed Telugu characters show that the algorithm is useful in getting a generalized form of the character symbols on the common multiple dissimilar fonts.

Keywords: Skeletonization, OCR, Multifonts, telugu.

1 Introduction

Skeletonization plays an important role for the analysis and recognition of binary images e.g. in single and multi-font Optical Character Recognition (OCR). Although, machine printed characters are uniform in size, position and pitch for any given font and are “OCR friendly”, multi-font OCR has always been proven to be a difficult problem because of the wide variety in which the characters are written in different fonts [1-4].

But then how do human beings manage to recognize the different fonts and even different people’s handwritings? This is because they somehow eliminate the redundant information that the different ornate fonts and handwritings have and recognize the basic structure of the character. Humans do not rely on pixel level features. They somehow capture the “overall” shape of the character. The same ability is necessary for a multi-font OCR engine to succeed in recognizing a variety of shapes. Figure 1 shows the skeleton of a character for two different fonts. It is evident from the figure that the skeleton looks the same in spite of the wide difference in fonts. Once this skeleton is captured it would be easier to build a recognizer that recognizes that the characters are same. An algorithm is proposed in this paper to compute the skeleton of a character.

A good skeletonization algorithm should possess the following properties: (1) preserving connectivity of skeleton, (2) converging to skeleton of unit width, (3) closely approximating the medial axis; (4) possessing insensitivity to boundary noise; (5)

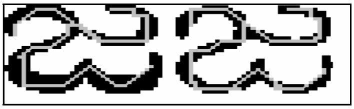


Fig. 1. Skeletons of a telugu character

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 393–399, 2009.

achieving high data reduction efficiency. Thinning is the most frequently used method to achieve the skeletonization goal. In the past several decades, many thinning algorithms have been developed and a comprehensive survey is presented in [5].

Some classical algorithms provide good results but still have some deficiencies. For example, some of them cannot preserve the connectedness of an image. In addition, some of these algorithms result in loss of information of the pattern. Huang et al. [6] overcome this information loss by integrating contour and skeleton of pattern. Another method based on connected component approach is proposed by Vijaya Kumar et al. [7]. A simple sequential thinning algorithm for peeling off pixels along contours is described by Govindan and Shivaprasad [8]. One more problem with thinning algorithms is deformation at crossing points. To solve this problem, a knowledge-based thinning algorithm (KBTA), was proposed by Li and Suen [9]. Block Adjacency Graph(BAG) structure is proposed by Suryaprakash[12] .

In this paper, a new skeletonization algorithm using modified Block Adjacency Graph (BAG) structure is proposed which can be used for getting a generalized form of character symbol for multiple fonts. The algorithm presented is not an iterative deletion of pixels. It ensures the connectedness of the image and also minimizes information loss. Although the approach can be used for any character image in general, computational performance on several popular fonts of an Indian script, Telugu, is presented.

The rest of the paper is organized as follows. The proposed approach for skeletonization is described in section 2. Quantitative evaluation is done is section 3. Results on various fonts and sizes are in section 4. Conclusions are derived in section 5.

2 The Proposed Approach

It is very difficult to balance the twin requirements of removal of as many pixels as possible and maintaining connectivity. Thus, the proposed approach relies on finding shape by approximating it as a sequence of carefully constructed segments. BAG can be created by classifying runs as merging, splitting, or continuing runs. The relevant concepts and definitions are as follows.

Horizontal and Vertical Runs-Run Length encoding keeps track of horizontal runs (Hruns) and vertical runs (Vruns) of black pixels, storing the coordinate of the first black pixel and the run length in the corresponding row or column.

Splitting, Merging and Continuous Runs-Hruns are classified by the number of Hruns above and number of Hruns below. If the number of runs above=0 and the number of runs below >1 then the runs are split. If the number of runs below=0 and above >1 then the runs are merged. The remaining Hruns are considered continuous.

Area of block: Number of black pixels in the block.

The proposed method with the help of pseudo-code is given below.

Step1: Isolate the character image say I from the document.

Step 2: Convert I to binary image I' using Otsu's method [10].

Step 3: Create horizontal and vertical runs of the image I' as shown in Figure 2(c).

Step 4: Mark runs as splitting, merging and continuous as defined above (Figure 2 (d)).

Step 5: Construct blocks as shown in Figure 2 (e):

(a) Merge all adjacent continuous runs to form a block till a split or merge run is encountered.

(b) Merge this split/merge run into the block if it is directly above or below the block so formed else start a new block.

Step 6: Compute centroids of each of the blocks as shown in Figure 2(f).

Step 7: To remove the spurious blocks/pixels which may result in skeletal legs, a block is removed if both the following conditions are true.

(a)Area of block is less than average size where average size is computed as

$$Avg.Size = \frac{\sum_{i=1}^{n} Area\ (b_i)}{n}$$

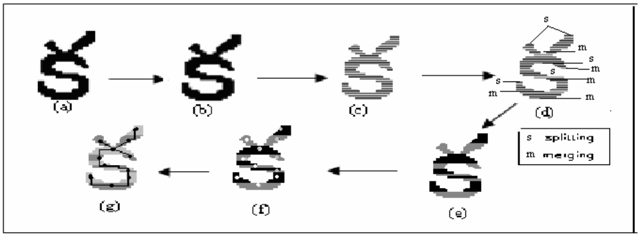
,where b_i denotes block i and n is the total number of blocks.

(b) The block is adjacent to only one block and number of black pixels on the path connecting the two centroids is less than k. Value of k depends on average size of character image.

Step 8: Connect the centroid points preserving the path as observed from the original character image as follows.

(c) If two adjacent centroid points, if they lie in same column or row, and path between them corresponds with original image, all pixels connecting them in that row or column are made black.

(d) If two adjacent centroid points do not lie in same row or column the connections as shown in Figure 3 can occur. So the path is computed using these two centroid points and the actual path present in the original image between them.



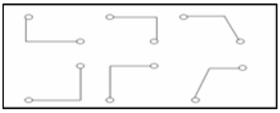


Fig. 3. Connecting paths

Fig. 2. Steps in proposed algorithm (a) Isolated character (b) After Otsu's thresholding (c) Horizontal runs and vertical runs images (d) runs marked as splitting, merging and continuing (e) Creation of blocks (f) computation of centroids (g) connecting centroid points

The identification of splitting, merging and continuous runs (step 4), conditions for noise removal (step 7) and connecting the centroids (step 8) are different.

3 Quantitative Evaluation and Experimental Results

In this section, a set of measures is defined to evaluate the performance of the proposed algorithm. Connectivity was first examined. Then, measures are computed to evaluate i) how closely the resulting skeleton approximates the true medial axis of the object, ii) how well the pattern converges to the desired one pixel-wide skeleton, and iii) how sensitive the resulting skeleton is to boundary noise. Finally, the data reduction efficiency is measured to evaluate the efficiency of the algorithm.

3.1 Connectivity

A connectivity algorithm is used to check that skeleton contains only one connected component [11]. In the proposed algorithm, since the connection of centroids of adjacent blocks is established following the possible paths; the skeleton is always connected.

3.2 Measure of Convergence to Unit Width

If the converged skeleton S_M does not contain any one of the patterns $\boldsymbol{Q}^k$ as shown in Figure 4, then S_M is one pixel wide. To measure the width of the resulting skeleton, m_t is defined as $m_t = 1 - \frac{Area\left[\bigcup_{1\le k\le 4} S_M Q^k\right]}{Area[S_M]}$ where Area is the operation that counts the number of black pixels. This measure has a nonnegative value less than or equal to 1, with $m_t = 1$ if S_M *is* a perfect unit-width skeleton.

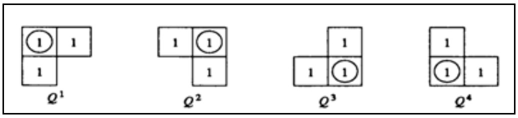


Fig. 4. Templates Q^k used to examine the width of the convergent skeleton

3.3 Measure of Medial Axis Representation

To evaluate the amount of variations between the skeleton and its ideal medial axis, a measure is defined as a function of the input image S and its resulting skeleton S_M. At each point p $\in$ S_M, the maximum digital disk included in S and centered at p is denoted as DD(p,r_p), where r_p represents the radius of a continuous disk whose digital image of the disk is DD centered at p. Note that DD(p, 0) = {p}by convention. Then, we have $S' = \bigcup_{p\in S_M} DD(p, r_p)$, where $S' \subseteq S$. Hence, the measure m_m, can be defined as $m_m = \frac{Area[S']}{Area[S]}$.

This measure has a nonnegative value less than or equal to 1 with equality if and only if the resulting skeleton contains the discrete medial axis.

3.4 Measure of Boundary Noise Sensitivity

Let *S* be the given noise-free binary image and *S''* be its noisy version generated by randomly adding and subtracting k unit-size points along the boundary of *S*. We define the signal-to-boundary noise ratio as $SBNR_k = \frac{Area[\partial S]}{Area[S''/S] + Area[S/S'']}$,where ∂S denotes the boundary of *S*. The terms *S''/S* and *S/ S''* are the set differences between the ideal image and noisy image, respectively. Thus, the error introduced by boundary noise at a particular $SBNR_k$ can be measured by the normalized quantity

$m_e(SBNR_k) = \min(1, \frac{Area[S_M / S''_M] + Area[S''_M / S_M]}{Area[S_M]})$,where S_M and S''_M are the resulting skeletons of S and S'', respectively. The measure is normalized between 0 and 1; a highly noise-sensitive algorithm will yield an m_e close to 1.

3.5 Measure of Computational Cost

A measure to evaluate both the data reduction efficiency and the computational cost is defined as $m_d = \min\left[1, \frac{Area[S] - Area[S_M]}{Area[S]}\right]$. This measure has a value between 0 and 1; a large value indicates high efficiency. The quantitative measures described above were used to evaluate the proposed algorithm and compared with parallel thining algorithm[13] on sample skeleton shown in figure 5.

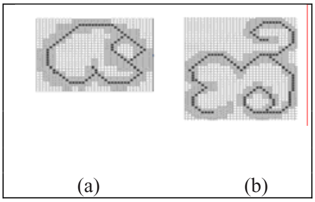


Fig. 5. Skeletons obtained by proposed algorithm

Table 1. Comparison of parallel thinging algorithm and proposed algorithm

Quantitative evaluation	Parallel thining Algorithm of fig. 5(a)	Proposed algorithm of fig. 5(a)	Parallel thining Algorithm of fig.5 (b)	Proposed algorithm of fig. 5(b)
m_t	0.99	0.99	0.99	0.99
m_m	1.00	0.99	0.99	1.00
m_d	0.77	0.80	0.82	0.84

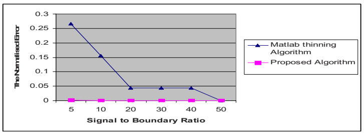


Fig. 6. Effect of boundary noise

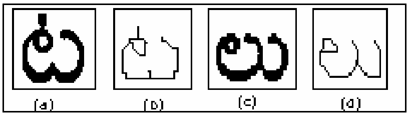


Fig. 7. (a) Telugu character Ta (b) Skeleton of Ta (c) Telugu character lu (d) Skeleton of lu

In another experiment, a random boundary noise was generated at various (signal boundary ratio) $SBNR_k$ levels. The experiment is summarized by plotting $SBNR_k$ against m_e in Figure 6.

4 Results on Different Telugu Fonts and Sizes

Telugu character samples are prepared using 3 fonts- Godavari, HarshaPriya and Hemlata and three font sizes: 25, 30, and 35. Figure 7 shows two confusingly similar characters 'Ta' and 'lu' in Godavari 25 and their skeletons. An experiment is performed using density features on 3x3 grids with Euclidean distance measure for computing the distance. Figure 7(a) is the base character for comparison in columns 3 and 4 while Figure 7(b) is for columns 6 and 7 of Table2. Comparing columns 5 and 8 depicting ratio's, we see that the distinction between confusingly similar characters is more prominent with skeletonized characters and as a result is less prone to errors in recognition.

Table 2. Distance of skeletonized and unskeletonized character ‘Ta’ (font Godavari and size 25) with character ‘lu’ of different fonts and sizes and with same character of different fonts and sizes

Fonts	Sizes	Distance of 7(a) with unskeletonized		Ratio (2)/(1)	Distance of fig.7 (b) with skeletons of		Ratio (4)/(3)
		‘Ta’ (1)	‘lu’(2)		‘Ta’(3)	‘lu’(4)	
Godavari	25	0	0.504	-	0	0.116	-
	30	0.208	0.39	1.87	0.033	0.1396	4.23
	35	0.1519	0.429	2.82	0.0484	0.1269	2.62
HarshaPriya	25	0.2394	0.456	1.90	0.056	0.1023	1.82
	30	0.2912	0.4857	1.66	0.047	0.095	2.02
	35	0.3433	0.542	1.57	0.032	0.1104	3.45
Hemlata	25	0.2617	0.4616	1.76	0.0273	0.1212	4.43
	30	0.3042	0.4151	1.36	0.0509	0.1318	2.58
	35	0.27	0.454	1.68	0.043	0.1296	3.01

5 Conclusion

Multi-font OCR has proven to be a tough problem, especially for Indian scripts because of the complicated scripts and the large variety of ways in which the characters are written in different fonts. One of the approaches to tackle this is to reduce the characters to their basic structural forms. In this paper, a method is proposed to do this utilizing a modified BAG data structure. Examples are provided to show that the method indeed amplifies the dissimilarity between different characters and the similarity between same characters in different fonts. This is conducive to better downstream operations for OCR like feature detection and recognition. Efforts are on to develop a complete OCR engine for multi-font Telugu OCR based on this simplification approach.

References

[1] Chan, C., Wong, P.: A branch and bound decision tree Bayes classifier for robust multi-font printed Chinese character recognition. In: IEEE Region 10th International Conference, November 11-13, vol. 1, pp. 267–271 (1992)

[2] Ben, A.N.: Multifont Arabic Characters Recognition Using Hough Transform and HMM/ANN Classification. Journal of Multimedia 1(2) (May 2006)

[3] Ho, T.K., et al.: A Computational Model for Recognition of Mutifont Word Images. Machine Vision and Applications 5, 157–168 (1992)

[4] Wang, J., Jean, J.: Resolving multifont character confusion with neural networks. Pattern Recognition 26(1), 175–187 (1993)

[5] Lam, L.S.W., Suen, C.Y.: Thinning Methodologies-A Comprehensive Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(9), 869–885 (1992)

[6] Huang, L., Genxun, W., Liu, C.: An Improved Parallel Thinning Algorithm., Proceedings of the Seventh International Conference on Document Analysis and Recognition, ICDAR (2003)

[7] Kumar, V.V., Srikrishna, A., Shaik, S.A., Trinath, S.: A New Skeletonization Method Based on Connected Component Approach. IJCSNS International Journal of Computer Science and Network Security 8(2) (February 2008)

[8] Govindan, V.K., Shivaprasad, A.P.: A pattern adaptive thinning algorithm. Pattern Recognition 20(6), 623–637 (1987)

[9] Li, B., Suen, C.Y.: A knowledge-based thinning algorithm. Pattern Recognition 24(12), 1211–1221 (1991)

[10] Otsu, N.: A Threshold selection method from Grey-level histograms. IEEE Transactions on Systems, Man and Cybernetics SMC-9(1) (January 1979)

[11] Efford, N.: Digital Image Processing. A practical introduction using java

[12] Kompalli, S.: A stochastic framework for font-independent Devnagari OCR. P.hd. thesis, SUNY Buffalo (January 2007)

[13] Guo, Hall, R.W.: Parallel thinning with two-subiteration algorithms. Comm. ACM 32(3), 359–373 (1989)

A Novel Approach for Detection of Alteration in Ball Pen Writings

Rajesh Kumar[1], Nikhil R. Pal[2], J.D. Sharma[3], and Bhabatosh Chanda[2]

[1] Directorate of Forensic Science, MHA, GOI, New Delhi, India
[2] ECSU, Indian Statistical Institute, Kolkata, India
[3] Dr. HSG University, Sagar (M. P.), India

Abstract. Addition or alteration to documents that have profound implication is very common. The technique that Forensic Document Examiners (FDEs) use for the examination of such documents is basically a physical examination. In this paper we consider the alteration detection as a two-class pattern recognition problem. Image processing techniques are used for feature extraction and a neural network based feature analysis technique is used for finding a set of discriminatory features. The results using a nearest neighbor classifier are very encouraging. The results also demonstrate the effectiveness of feature analysis.

Keywords: Alteration, ball pen, feature analysis, image processing.

1 Introduction

Although world is moving from the paper age to paperless age, billions of ball point pens are sold every year and people are frequently using these pens for writing and signing on various documents. Hence, number of forgeries involving these materials is also high. A little alteration by similar color ink in the amount written on a cheque can do havoc. Everyday at some part of the world people are victimized by such a white collar crime.

The techniques that FDEs are using in routine cases is basically a physical examination through different light sources ranging from ultraviolet to infrared and some kind of conventional optical filters. Osborne [1], the father of forensic document examination, himself suggested the physical examination of altered documents in various lighting conditions and also the use of some chemical examination. The work on alteration detection using ink analysis is being done in two major pathways. First is non-destructive kind of examination using techniques like Fourier-Transform Infrared (FTIR) microscopy, IR luminescence and laser examination. The second approach is based on destructive techniques ranging from basic and modern chromatographic techniques to Neutron activation analysis [2]. Application of image processing and pattern recognition for this problem is still limited to the extent of using some image enhancement techniques as an aid to the FDEs.

In this paper we have taken the alteration detection as a two-class problem. A machine learning technique is used for feature analysis and the nearest neighbor classifier is used for detection.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 400–405, 2009.

2 Proposed Methodology

The main type of cases that the FDEs get is insertion of some strokes on preexisting strokes. To simulate this problem 10 different ball point pens of different brands are used. As quality of paper also has some influence on writing stroke and we are interested in detecting whether pens are same or different, the same brand of paper is used.Using combinations of these 10 pens we have prepared $^{10}C_2 = 45$ combinations of two intersecting strokes on the paper [see Fig.1(a)]. To eliminate bias on the order of strokes, the order of strokes is interchanged to create another 45 combinations. Two copies of each of these 90 combinations are then created. The pictures of intersecting strokes are enlarged 80 times and corresponding images are captured using VSC 5000.

The background area due to paper is removed by simple graylevel thresholding method. Each intersecting stroke in the image is then divided into two non-overlapping segments which are used for feature extraction.

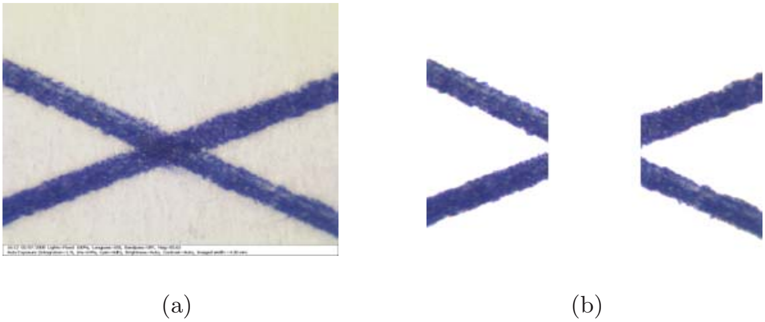

(a) (b)

Fig. 1. A typical sample image data (a): Image of intersecting strokes (b): segmented strokes used for feature extraction

2.1 Feature Extraction

Our aim is to extract some features that are close to what can be perceived by human and can be used to distinguish between different pens. Keeping this in mind we have chosen $\mathrm{Y}C_bC_r$ and opponent chromaticity space (rg, yb) to compute various features. The selection of two color spaces simultaneously can be justified through opponent process theory of color vision. According to that theory, the trichomatic theory of color vision does not explain all aspects of color vision. The reason for not seeing combinations like reddish-green or yellowish-blue, though that can be formed in $\mathrm{Y}C_bC_r$ space, is that the opponent responses are controlled by the opponent neurons. So, considering both the spaces one can get complete color description of an object. A survey of trichomatic color spaces can be found in [3,4], while opponent chromaticity space is described in [4,5].

We compute various moments of joint distribution of various color channels as color features and various moments of graylevel co-occurence matrix (glcm) as texture features. In particular, we calculate Legendre and geometric moments [6,7] up to third order for each of rg-yb and C_b-C_r spaces. Thus 18 moment features from each of the two spaces in addition to mean and standard deviation for each channel are the color features of each channel. We also compute three sets of texture features namely contrast, homogeneity and energy from glcms [8] defined for four distances (10, 20, 30 and 40) and involving all eight directions. Thus a total of 106 features are extracted for each stroke segment.

2.2 Detection of Alteration

We are considering alteration detection as a two-class problem: class-I, when the same pen is used to create two intersecting strokes, and class-II, when two different pens are used for the creation of the two strokes. Each image in the data set is having two intersecting strokes and each stoke is segmented into two segments. Thus, ${}^4C_2$= 6 feature vectors are created from each image combining the features of 2 stroke segments at a time. By toggling the feature vectors of each stroke segment, another 6 feature vectors are obtained from the same image. Thus a total of 12 feature vectors each with 212 features are generated from each image. These 12 feature vectors are included into class-I (same pen stroke) or class-II (different pen stroke) according to their combination.

The nearest neighbor classifier is used for the detection of alteration due to its simplicity and robustness. One particular pen P_k is kept out and the data for remaining 9 pens are used as the training data. The process is repeated for each pen $P_k; k = 1, \cdots, 10$. By doing so, one can avoid bias towards any pen.

2.3 Feature Analysis

We shall see later that the detection results with the nearest neighbor classifier using all features exhibit some undesirable behavior when a particular pen is left out for testing. This could be because of presence of some poor feature(s) that may be affecting the performance noticeably when a particular pen is left out. Moreover, use of all 212 features may not be needed for this problem. For an easy identification of the decision making system as well as for better generalization it is desirable to use just an adequate set of features, possibly with limited redundancy. Hence, with a view to improve the overall classification performance and with a hope to realize a more uniform performance when different pens are left out, we do some feature analysis.

There are many feature selection techniques available in the literature [9]. Here we use a Multilayered Perceptron (MLP) based feature selection method, that we call FSMLP, in short [10]. In this method we associate an adaptive gate to each input node (hence each feature) of the network. The gate is modeled by a monotonic differentiable function $g()$ whose range is [0,1]. Each gate has a tunable parameter. The degree to which a gate is opened determines the goodness of the feature. Unlike conventional MLP, here each input node modulates the input feature value by computing the product of the input feature and the gate function value. This modulated input is then passed into the next layer of the network. The FSMLP begins its training assuming that all features are bad features. In other words, at the beginning, each modulated feature takes a value of nearly zero and that is passed into the higher layers of the network. We want to emphasize on the word *almost zero*. If the value is equal to zero then the learning cannot proceed. The learning process uses gradient descent to minimize the classification error. Therefore, the features which can reduce the error faster are likely to get their associated gates opened faster; while the gates associated with bad features (that cannot reduce the error) are likely to get closed more tightly. In all our experiments, we train the FSMLP only for 1000 iterations. There is no need to train FSMLP till error becomes very low. We can stop when the training error is just satisfactory as our objective is only to pick up the good features.

An important advantage of this method over many others is that it can account for the non-linear interaction between the features as well as that between features and the tool (here neural network). For details, please see [10]. Use of MLP type networks raises many issues like choice of network size etc. Here we have made a few experiments and found that a choice of 8 or 9 hidden nodes is good enough. Hence, in all our experiments, we have used 8 hidden nodes. Since such a network uses gradient search, depending on the initial conditions, different runs can result in different sets of features, each of which may be equally good. This can particularly happen when there are many correlated features. Hence, we proceed as follows. A pen, P_k, is kept out for testing. The remaining nine pens are then used for feature selection and classifier design. Using these nine pens, the FSMLP is run 10 times. Each run R generates a gate opening value for each feature, f as $g_f^R; R = 1, \cdots, 10; f = 1, \cdots, 212$. Now we compute $g_f = \sum_{R=1}^{10} g_f^R$ as the composite importance for feature f and use these importance value to select a set of features for any experiment with pen P_k left out. Using these selected features we compute the performance of a nearest neighbor classifier. This process is repeated for each of the 10 pens, $P_k; k = 1, \cdots, 10$.

3 Results and Discussion

The second column of Table 1 shows the performance of a nearest neighbor classifier using all 212 features. Training is performed by keeping a particular pen P_k out and testing on pen P_k. One can see that when the pen P_1 is kept out the performance decreased to 69.41% while overall performance is around 78%. The declined performance of the pen P_1 with all 212 features motivated us to do feature analysis.

Figure. 2 shows the overall performance of the nearest neighbor classifier using different sets of features selected by FSMLP, the detailed performance is included in Table 1. Figure. 2 reveals that the performance of the classifier decreases in either side of 60 features. Thus only 60 ranked features are sufficient and good to discriminate strokes using the same pen or different pens. In the set of top 60 features, contrast in each of the five channels except in C_b, Homogeneity in C_b, mean of intensity in rg and Y, energy in rg and yb, third order geometric

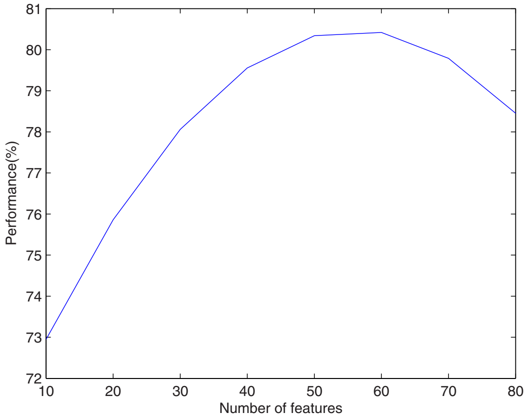


Fig. 2. A sketch of performance of a NN classifier with different number of features

Table 1. Performance(%) of a nearest neighbor classifier

No.of features	212	80	70	65	60	55	50	40
P_1	69.41	69.71	70.59	72.06	72.65	72.06	72.65	71.76
P_2	85.71	81.71	81.71	80.57	82.57	81.71	80.86	78.29
P_3	73.89	75.00	75.83	77.22	79.17	78.33	78.89	76.11
P_4	74.55	76.06	76.97	78.48	77.27	77.58	76.67	74.85
P_5	93.13	95.94	96.56	96.56	96.25	96.25	97.19	96.88
P_6	82.35	81.18	79.71	79.41	78.24	78.53	77.94	77.06
P_7	79.44	88.89	89.72	92.50	92.50	91.94	92.78	94.72
P_8	74.44	74.72	75.00	75.00	75.83	77.50	74.42	75.00
P_9	74.84	74.52	74.52	75.48	75.16	76.45	77.10	77.42
P_{10}	72.00	71.14	73.43	73.71	73.14	70.57	72.29	70.86
Total accuracy	77.98	78.89	79.40	80.10	80.28	80.09	80.11	79.29

Table 2. False Negative and False Positive rates(%) of a NN classifier

No.of features	212	80	70	65	60	55	50	40
False Negative	26.22	24.87	24.23	23.48	23.28	23.49	23.17	25.29
False Positive	05.22	06.08	06.06	05.59	05.48	05.59	06.78	06.29

and Legendre moments in rg-yb and one particular geometric moment in C_b-C_r appeared as the most consistent candidates for all pens.

One can easily compare the performance before and after the feature analysis as shown in Table 1. As expected, not only the overall performance increased to 80.28% but also the performance of the first pen has improved to 72.65%. Beside this we can see the uniformity in performance over all the pens.

Moreover, the best part of the classification is its low false positive rate (Table 2), that is very important for making legal decisions, where the motto is, *no innocent person should be convicted*. From Table 2, the impact of feature analysis can also be seen on false positive and false negative rates. We find that with 60 selected features the false negative rate goes down and the false positive rate is comparable to that of the case with all features.

4 Conclusion and Future Work

In this study we have attempted to solve an important forensic problem using image processing and pattern recognition techniques and achieved an accuracy of more than 80%. Moreover, we have obtained a low false positive rate of 5.48. which is good for making legal decisions. Our approach is quite novel and reliable as far as forensic document examination is concerned. The proposed algorithm can be an aid to FDEs for the alteration detection in combination with other conventional methods.

We shall continue our investigation to improve the performance as well as to extract and select robust features. Further, we are planning to assign a confidence value to each decision so that in case of low confidence additional information can be sought. Such a system would be of great help to forensic community.

References

1. Osborne, A.S.: Questioned Documents. Boyd Printing Co., New York (1929)
2. Ellen, D.: The Scientific Examination of Documents Methods and Techniques, 2nd edn. Taylor and Francis, London (2003)
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Printice Hall, New Jersey (1996)
4. Koschan, A., Abidi, M.: Digital Color Image Processing. Wiley Interscience, New Jersey (2008)
5. Berens, J., Finlayson, G.D., Qiu, G.: A Statistical Image of Color Space. In: IEE Proc. International conference on Image Processing and its Application (1), pp. 348–352 (1999)
6. Yap, P.T., Parmesaran, R.: Content Based Image Retrieval using Legendre Chromaticity Distribution moments. In: IEE Proc. Visual Image Signal Processing, pp. 17–24 (2006)
7. Teh, C.H., Chin, R.T.: On Image Analysis by the Methods of Moments. IEEE Transaction on Pattern Analysis and Machine Intelligence 10(4), 496–513 (1988)
8. Haralick, R.M., Bosley, R.: Texture Features for Image Classification. In: Proc. Third ERTS symposium, NASA, pp. 1219–1228 (1973)
9. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman and Hall/CRC, Taylor and Francis (2007)
10. Pal, N.R., Chintalapudi, K.K.: A Connectionist System for Feature Selection. Neural, Parallel and Scientific Computations (5), 359–382 (1997)

Resolving Ambiguities in Confused Online Tamil Characters with Post Processing Algorithms

A.G. Ramakrishnan and Suresh Sundaram

Medical Intelligence and Language Engineering Laboratory
Indian Institute of Science, Bangalore, India
{ramkiag,suresh}@ee.iisc.ernet.in

Abstract. This paper addresses the problem of resolving ambiguities in frequently confused online Tamil character pairs by employing script specific algorithms as a post classification step. Robust structural cues and temporal information of the preprocessed character are extensively utilized in the design of these algorithms. The methods are quite robust in automatically extracting the discriminative sub-strokes of confused characters for further analysis. Experimental validation on the IWFHR Database indicates error rates of less than 3 % for the confused characters. Thus, these post processing steps have a good potential to improve the performance of online Tamil handwritten character recognition.

Keywords: Confusion Pairs, Sub stroke Extraction and analysis, Fourier Descriptors, Online handwritten character recognition.

1 Introduction

Tamil is a popular classical language spoken by a significant population in South East Asian countries. There are 156 distinct symbols in Tamil [1]. As far as earlier work on recognition of online Tamil characters is concerned, Deepu *et al.* [2] generate class specific subspaces using principal component analysis, while Niranjan *et al.* [1] have employed dynamic time warping for matching unequal length feature sequences. Hidden Markov models for recognition have also been reported in [3] [4]. In a recent work, we have studied the performance of the 2DPCA Algorithm [5], which was originally proposed for face recognition.

Each of the above schemes is found to give nearly similar generalization performances on a given test data. Most of the misclassifications of the given data, in general, are attributed to the fact that Tamil has many symbols that look visually similar. Any classifier that works on features at a global level fails to capture finer nuances that make these symbols distinct. One way to circumvent this drawback would be to incorporate a post processing step that employs local features to reduce the degree of confusion between frequently confused characters, and thereby improves the overall performance of the recognition. Specifically, this paper proposes algorithms for disambiguating frequently confused symbols. The approaches are developed, taking into account, the popular writing / lexemic styles of modern Tamil script. They can be applied irrespective of the nature of the classifier used for the recognition.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 406–413, 2009.

In a system that deals with recognition at the word level one could use language models. However, when recognizing isolated characters, one is devoid of any such additional information that can be used to correct errors. Thus, we resort to the use of robust structural features for post recognition disambiguation.

2 Confusion Pair Analysis

A careful analysis of the confusion matrices, obtained with different classifier frameworks, suggests that the frequently confused Tamil characters can be manually grouped into two categories A and B [5]. Accordingly, we propose appropriate post-processing techniques to each group of the confusion pairs. The confusions in Group A appear between pairs of Tamil consonant-vowel combinations sharing the same base consonant but different vowel modifiers. The confused strokes contributing to errors are ி and ீ. These correspond to the modifiers of the vowels இ and ஈ respectively.

Apart from Group A pairs, there exist other pairs that differ predominantly in the end such as (ஐ ஜ) and (க ச). The structure to be analyzed for these pairs is strikingly different from those belonging to Group A. Consequently, these pairs are placed in Group B. Also, there exist certain character pairs that differ only at the start and middle of the trace such as (ஏ ர) (ண ள) and (மு ழு). These are also incorporated in Group B. As stated before, the spatial and temporal information provided by the online data is extensively utilized for the design of the algorithms. Prior to designing the appropriate post processing technique, the raw Tamil character is subject to pre processing modules [2] such as smoothing, size normalization and resampling.

3 Disambiguation of Group A Confusion Pairs

In this section, we outline the post processing algorithm for disambiguating the vowel modifiers ி and ீ for a given base consonant. Popular writing styles of modern Tamil script suggest that the vowel modifiers ி and ீ always form the last stroke in any multistroke consonant-vowel combination character. However, for CV combinations written as a single stroke (where the vowel modifiers are written as a continuation of the base consonant), a subset of sample points, carefully chosen before the final PEN UP signal, is taken to be the vowel modifier. It is worth re-emphasizing that the confused pairs in Group A correspond to CV combinations sharing the same base consonant (BC). Let ω_1 and ω_2 denote the class labels of BC+ ி and BC+ ீ combinations, respectively. We outline below the algorithm proposed for distinguishing ω_1 and ω_2. Note that as soon as any 'If' condition in the algorithm is satisfied, the corresponding class label is assigned to the CV combination and we terminate.

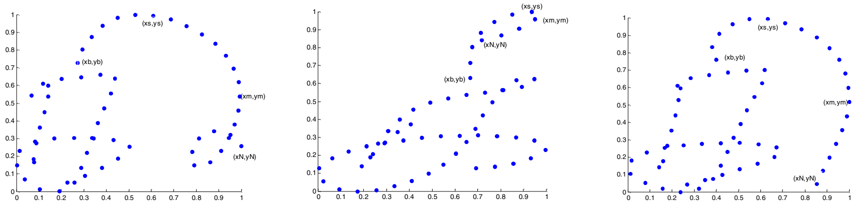
For the preprocessed CV combination resampled to N points, let $S = \{(x_i, y_i)\}_{i=b}^{N}$ denote the pen coordinates of the extracted vowel modifier. Here (x_b, y_b) denotes the starting sample point of the vowel modifier. A point (x_i, y_i) in S is said to be an 'interest point' if the following two conditions are satisfied.

(i) $y_i < y_{i-1}$ and $y_i < y_{i+1}$.

(ii) $x_{i+1} < x_i$. (1)

```
Using the aforementioned condition, compute the number of interest points $I$.
    Find the sample point $(x_s, y_s)$ satisfying the relation $y_s = \max_{i>=b} y_i$ .
    If $(x_s\ y_s)$ corresponds to the last sample point of the modifier,
            Accept class $\omega_2$ if $I$>0 and $\omega_1$ if $I$=0.
    End
    If $I > 0$
        Assign character to class $\omega_2$ (Fig. 1(a)).
    End
Locate the sample point $(x_m, y_m)$ satisfying the relation $x_m = \max_{i>s} x_i$.
Define the quantity $r = x_m - x_N$                                        (2)
If $r \geq \varepsilon$ and $y_N > y_b$
    Assign the character to class $\omega_2$ (Fig. 1(b));
else
    Assign it to class $\omega_1$ (Fig.1(c)).
End
```

Here ε is a threshold, empirically set to 0.07.



(a) (b) (c)

Fig. 1. Disambiguation of (Group A) confused pairs by structural features

(a) This sample is assigned to class கீ (by Step 2). Here $I = 1$. (b) This sample is assigned to கீ (by Step 3). Here $r > \mathcal{E}$, $y_N > y_b$ and $I = 0$ (c) This sample is assigned to கி (by Step 3). Here $r > \mathcal{E}$, $y_N < y_b$ and $I = 0$.

4 Disambiguation of Group B Confusion Pairs

The discriminative substrokes of confused pairs in Group B may fall in the middle of the trace. Hence, the post processing algorithms for this group must be able enough to automatically extract and analyze parts of strokes that differ. Unlike in Group A, where a single algorithm is used for disambiguation, separate post processing algorithms are generated for each confusion pair in Group B. The need for using dedicated post processing scheme to disambiguate a specific confusion pair arises mainly due to the structural characteristics of the pair under consideration. For example, the post processing scheme for extracting the discriminative sub stroke of the (ண ள) pair will be different from that used for extracting the sub stroke in (ழ ழு) pair. Note that, however, in both these pairs, the finer nuance that makes the confused characters distinct is at the middle of the online trace. In conclusion, for a given pair, the proposed algorithms for Group B take into consideration the temporal information, writing styles and structural cues of the confused characters. The illustrations that follow will throw light on the effectiveness of the methods in both sub stroke extraction and analysis.

A) Consider the characters ஜ and ஐ. Instead of feeding the (x,y) coordinates of these characters as a whole to the post- processing module, we focus on the shape of sub strokes forming the tails of these characters and extract Fourier descriptors from them. In order to extract the shape of interest, we compute the length of the character and divide it to 4 equal parts (segments). Sample points that lie in the last segment form the tail of the character and are resampled to 30 points before deriving the features. The number of Fourier coefficients chosen is set empirically to 10. A nearest neighbor classifier is used to obtain the final recognition label of a test character.

B) Consider the characters ண and ள. Since the subtle difference in these characters is observed in the middle of the trace, Fourier descriptors do not form a robust feature for discrimination. Our post processing algorithm first automatically extracts the substroke of interest as follows. Let $\{(x_i, y_i)\}_{i=1}^{N}$ be the sample points of preprocessed character (ண or ள). Locate the first minimum (x_c, y_c) that satisfies the condition $y_c < y_{c-1}$ and $y_c < y_{c+1}$. Starting from (x_c, y_c), move along the trace and locate the point (x_b, y_b) whose x coordinate just exceeds x_c. A substroke is extracted with (x_b, y_b) as the starting point. The extracted substroke can thus be described as

$$S = \{(x_i, y_i)\}_{i=b}^{N}. \tag{3}$$

Having extracted the substroke *S*, we shift our focus to analyzing the same. Let (x_{int}, y_{int}) be the first encountered minimum in *S* for which

$$y_{int-1} < y_{int} < y_{int+1}. \tag{4}$$

Fig. 2 provides a pictorial illustration of the aforementioned explanation.

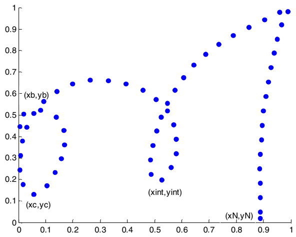


Fig. 2. Illustration of how (x_{int}, y_{int}) is located in the character ண

The following conditions are applied to the point (x_{int}, y_{int}) for finalizing our decision and thereby resolving the confusion.

1) If $x_{int-1} < x_{int+1,}$ assign symbol to ண. (Fig. 3 (a))

2) If the angle between successive pen directions defined at (x_{int}, y_{int}) is greater than 150 deg, assign

symbol to ண. (Fig. 3 (b))

Angle between successive pen directions at (x_{int}, y_{int}) refers to the angle formed by the vectors

$(x_{\text{int}} - x_{\text{int}-1} \quad y_{\text{int}} - y_{\text{int}-1})$ and $(x_{\text{int}+1} - x_{\text{int}} \quad y_{\text{int}+1} - y_{\text{int}})$.

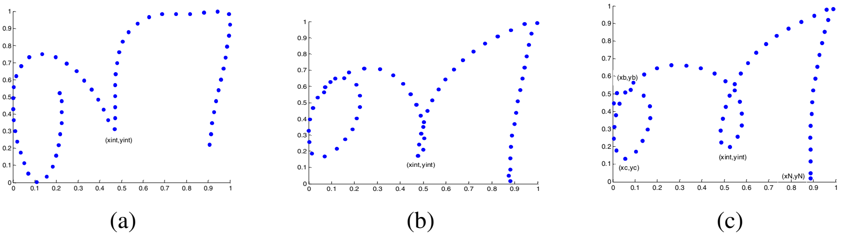
3) If points in the neighborhood of (x_{int}, y_{int}) are sufficiently close to each other (less than a threshold), the character is assigned to ண. (Fig. 3 (b))

The 'closeness' condition can be interpreted as follows: Let $W = \{(x_i, y_i)\}_{i=\text{int}-3}^{\text{int}+3}$ be a window size of 7 samples centered at (x_{int}, y_{int}). Using this window, compute three distances D_1, D_2 and D_3 as defined below

$$D_j = \text{dist}\,((x_{int-j} \quad y_{int-j})\,(x_{int+j} \quad y_{int+j})) \qquad j = 1, 2, 3 \tag{5}$$

Our final decision for deciding the label of the character can be formulated as follows: Assign character to ண if $\sqrt{D_1^2 + D_2^2 + D_3^2} < 0.1$. (6)

4) If neither condition hold good, character is assigned to ண. (Fig. 3 (c)).



(a) (b) (c)

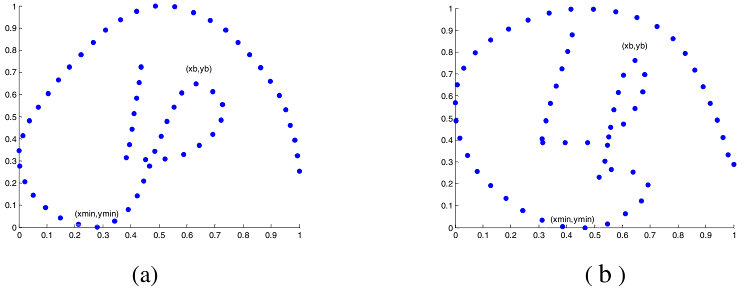
Fig. 3. (a) (b) indicate samples of ள that satisfy at least one of the conditions (1) (2) (3). Fig. 3 (c) depicts the sample ள that violates the above conditions.

C) We describe the strategy proposed to reduce the confusion between the symbols (ழ, ழு). The intelligence for segmenting out the substroke of interest, S is described as follows. Locate the first sample point (x_b, y_b) for which the following 2 conditions are simultaneously satisfied.

(i) $y_{b+1} < y_b$ and $y_{b-1} < y_b$.

(ii) $x_{b+1} < x_b < x_{b-1}$. (7)

(x_b, y_b) serves as the starting point for our substroke. The minimum y coordinate of the symbol y_{min} is considered to be the end point. Stated in another way, $S = \{(x_i, y_i)\}_{i=b}^{\min}$. Fig. 4 (a) (b) shows snapshots of ழ and ழு with the extracted sub strokes.



(a) (b)

Fig. 4. (a) (b). Samples of ழ and ழு with the desired sub strokes respectively.

We now proceed with analyzing the substroke S as follows. Calculate the number of sample points, f in S for which $y_{i+1} > y_i$. If f>0, the character is assigned to ழு. Otherwise ழ is selected. Using this condition, we see that the character in Fig. 4 (a) is assigned to ழ since f=0 for this sample. Fig. 4(b) provides a sample for which f=1. Hence, by our logic, it is assigned to ழு.

5 Experimental Results

The aforementioned techniques are tested on the IWFHR Competition dataset [6]. Table 2 depicts the recognition accuracies obtained by incorporating the post processing scheme for disambiguating confused characters. It can be noted that the ambiguity amongst the symbols have been resolved significantly, as indicated by the substantially high classification rates. The algorithms can also be extended to disambiguate confusion quadruples like (ணி ளி ணீ ளீ). Here, we first invoke the appropriate post processing algorithms to distinguish the vowel modifiers ி and ீ (Group A pair disambiguation) and then discriminate the base consonants ண and ள (Group B pair disambiguation). Employing the scheme to the quadruple gives a recognition rate of 96.2 %. However, it is important mentioning here that the accuracies quoted in Table 2 are applicable only to the confused characters in the pair under consideration. Overall, we have observed an improvement of 1% in the accuracy of the recognition system. Needless to say, the post processing methods can be applied irrespective of the classifier used for the recognition, though the nature of the confusion matrix may slightly vary.

Table 2. Recognition accuracies after invoking appropriate post processing algorithms to confusion pairs

GROUP A		GROUP B	
(நி நீ)	99.2%	(ஐ ஜ)	98.9%
(ணி ணீ)	98.5%	(ண ள)	98.2%
(ளி ளீ)	98.1%	(ணி ளி)	98.3%
(னி னீ)	97.6%	(ணீ ளீ)	97.2%
(கி கீ)	99.1%	(ழு ழூ)	97.5%

In this work, we have designed post processing algorithms to reduce the ambiguity between frequently confused Tamil characters. Structural cues are utilized in the design of these methods, which can be applied independent of the classifier used for the recognition. Future areas of research would involve the utilization of language models as a post processing module in place of structural cues.

Acknowledgements. The authors would specially like to thank TDIL, DIT for funding this research work.

References

1. Joshi, N., Sita, G., Ramakrishnan, A.G., Madhavanath, S.: Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition. In: Proc. Intl. Workshop Frontiers Handwriting Recog., pp. 444–449 (2004)
2. Deepu, V., Madhavanath, S., Ramakrishnan., A.G.: Principal Component Analysis for Online Handwritten Character Recognition. In: Proc. Intl. Conf. Pattern Recog., vol. 2, pp. 327–330 (2004)
3. Toselli, A.H., Pastor, M., Vidal, E.: On-Line Handwriting Recognition System for Tamil Handwritten Characters. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (eds.) IbPRIA 2007. LNCS, vol. 4477, pp. 370–377. Springer, Heidelberg (2007)
4. Prasad, S.: Online Handwriting Recognition for Tamil. Masters Thesis Report 2008. Indian Institute of Science, Bangalore, India (2008)
5. Sundaram, S., Ramakrishnan, A.G.: Improvement of Online Tamil Character Recognition Engine using Post Processing Methods. Accepted for publication in: Proc. Tenth Intl. Conf. on Doc. Anal. and Recog., ICDAR (2009)
6. HP Labs Isolated Handwritten Tamil Character Dataset, `http://www.hpl.hp.com/india/research/penhw-interfaces-1linguistics.html#datasets`

Hidden QIM Watermarking on Compressed Data Using Channel Coding and Lifting

Santi P. Maity, Claude Delpha, Sofiane Braci, and Rémy Boyer

Laboratiore des Signaux et Systèmes (L2S), CNRS, Universit*é* Paris-Sud XI (UPS),
SUPELEC, France
{santiprasad.maity, claude.delpha,
sofiane.braci,remy.boyer}@lss.supelec.fr

Abstract. This paper investigates the scope of application of channel coding on compressed host data for watermarking using dither modulation (DM) based quantization index modulation (QIM). Lifting based wavelet is used to decompose the encoded compressed data in integer coefficients. The relative gain on imperceptibility and robustness performance are reported for direct watermark embedding on entropy decoded host, using repetition code, convolution code, and finally the combined use of channel codes and lifting. Simulation results show that 6.24 dB (9.50 dB) improvement in document- to-watermark ratio (DWR) for watermark power at 12. 73 dB (16.81 dB) and 15 dB gain in noise power for watermark decoding at bit error rate (BER) of 10^{-2} are achieved, respectively over direct watermarking on entropy decoded data.

1 Introduction

Nowadays watermarking on compressed host data becomes highly demanding as it is rare to encounter any kind of multimedia signal in a raw, uncompressed format. However, watermarking and compression are antagonistic in characteristics and many watermarking methods developed for the uncompressed host data may not be suitable for the compressed data [1]. Least significant bits (LSB) of variable-length codes (VLC) in MPEG stream [2], JPEG data [3], entropy coded stream [4], and inter-block correlation of the selected DCT coefficients for JPEG compressed data [5], [6] are used for watermarking. Wu et al [7] maximizes robustness in the context of joint watermarking and compression (JWC).

This work applies channel coding on compressed host data for watermarking. Channel coded compressed host data is decomposed by discrete wavelet transform (DWT) using lifting to generate lossless integer coefficients. Watermark information is casted using dither modulation (DM) based quantization index modulation (QIM) for ease of implementation. The relative gain on imperceptibility and robustness performance are reported for direct watermark embedding on entropy decoded host, using repetition code, convolution code, and finally the combined use of convolution code and lifting.

The rest of the paper is organized as follows: Section 2 describes the proposed watermarking method on compressed data. Section 3 presents performance evaluation and discussion. Finally, the paper is concluded in Section 4 along with the scope of future work.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 414–419, 2009.

2 Proposed Watermarking Method on Compressed Data

Fig. 1(a) and Fig. 1(b) show the basic principle and the block diagram representation of the proposed watermarking scheme, respectively.

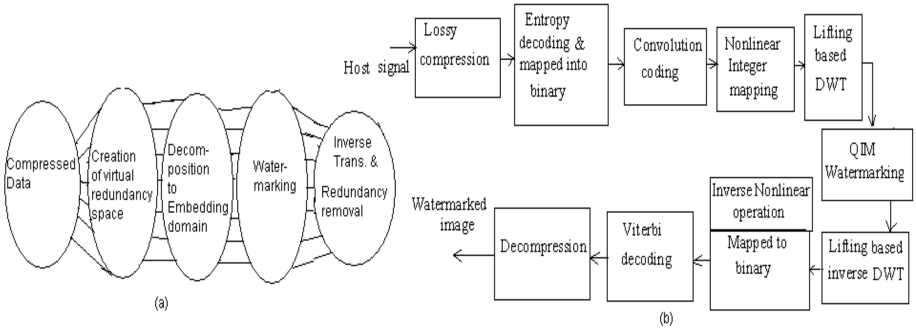

Compressed Data
Creation of virtual redundancy space
Decom-position to Embedding domain
Water-marking
Inverse Trans. & Redundancy removal
(a)
Host signal
Lossy compression
Entropy decoding & mapped into binary
Convolution coding
Nonlinear Integer mapping
Lifting based DWT
QIM Watermarking
Lifting based inverse DWT
Inverse Nonlinear operation
Mapped to binary
Viterbi decoding
Decompression
Watermarked image
(b)

Fig. 1. (a) Basic principle and (b) block diagram representation of the proposed watermarking scheme on compressed data

The lossy compressed host signal is entropy decoded. The non-zero quantized discrete cosine transform (DCT) coefficients are then mapped to the binary data. The binary data so obtained is then encoded using convolution coding to map ‘k’ bits into ‘n’ bits ($n \geq k$) in order to create a redundancy space in the compressed data and can be made use for watermarking. The convolution code is chosen as it operates on serial data and also uses memory system which in the present case helps to create correlation among the sample coefficients. The channel coded binary data is then converted to integer coefficients through a simple, easily implementable and reversible non-linear mapping like binary-to-decimal conversion. The integer coefficients thus obtained undergo lifting based discrete wavelet transform using 5-tap/3-tap filter coefficients [8], however, other lifting based DWT filters can also be used. Lifting-based filtering offers a benefit of very simple filtering operations for which alternately odd sample values of the signal are updated with a weighted sum of even sample values and even samples values are updated with a weighted sum of odd sample values [8] as below.

$$y(2n+1) = x_{est}(2n+1) - \lfloor \frac{x_{est}(2n) + x_{est}(2n+2)}{2} \rfloor \tag{1}$$

$$y(2n) = x_{est}(2n) - \lfloor \frac{y(2n-1) + y(2n+1) + 2}{2} \rfloor \tag{2}$$

where x_{est} is the extended input, y is the output signal and $\lfloor a \rfloor$, indicate the largest integer not extending ‘a’. Lifting on coded DCT coefficients makes watermarking compatible to both JPEG and JPEG 2000 compression operations.

A binary message ‘W’ is used as watermark and two dither sequences, with length L, are generated pseudo randomly with step size ($\varDelta$) as follows:

$$d_q(0) = \{\Re(key) \star \varDelta\} - \varDelta/2, \qquad 0 \le q \le L-1 \tag{3}$$

$$d_q(1) = \begin{cases} d_q(0) + \varDelta/2 \text{ if } d_q(0) < 0 \\ d_q(0) - \varDelta/2 \text{ if } d_q(0) \ge 0 \end{cases} \tag{4}$$

where $\Re(key)$ is a random number generator. The $q-th$ watermarked DWT coefficients S_q is obtained as follows:

$$S_q = \begin{cases} Q\{X_q - d_q(0), \varDelta\} + d_q(0) \text{ if } W(i,j) = 0 \\ Q\{X_q + d_q(1), \varDelta\} - d_q(1) \text{ if } W(i,j) = 1 \end{cases} \tag{5}$$

where X_q is the original channel coded q-th DWT coefficients, Q is a uniform quantizer (and dequantizer) with step $\varDelta$, and $W(i,j)$ is the (i,j)-th pixel of the watermark. Inverse lifting based DWT (IDWT) is then applied on the watermarked coefficients. Inverse non-linear operation maps each integer signal into binary data. The Viterbi decoding is then applied on the binary data to map each ‘n’-bits into ‘k’ bits. This operation is done for the inverse operation of channel coding i.e. for redundancy removal and not for watermark decoding. The Viterbi decoding is used due to its highly satisfactory bit error performance, high speed of operation, ease of implementation, low cost and fixed decoding time. We call this watermarking as “hidden QIM” as the information embedding process shown in Fig. 1(a) and Fig. 1(b) consist of (i) preprocessing of the compressed host data (using channel coding), (2) QIM embedding, and (3) post processing using channel decoding to form the composite signal [9].

The watermark information can be extracted from the compressed data using the following rule.

$$\begin{aligned} A &= \sum_{q=0}^{L-1} (\mid Q(Y_q - d_q(0), \varDelta) + d_q(0) - Y_q \mid) \\ B &= \sum_{q=0}^{L-1} (\mid Q(Y_q + d_q(0), \varDelta) - d_q(0) - Y_q \mid) \end{aligned} \tag{6}$$

where Y_q is the q-th DWT coefficient (possibly noisy due to transmission channel) of the received signal. A watermark bit $\hat{W}(i,j)$ is decoded using the rule:

$$\hat{W}(i,j) = \begin{cases} 0 \text{ if} \qquad A < B \\ 1 \text{ otherwise} \end{cases} \tag{7}$$

3 Performance Evaluation and Discussion

Fig. 2(a) shows (256 × 256), 8-bits/pixel gray scale watermarked Lena image when binary watermark image of size (32 × 32) (shown in Fig. 2(b)) is

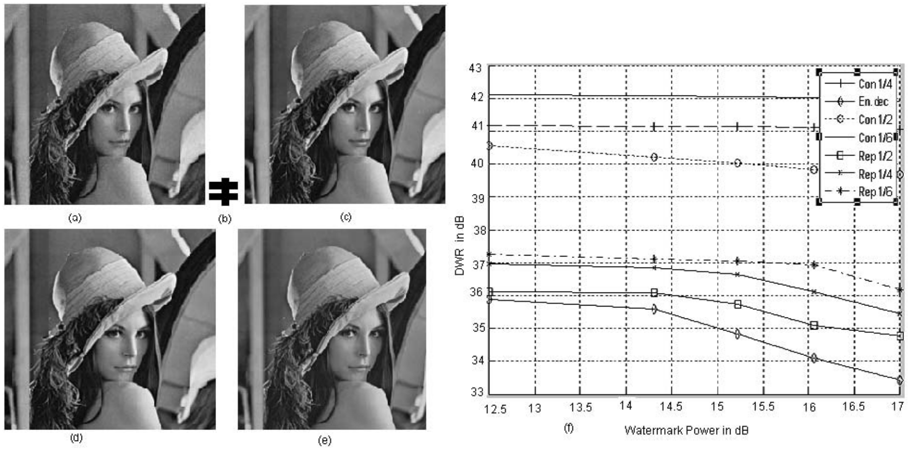


Fig. 2. (a),(c),(d) and (d) Watermarked images after embedding on entropy decoded coefficients, using lifting and convolution codes at rate 1/2, 1/4 and 1/6, respectively, (b) binary watermark image,and (f) DWR vs watermark power

embedded on entropy decoded DCT coefficients. Although experiment is carried out over large number of images, imperceptibility-robustness performance is reported here only for Lena image due to space limitation. Figs. 2(c)-2(e) show the watermarked images after embedding binary watermark using both lifting and convolution codes at rate r=1/2, 1/4 and 1/6, respectively and at quality factor 60. The numerical values of DWR for the Figs. 2(a), 2(c)-2(e) are 36.89 dB, 41.23 dB, 42.67 dB and 43.12 dB, respectively while the corresponding MSSIM (Mean Structural SIMilarity) index values are 0.9388, 0.9751, 0.9815 and 0.9856, respectively at watermark power 12.73 dB.

Fig. 2(f) and Fig. 3(a) show the graphical representation for DWR with change in watermark power at different code rates. It is quite clear from both the graphs that significant improvement in DWR is achieved due to the use of channel coding compared to the direct embedding of watermark information on the entropy decoded coefficients. The improvement is found to be higher in case of convolution codes compared to the repetition codes. The use of lifting in both cases shows relative improvement in DWR of the order of $\sim$ 0.5 dB but benefits in other way. A careful inspection on Fig. 2(f) and Fig. 3(a) show that the use of lifting with channel coding, particularly for convolution coding, maintains high DWR values even with large increase in watermark power leading to a significant improvement in BER performance against AWGN attack. The overall high DWR is achieved due to channel coding which is further augmented through the correlations among the sample coefficients due to the use of lifting and can be well explained by Eq. (1) and Eq.(2).

Fig. 3(b) shows BER performance and can be explained mathematically. The bit error rate (P_e) in watermark detection, for more general case of M-PAM signaling [10], is expressed as

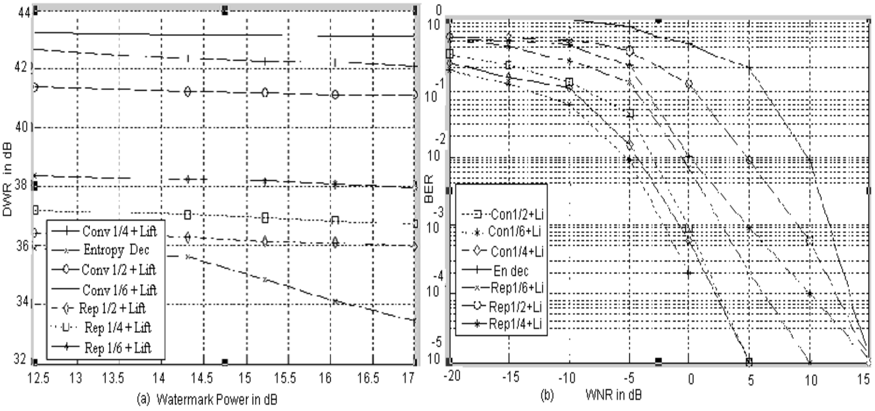


Fig. 3. (a) DWR vs watermark power for direct embedding on entropy decoded data and using both channel coding and lifting, (b)BER performance at different watermark-to-noise (WNR) power in dB

$$P_e = \frac{2(M-1)}{M}\Upsilon(\sqrt{\frac{Nd_0^2}{4\sigma_x^2}}) \tag{8}$$

where M corresponds to M-PAM (for the present case M=2), N is the gain in code rate in terms of the number of cover signal points over which each watermark bit is embedded, d_0^2 indicates the watermark power, $\Upsilon(.)$ indicates the complementary error function and σ_x^2 is the variance of the embedding coefficients. BER performance for watermarking on entropy coded data is poor as in such case N=1 and σ_x^2 is high. On the other hand, low (P_e) value for the decoded watermark is achieved due to two-fold advantages, namely large N-values due

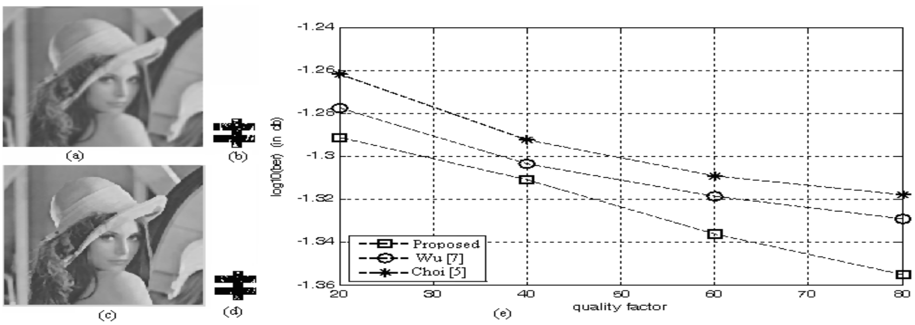


Fig. 4. (a), (c): Watermarked images after mean and median filtering, respectively, (b)and(d): Extracted watermarks from (a), and (c), respectively, (e) Robustness performance comparison against JPEG compression

to code rates and low σ_x^2 value compared to normal DWT coefficients as integer channel coded data is converted to integer coefficients due to the use of lifting.

Fig. 4(a) and Fig. 4(c) show the watermarked images with DWR 21.41 dB and 24.03 dB, respectively obtained after mean and median filtering operations using window sizes (11×11), while the extracted watermark images are shown in Fig. 4(b) and 4(d), respectively with BER values 0.0016 and 0.0011, respectively. Fig. 4(e) shows significantly improved robustness performance for our scheme compared to Choi [5] and Wu [7] method against JPEG compression operation. Simulation results also show similar robustness against JPEG 2000 compressions.

4 Conclusions and Scope of Future Works

A novel QIM watermarking is proposed using channel coding and lifting. Channel coding offers improvement both for imperceptibility as well as BER performance while lifting contributes much on BER performance. Present work is going on to implement the proposed algorithm for the compressed audio watermarking application. Future work may be carried out to design capacity optimized hidden watermarking scheme on the compressed data and real-time implementation through VLSI design using FPGA or ASIC.

Acknowledgement. The authors would like to thanks the French National Research program via the ANR project MEDIEVALS for Funding.

References

1. Wong, P.H., Au, O.C.: A capacity estimation technique for JPEG-to-JPEG image watermarking. IEEE Trans. on Cir. & Sys. for Video Tech. 13, 747–752 (2003)
2. Langelaar, G.C., Setyawan, I., Lagendijk, R.L.: Watermarking digital image and video data. IEEE Signal Proc. Mag. 17, 20–46 (2000)
3. Fridrich, J., Goljan, M., Chen, Q., Pathak, V.: Lossless data embedding with file size preservation. In: Proc. EI, Security, Steganography, Watermarking Multimedia Contents VI, vol. 5306, pp. 354–365 (2004)
4. Mobasseri, B.J., Berger, R.J.: A foundation for watermarking in compressed domain. IEEE Signal Proc. Lett. 12, 399–402 (2005)
5. Choi, Y., Aizawa, K.: Digital watermarking using interblock correlation. In: Proc. IEEE Int. Conf. Inf. Tech.:Coding and Compr., pp. 133–138 (2000)
6. Luo, W., Heileman, G.L., Pizano, C.E.: Fast and robust watermarking for JPEG files. In: Proc. IEEE 5th Southwest Symp. Image Anal. and Interp., pp. 158–162 (2002)
7. Wu, G., Yang, E.H.: Joint watermarking and compression using scalar quantization for maximizing robustness in the presence of additive gaussian attacks. IEEE Tran. Signal Proc. 53, 834–844 (2005)
8. Adams, M.D., Kossentini, F.: Reversible integer-to-integer wavelet transforms for image compression: performance evaluation and analysis. IEEE Trans. Image Proc. 9, 1010–1024 (2000)
9. Chen, B., Wornell, G.W.: Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Trans. on Information Th. 47, 1423–1443 (2001)
10. Voloshynovskiy, S., Pun, T.: Capacity security analysis of data hiding technologies. In: Proc. IEEE Int. Conf. Multimedia and Expo., Laussanne, Switzerland, pp. 477–480 (2002)

MRF Based LSB Steganalysis: A New Measure of Steganography Capacity

Debasis Mazumdar[1], Apurba Das[1], and Sankar K. Pal[2]

[1] CDAC, Kolkata, Salt Lake Electronics Complex, Kolkata, India
{debasis.mazumdar,apurba.das}@cdackolkata.in
[2] Indian Statistical Institute, Kolkata, India
sankar@isical.ac.in

Abstract. A new steganalysis algorithm is described based on the MRF model of image LSB plane. In this framework the limitation of the Cachin's definition of the steganography capacity is quantified and a new measure is proposed.

Keywords: Steganography, Cover object, Stego object, Ising model, partition function, ϵ-security.

1 Introduction

The classical definition of steganography is statistical and not perceptual [1]. Determining steganography capacity is an important research topic. As per Cachin's [2] definition a steganography scheme is said to be ϵ-secure if the Kullback-Leibler(K-L) divergence between the stego and the cover object is at most ϵ. However, in this model the statistical properties of the cover image has not been considered as a parameter. In the present work an attempt has been made to quantify the limitations of the Cachin's formula experimentally and a new measure of steganography capacity which is more relevant to the real life data is proposed.

We have developed a new steganalysis algorithm in which the LSB plane is modeled as a 2D Ising lattice, and Gibbs Markov statistical distribution is considered to be the a-priory distribution over the lattice. Using this model we have described the limitation of Cachin's ϵ-security in case of natural images. In the framework of Ising model the local randomness in the LSB plane is studied and its influence on the steganography capacity is computed. Based on these a new measure of steganography capacity is then derived. It is also shown that our theoretical model is coherent and in perfect agreement with the real life situations.

2 Statistical Structure of the LSB Plane

The statistical structure of the LSB plane is apparently very random, possessing no correlations, even among the neighbors. The heterogeneity or the disorderliness present in the LSB plane makes the process of detecting tampered bits more and more complex.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 420–425, 2009.

In Table-1 we represent some natural images, distribution of 1 and 0 in their LSB plane (white and black dots correspond to LSB value 1 and 0 respectively) and coefficient of kurtosis(α_4) of the distribution. It is clearly observed that, in some images, namely the first two, the coefficient of kurtosis is less than 3 and their LSB plane are highly random. In the other cases, the coefficient of kurtosis are more than 3 and long sequences of homogeneous LSBs are present. The images for which $\alpha_4 < 3$, detection of tampered bit is complicated. In the other set of images where $\alpha_4 > 3$, the detection is easier. The sharpness of the distribution in the latter case helps us in detecting the outliers easily. In order to develop a coherent steganalysis algorithm based on this statistical property of the LSB plane, we describe it as a 2D ising lattice having Gibbs-Markov probability distribution.

Table 1. Coefficient of kurtosis of the distribution of '1' and '0' in the LSB plane for different images

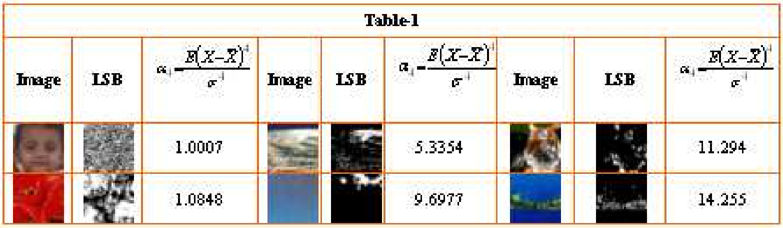


Table-1

Image	LSB	$\alpha_4 = \frac{E(X-\bar{X})^4}{\sigma^4}$	Image	LSB	$\alpha_4 = \frac{E(X-\bar{X})^4}{\sigma^4}$	Image	LSB	$\alpha_4 = \frac{E(X-\bar{X})^4}{\sigma^4}$
		1.0007			5.3354			11.294
		1.0848			9.6977			14.255

3 Ising Model of the LSB Plane

In our model, the LSB plane is represented as a 2D Ising lattice having the pixel values ±1. LSB based steganographic techniques either change the pixel values by ±1 or leave them unchanged. In this Ising lattice, let us introduce the notation η_{ij} representing the neighborhood of the $(i,j)^{th}$ pixel. If η_{ij} is perfectly homogeneous and isotropic, the entire lattice will be perfectly self-similar with respect to a basic building block, otherwise the heterogeneity or disorderliness sets in. We will show in the later part of the present paper that the fluctuation in the configuration of η_{ij} plays a significant role in determining the steganography capacity of the cover image. In order to introduce further geometrical notion into the model let us define $\Re_{ij}$ as the set of cliques associated with the neighborhood η_{ij}. The examples of neighborhood system and their associated cliques are discussed in many books and technical papers [3]. Let, (ξ, τ, ρ) define a Markov Random Field (MRF) over this Ising lattice, where ξ defines the set of all possible configurations, called the sample space, τ is the collection of subsets of ξ and ρ is the Gibbs distribution on this lattice, defined as:

$$\rho(x) = \frac{1}{Z} e^{-u(x)} \tag{1}$$

where u(x)=$\sum_{\forall(i,j)} u_C(x_C)$. Here u(x) is the potential function, $u_C(x_C)$ is the energy associated with the cliques C$\in \Re_{ij}$, and Z=$\sum_{\forall x} e^{-u(x)}$ is the normalizing constant, popularly known as the partition function. In our present work we have utilized a particular class of MRF with pair-wise interaction to model the LSB plane. This class of model has been extensively used for restoration of degraded binary images [4]. These applications have a close resemblance to our present problem. The joint density function (Eq. 1), in our model considers the first order neighborhood system where in each pixel interacts with its immediate neighbors as shown in the Fig. 1. We further assume that, the clique interaction energy follows the Hebbian rule [3]. Let, β_h and β_v be the coupling intensities in the horizontal and vertical direction respectively and $(i, j), (i\prime, j\prime)$ is a clique. The generalized expression of the interaction potential is:

$$u(x_{ij}) = \beta_h \sum x_{ij} x_{i\prime j\prime} + \beta_v \sum x_{ij} x_{i\prime j\prime} \tag{2}$$

We describe in the next section the algorithm to estimate the values of β_h and β_v for any image.

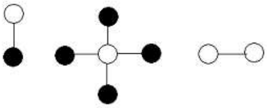


Fig. 1. Horizontal and vertical cliques

4 Estimation of the MRF Parameters and Detection of Tampered Bits

The problem of parameter estimation is an aspect of the research related to the modeling of any image using MRF. The principal cause of criticality lies in the computation of the partition function Z. In the present work EM(Expectation Maximization) based algorithm proposed by Besag et. al [5], is used to estimate the MRF parameters. Once β_h and β_v is estimated, the probability of occurrence $\rho(x)$ of each pixel is calculated using Eq. 1. If $\rho(x)$ is found to be significantly small, the corresponding pixel is considered as an outlier with respect to its neighborhood and declared as a tampered bit. In a complete image if the number of tampered bits crosses a threshold, then the image is declared as a stego image. In the following it will be shown that, the local randomness of the bit pattern around a pixel of the LSB plane is an important parameter in determining the accuracy of computation of $\rho(x)$. If for some random bit pattern gross inaccuracy occurs in the computation of $\rho(x)$ then its detection as stego/non-stego will be grossly inaccurate. We can say that such bit pattern offers more security to be considered as a secured cover image and the steganography capacity can be considered as a function of the % error of detecting tampered and untampered bits. In the next section we describe the influence of local randomness in the LSB plane on the estimation of tampered/ un-tampered bits. Based on this result a new measure of steganography capacity is also proposed.

5 Run Statistics of the LSB Plane and a New Measure of Steganography Capacity

The new steganalysis algorithm as described above, is tested over a large number of natural images. For space restriction a few results are presented here. In Fig. 2 the % of erroneously detected tampered and un-tampered bits is plotted as a function of the co-efficient of kurtosis (α_4), of the LSB plane. It is observed that for the images with high co-efficient of kurtosis α_4, detection of tampered/ un-tampered bits is less erroneous. It is also to be noted that, for small values of α_4, the detection accuracy is very poor and nearly the same for both the stego and non-stego images. As α_4 increases, the discernibility between the stego and non-stego images increases. To bring further clarity into this explanation we use the run statistics of images. We assume that any image can be considered as a set of rows (or columns). In each row, let there be N partition points (rank) which separate the sequences of homogeneous pixels (Fig. 3a. If in a row, there are N_A number of +1 pixels and N_B number of -1 pixels, the z-score of the Run-statistics is defined as [6]:

$$z = \frac{N - \mu_v}{\sigma_v^2} \tag{3}$$

where,

$$\mu_v = \frac{2N_A N_B}{N_A + N_B} + 1 \quad \mathit{and} \quad \sigma_v = \frac{2N_A N_B(2N_A N_B - N_A - N_B)}{(N_A + N_B)^2(N_A + N_B - 1)} \tag{4}$$

Let, the null hypothesis H_0 be that 'the row is random' and H_1 be the alternative hypothesis that 'the row is homogeneous'. For a two tailed test at the significance level 0.05, a row will be considered as random when $-1.96 \leq z \leq 1.96$ otherwise it will be considered that the row has long run length. In Fig. 3b the percentage

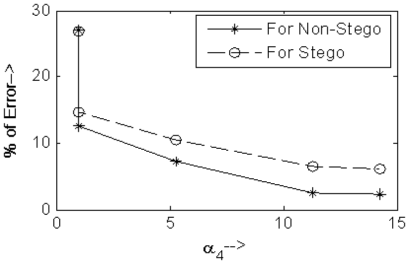
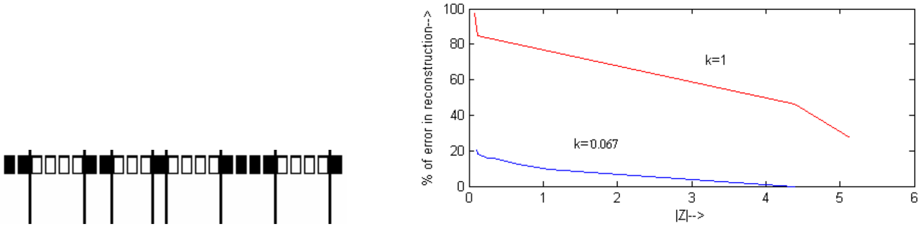


Fig. 2. Change in reconstruction error with coefficient of kurtosis



(a) (b)

Fig. 3. (a) Partition points(rank) separating the sequences of homogeneous pixels, and (b) Change in reconstruction error(%) with Z-score

error(e) in reconstruction is plotted as a function of the z-score. The curve is estimated numerically and a polynomial is obtained which fits into the curve closely.

$$e = k(-15.386|Z|^3 + 145.275|Z|^2 - 359.339|Z| + 125.5252) \qquad (5)$$

where k is a parameter which takes different positive non-zero values for different set of images. We now define the modified measure of steganography capacity as $\beta_\aleph$ where

$$D(P_S \| P_C) + \frac{1}{k(-15.386|Z|^3 + 145.275|Z|^2 - 359.3393|Z| + 125.525)} < \beta_\aleph \quad (6)$$

$D(P_S \| P_C)$ represents the K-L divergence between the cover and stego image as taken into consideration by Cachin in his definition of ϵ-security.

6 Experimental Results

We compute the K-L divergence between a pair of stego and non-stego images using the formulae:

$$D(P_C \| P_S) = log\frac{Z_S}{Z_C} - E_C\left[u_C - u_S\right] \qquad (7)$$

where Z_S and Z_C are the partition functions of the stego and cover image respectively and u_S and u_C are the clique potentials in the stego and cover image. $E_C[]$ represents the expectation of any random variable with respect to the distribution of the non-stego images. The K-L divergence for a number of natural images are computed. The images are chosen cautiously such that for each of them the K-L divergence remain nearly the same for particular amount of packing density. Following Cachin's definition, it is expected that, the % of error in bit reconstruction will remain same for all these images. In the Fig. 4 we see that the experimental curve is significantly deviated from the theoretical curve following Cachin's formulae. While the curve obtained using our formula is in close agreement with the experimental curve.

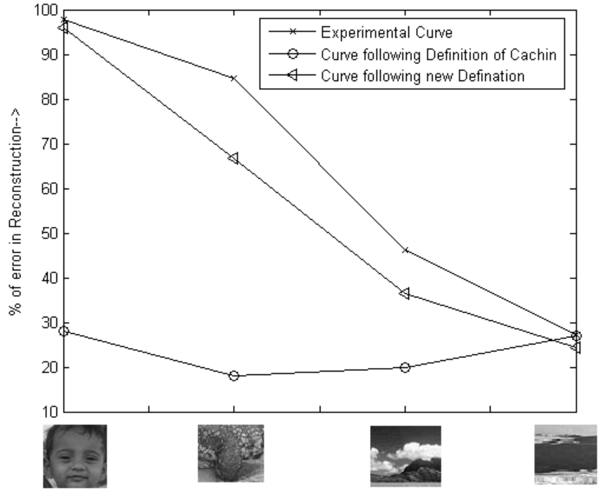


Fig. 4. Variation of the reconstruction error in different images. Experimental curve, curve obtained using Cachin's formula and the proposed measure of steganography capacity.

References

1. Mazumdar, D., Mitra, S., Dhali, S., Pal, S.K.: A chosen plaintext steganalysis of hide4pgp v 2.0. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 459–464. Springer, Heidelberg (2005)
2. Cachin, C.: An information-theoretic model for steganography. Information and Computation 1, 41–56 (2004)
3. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and baysian restoration of images. IEEE Trans. Pattern Anal. and Mach. Intell. 16, 721–741 (1984)
4. Yamasaki, T., Shirazi, M.N., Noda, H.: Adaptive restoration of degraded binary mrf images using em method. IEEE Transaction on Information and Systems E76-D, 259–268 (1993)
5. Besag, J.E.: Spetial interaction and the statistical analysis of lattice system. J. Roy. Statis. Soc. B, 36, 192–336 (1974)
6. Spiegel, M.R., Schiller, J., Srinivasan, R.A.: Theory and Problems of Probability and Statistics, 2nd edn. Tata McGraw-Hill Publishing Company Ltd. (1998)

Partial Encryption on SPIHT Compressed Images

Nidhi Taneja[1], Balasubramanian Raman[2], and Indra Gupta[1]

[1] Department of Electrical Engineering, IIT Roorkee, India
nidhi.iitr@gmail.com, indrafee@iitr.ernet.in
[2] Department of Mathematics, IIT Roorkee, India
balarfma@iitr.ernet.in

Abstract. Recently, Multimedia security paradigm has shifted towards encryption techniques combined with compression systems to meet the constraints of wireless networks. Several techniques of encryption with compression have been proposed to reduce processing time, but either they are insecure or computationally intensive. An efficient partial encryption technique based on SPIHT compressed bitstream is thus proposed to provide an efficient encryption for digital images. A satisfactory level of security is achieved with encryption of less than 0.5% of data for image encoded at 0.8 bpp by encrypting the bits as well as their locations. The proposed technique has large key space and can withstand approximation and brute force attacks.

1 Introduction

The past decade has witnessed an intensified growth in the development of joint compression and encryption techniques. Traditionally, multimedia data is compressed using a suitable compression algorithm followed by encryption of compressed output with an independent encryption algorithm. Selective encryption is mostly used to provide faster encryption by encrypting only a small portion of multimedia data, either from the final results or from the intermediate steps of compression system. This process must then be reversed by the decoder.

Methods have been proposed to combine partial encryption and compression to reduce the processing time, but either they are insecure or computationally intensive. Thus, a SPIHT based partial encryption technique is proposed that provides adequate security and consumes less computational resources. The proposed algorithm encrypts only the bit values, however, the encrypted bit locations can also be encrypted, providing a two tier security. Analysis has shown that the proposed algorithm gives a good perceptual degradation, low PSNR (Peak Signal to Noise ratio) value, a large key space, and can withstand brute force and approximation attacks. Subsequent section discusses the existing Set partitioning in Hierarchical Tree (SPIHT) based encryption techniques, proposed algorithm and its security analysis.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 426–431, 2009.

2 Existing SPIHT Based Encryption Techniques

Most of the compression algorithms generally decompose their input into important and unimportant parts to facilitate efficient encoding. Efficient encryption of only important part makes it difficult for a cryptanalyst to retrieve the correct image from the unencrypted data alone.

Cheng *et al.* [1] perceived that wrong initial information of two highest pyramid levels will disturb the entire tree structure formed and proposed to encrypt threshold level n and significant information of two highest pyramid levels in the SPIHT [2] encoded bitstream. Said [3] has argued that increased dependency does not necessarily mean high security. Said proved that an acceptable quality of image can be retrieved if the encrypted image and its respective watermarked thumbnail image is available.

Another partial encryption approach based on Color-SPIHT compression technique is proposed by Martin *et al.* [4]. The author proposed to encrypt all the LIP and LIS significant bits encountered during first K sorting passes. The confidentiality *vs.* processing overhead can be controlled by choosing K during encoding. Although, encryption time can be reduced by selecting only a part of the LIS bitstream.

Norcen and Uhl [5] proposed random permutation of wavelet coefficients in all the subbands before compression by SPIHT or JPEG2000 encoding. However, compression ratio reduces by 27% even when the permutation keys are not embedded in the encoded bitstream. Encryption of 3D-SPIHT coded video is proposed by Lian *et al.* [6] using sign bit encryption and two stages of wavelet coefficient confusion. But, coefficient confusion before encoding adversely affects the compression ratio.

To alleviate the problems incurred in above discussed techniques, a partial encryption technique based on SPIHT compression is proposed that gives high computational efficiency without adversely affecting the compression performance.

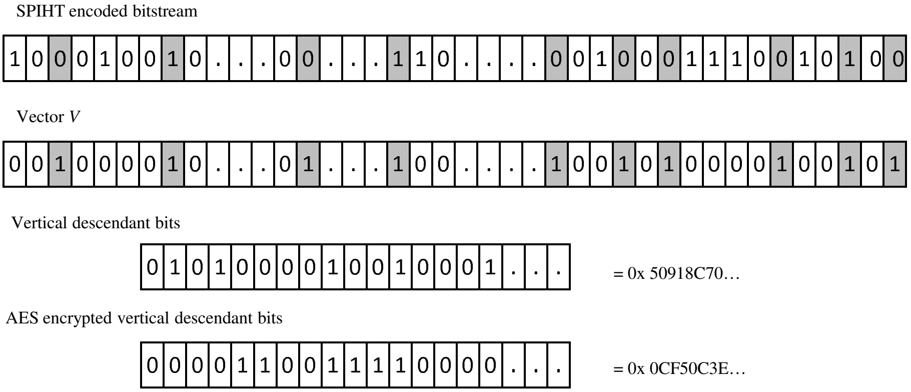
3 Proposed Scheme for Image Encryption

SPIHT generates an encoded bitstream which gives information regarding sign bits, refinement bits, significance of pixels, and significance of sets. The SPIHT encoded bitstream can be divided into three small bitstreams, denoting the bits belonging to LIP, LIS and LSP. The LIP bits denotes the ordered set of bits obtained during the first phase of sorting pass, where coefficients in LIP are tested for significance, LIS denotes the ordered set of bits obtained during the second phase of sorting pass where entire trees are tested for significance, and LSP denotes the ordered set of bits obtained during the refinement pass.

The decoder should consider every bit correctly to faithfully retrieve the image. A slight change in this encoded bitstream disrupt the tree structure, giving a wrongly decoded image. It is observed that encryption of sign bits and refinement bits do not contribute much to the security of images. Whereas, encryption of all the significant information consumes a lot of computational resources or time.

Thus, encryption of few significant bits representing the vertical descendants of LIS coefficients is proposed. This leads to only minor modifications in the tree structure, but degrades the image completely. The main advantage of the proposed scheme is that the number of bits to be encrypted can be controlled by the user to make an efficient trade off between computational resources and security. Secondly, the location of encrypted bits is not fixed and comes out to be different for different images. To provide sufficient security level, encryption is performed using a block cipher instead of stream cipher.

To identify the encrypted bits, a binary vector V is created during encoding, where $V_j = 1$ simply depicts that j^{th} bit in the encoded bitstream represents



SPIHT encoded bitstream

1 0 0 0 1 0 0 1 0 . . . . 0 0 . . . . 1 1 0 . . . . . 0 0 1 0 0 0 1 1 1 0 0 1 0 1 0 0

Vector *V*

0 0 1 0 0 0 0 1 0 . . . . 0 1 . . . . 1 0 0 . . . . . 1 0 0 1 0 1 0 0 0 0 1 0 0 1 0 1

Vertical descendant bits

0 1 0 1 0 0 0 0 1 0 0 1 0 0 0 1 . . . = 0x 50918C70…

AES encrypted vertical descendant bits

0 0 0 0 1 1 0 0 1 1 1 1 0 0 0 0 . . . = 0x 0CF50C3E…

Fig. 1. Systematic diagram for encryption

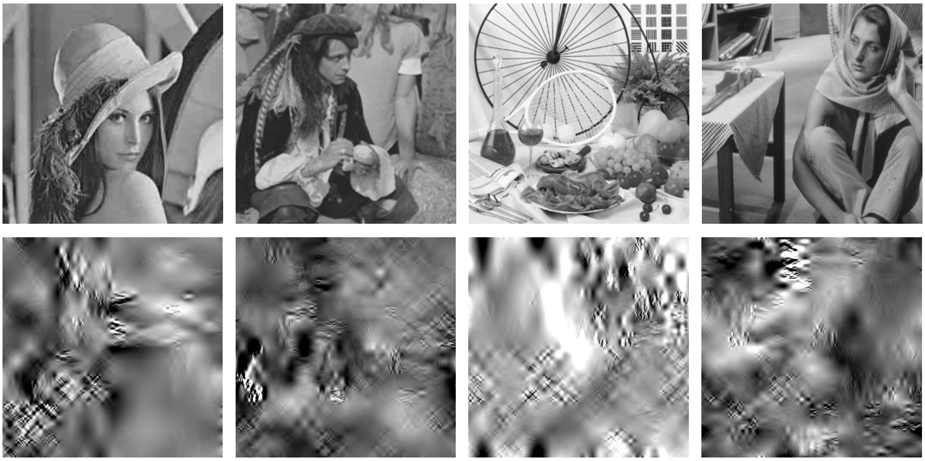


Fig. 2. Results for test images (left to right: lena, pirate, mix, barbara), first row: original image, second row: their encrypted counterparts

a vertical descendant bit. From the encoded bitstream, x number of bits for which $V_j = 1$ are grouped together, where x denotes the number of bits to be encrypted and controlled during encoding. These selected bits are then encrypted using AES, and placed back at their original positions in the encoded bitstream. The reconstructed image by decoding of this encrypted bitstream is completely incomprehensible. A systematic diagram depicting the encryption procedure is shown in fig. 1 and the results obtained by encryption are shown in fig. 2.

Other than the key used during AES, the vector V also becomes a key for the decryption process and transmitted along with the encrypted bitstream. Vector V is a variable length vector, different for different images, and also varies for different decomposition levels or threshold value in case of same image. This vector V can be encoded using (run, EOP) symbols which can further be protected by any public key encryption process. For decryption, the encrypted bits are selected from the encoded bitstream by comparison with the received vector V. Reverse process is carried out to reconstruct the original image.

4 Results and Discussions

An ideal encryption scheme should (a) ensure no information leakage from the cipher image (b) exhibit high key sensitivity, and (c) be able to withstand various cryptanalytic attacks. To analyze the security of the proposed technique, certain experiments on various grayscale images are perfromed. Fig. 2 shows the cipher images obtained after AES encryption of first 128 vertical descendants bits (less than 0.5% of the compressed bitstream).

The amount of degradation produced by the employed encryption technique is measured using Peak Signal to Noise Ratio (PSNR). The proposed encryption technique gives a PSNR value between 7dB-15dB, which is well within the satisfactory limits. The PSNR value obtained for various grayscale images are shown in table 1.

Table 1. PSNR values obtained

cameraman	*lena*	*pirate*	*barbara*	*woman*	*peppers*	*huts*	*wheel*	*mix*	*montage*	*butterfly*
9.89	12.29	13.09	12.52	14.19	8.75	10.80	12.44	9.96	7.84	13.76

During analysis, it is observed that in order to get first 128 vertical descendant bits, average length of vector V comes out to be 1500. In order to break the cipher, an intruder needs to know these encrypted bit locations which require $^{1500}C_{128}$ *i.e.* 1372! attempts. Even after knowing the bit locations, an opponent requires AES key to correctly decrypt the image. Collectively, a large key space is generated in the proposed cryptosystem, which is sufficient enough to resist brute force attacks.

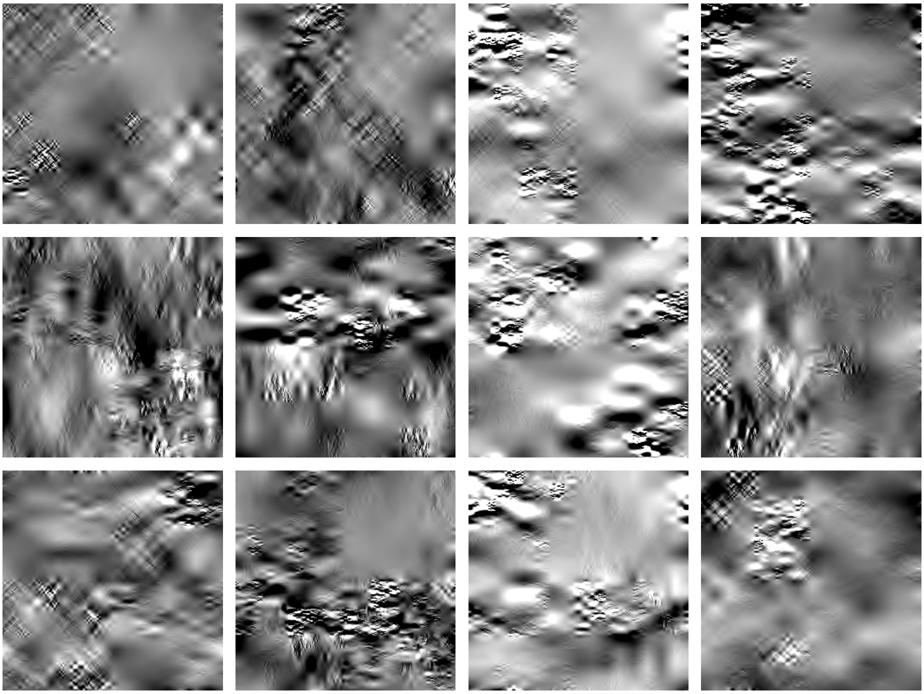


Fig. 3. Results for decrypted images (left to right: lena, pirate, mix, barbara), first row: wrong AES key (only single bit change in encryption and decryption key), second row: wrong vector V (only 5 bit locations are wrongly received), last row: approximation attack on the image (few encrypted values are replaced by 0 or 1, then decrypted and decoded)

The key sensitive nature of the proposed algorithm is tested by slightly varying the key used during encryption. In the first case, single bit change in the AES key at the LSB position is introduced. And, in the second case, the vector V used during decryption is changed. Randomly five bit locations are made different from the actual vector V. Fig. 3 shows the obtained result in both the cases and it can be seen that decryption using wrong key gives totally incomprehensible image.

The strength of encryption technique is also verified by launching approximation attack [7] in which all encrypted bit positions are replaced by a constant value *i.e.* 0's or 1's. The retrieved degraded images after approximation attack (fig. 3(last row)) do not reveal any information about the original image, thus, verifying the strength of proposed technique.

Use of AES encryption makes the proposed technique scalable in nature with first n bits forming the base layer, protected by a strong cipher. The rest of the bitstream constitute the enhancement layer providing bit level scalability. This enhancement layer can also be encrypted using a lightweight encryption scheme like XOR function.

4.1 Comparison with Existing Techniques

The proposed scheme is different from the scheme in [1,4] in three important ways: (a) computational overhead is reduced by encrypting only vertical descendant depicting bits of LIS coefficients, (b) the level of confidentiality achieved *vs.* computational overhead can be controlled by controlling the number of bits to be encrypted, and (c) encryption of vector V provides an additional security level. Further, the schemes in [5,6] are implemented before encoding, which adversely affects the compression ratio whereas the proposed technique is implemented on compressed bitstream without affecting the compression efficiency.

5 Conclusion

This paper proposes a partial encryption approach that encrypts selected data from the SPIHT compressed bitstream. Security is achieved by not only encrypting selected part of compressed bitstream, but can also be increased by encrypting the bit locations of encrypted bits. A good trade-off between security and computational resources is achieved without adversely affecting the compression ratio. The proposed technique provides a large key space, sufficient level of perceptual degradation, can withstand approximation attacks and brute force attacks.

References

1. Cheng, H., Li, X.: Partial encryption of compressed images and videos. IEEE Trans. Signal Process. 48(8), 2439–2451 (2000)
2. Said, A., Pearlman, W.A.: A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans. Circuits & Systems Video Tech. 6, 243–250 (1996)
3. Said, A.: Measuring the strength of partial encryption schemes. In: IEEE Int. Conf. Image Process., vol. 2, pp. 1126–1129 (2005)
4. Martin, K., Lukac, R., Plataniotis, K.N.: Efficient encryption of wavelet-based coded color images. Patt. Recogn. 38, 1111–1115 (2005)
5. Norcen, R., Uhl, A.: Encryption of wavelet-coded imagery using random permutations. In: Proc. IEEE Int. Conf. Image Process., pp. 3431–3434 (2004)
6. Lian, S., Sun, J., Wang, Z.: A secure 3D-SPIHT codec. In: Proc. European Signal Process. Conf., pp. 813–816 (2004)
7. Mao, Y., Wu, M.: A joint signal processing and cryptographic approach to multimedia encryption. IEEE Trans. Image Process. 15(7), 2061–2075 (2006)

Evaluation of Feature Selection Measures for Steganalysis

G.K. Rajput and R.K. Agrawal

School of Computer and Systems Sciences,
Jawaharlal Nehru University, New Delhi-110067
gauravrajput1@gmail.com, rka@mail.jnu.ac.in

Abstract. Steganalysis has attracted researchers attention overwhelmingly in last few years which discriminate stego images from non-stego images. The performance of a Steganalysis depends not only on the choice of classifier but also on features that are used to represent the image. Features extracted from images may contain irrelevant and redundant features which makes them inefficient for machine learning. Relevant features not only decrease the processing time to train a classifier but also provide better generalization. In this paper, kullback divergence measure, chernoff distance measure and linear regression are used for relevant feature selection. The performance of steganalysis using different measures used for feature selection is compared and evaluated in terms of classification error and computation time of training classifier. Experimental results show that Linear regression measure used for feature selection outperforms other measures used for feature selection in terms of both classification error and compilation time.

Keywords: Steganalysis, Feature Selection, Linear Regression, Kullback Divergence, Chernoff Distance Measure.

1 Introduction

Information hiding techniques have gained enormous popularity among researchers due to increasing threats for security, especially after 9/11 incident. Research community is involved in study of methods of secure communication as well as methods of detection of the covert communication between the two parties. Steganography is commonly used technique for data hiding in carrier signals. The carrier signal undergoes modification once the data to be sent is embedded using some technique. Images are often used as a carrier because of their extensive availability with high resolution of pixels. In recent years, detecting the presence of hidden messages has posed significant challenges to research community [2,3,4,6,10,17]. Steganalysis is art of discriminating such suspicious signals from a large number of inoffensive signals over a communication channel. Several techniques for steganalysis were proposed in literature [4,11,18]. Since the data hiding may involve different image formats, different embedding algorithms, and various steganographic keys, steganalysis turns out to be more difficult and challenging task. Recently, Farid [2,3] used wavelet-like decomposition to build higher-order statistical models of natural images in terms of mean, variance, skewness, and kurtosis of coefficients of wavelet subbands as features and employed classifier to discriminate between stego and non-stego images. However, the resultant features may contain noisy,

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 432–439, 2009.

irrelevant or redundant features which makes them inefficient for machine learning. In fact, the presence of irrelevant and redundant features may deteriorate the performance of the classifier and requires high computation time and other resources for training and testing the data. Hence, in order to enhance the performance of stegoanalysis in terms of accuracy and time required to detect, there is need to identify a set of relevant features.

Feature selection is used to remove such noisy, irrelevant, and redundant features. There are two major approaches to feature selection: filter and wrapper approach [5, 7, 8]. Most filter methods employ statistical characteristics of data for feature selection which requires less computation. It independently measures the importance of features without involving any classifier. Since, the filter approach does not take into account the learning bias introduced by the final learning algorithm, it may not be able to select the most relevant set of features for the learning algorithm. On the other hand, wrapper methods tend to find features better suited to the predetermined learning algorithm resulting in better performance. But, it tends to be computationally more expensive since the classifier must be trained for each candidate subset.

Feature ranking approaches have been widely investigated for feature selection [12, 14, 15] in literature. Since in most of feature ranking approaches features are evaluated using statistical characteristics of the data, different feature ranking methods measure different characteristics of data. Therefore, the informative features selected by different ranking methods may be different. Another disadvantage associated with feature ranking methods is that they ignore the correlation among the features because of their univariate approach. Hence the selected features subset may have low discriminatory capacity and increased redundancy. In literature to remove redundancy a forward/backward feature selection method or its combinations are used with a measure that selects relevant and non redundant features. Among the most widely used filter methods [6] for feature selection, there are techniques based on statistical separability measures which allow one to select a suitable subset of features by assigning the degree of interclass separability associated with each subset considered. In particular, Kullback Divergence, Chernoff distance measures and linear regression are commonly employed by research community. In this paper, we compare and evaluate these measures to determine relevant features for steganlaysis.

Our work is organized as follows: Feature extraction using higher order statistical model is included in section 2. A brief introduction of separability measures employed for features selection techniques are discussed in section 3. Experimental results on a database of natural images are shown in section 4 and section 5 contains conclusions.

2 Feature Extraction Using Higher Order Image Statistics

Although the presence of embedded message is most often not detectable to the human eye, but it may nevertheless changes the statistics of an image. The distortions in the resulting stegoimages can be analyzed by comparing the statistical properties of both cover and stegoimages [18, 19] . In literature, techniques [19, 20] are available to detect such changes based on first order statistical distributions of intensity or transform coefficients. The disadvantage of this analysis is that simple counter-measures that match first order statistics are likely to thwart detection. Farid [2] has pointed out that steganalysis based on higher-order statistical models may detect stegoimages. It has been observed across a large number of natural images that there exist strong higher-order statistical regularities within a wavelet-like decomposition. The embedding of a message may significantly change the statistics of image and thus becomes detectable.

The decomposition of image is possible by using separable quadrature mirror filters (QMFs). Thus, the frequency space is divided into multiple scales and orientations. This can be accomplished by applying separable lowpass and highpass filters along the image axes generating a vertical, horizontal, diagonal, and lowpass subband. The diagonal, horizontal, and vertical subband at scale $i = 1, 2, ..., n$ are represented as $D_i(x, y)$, $H_i(x, y)$, and $V_i(x, y)$ respectively. Subsequent scales are obtained by recursively filtering the lowpass subband. Farid [2,3] pointed out that using above decomposition the statistical model containing the mean, variance, skewness and kurtosis of the subband coefficients for each orientation and scales can be obtained for $i = 1$ to n. This characterizes the basic coefficient distributions statistically. The second set of statistics is based on the errors in an optimal linear predictor of coefficient magnitude. It is pointed out [2,3] that the subband coefficients of the image are correlated to their spatial, orientation and scale neighbors. Taking this into account,$V_i(x, y)$, a vertical band at scale i , can be represented in terms of neighboring pixels in spatial domain as:

$$\begin{aligned} V_i(x, y) = w_1 V_i(x-1, y) + w_2 V_i(x+1, y) + w_3 V_i(x, y-1) + w_4 V_i(x, y+1) \\ + w_5 V_{i+1}(x/2, y/2) + w_6 D_i(x, y) + w_7 D_{i+1}(x/2, y/2) \end{aligned} \tag{1}$$

Where w_k denotes scalar weighting values. In more compact form, it can be expressed as

$$\boldsymbol{V} = Q\boldsymbol{w}, \tag{2}$$

Where $\boldsymbol{w} = (w_1, w_2, ..., w_7)^T$, the vector $\boldsymbol{V}$ contains the coefficient magnitudes of $V_i(x, y)$ strung into a column vector and the columns of the matrix Q contain the neighboring coefficient magnitudes as specified in (1) also strung out in column vectors. To determine coefficients of quadratic error function is defined [2,3] as

$$E(\boldsymbol{w}) = [\boldsymbol{V} - Q\boldsymbol{w}]^2 \tag{3}$$

This error function $E(\boldsymbol{w})$ can be minimized by differentiating equation (3) and substituting it equal to zero yields:

$$\boldsymbol{w} = (Q^T Q)^{-1} Q^T \boldsymbol{V} \tag{4}$$

From the above, order statistics such as the mean, variance, skewness, and kurtosis can be evaluated. Similarly, the above procedure can be repeated to get the subbands $H_i(x, y)$, $D_i(x, y)$. Since, there are four statistics and linear predictor for three subband are computed for $(n-1)$ levels , we have total $12(n-1)$ error statistics and $12(n-1)$ coefficient statistics. Thus, these $24(n-1)$ statistics altogether will form a feature vector of the image.

3 Forward Feature Selection Techniques

Feature ranking is commonly used to determine a subset of relevant features. However, the disadvantage of feature ranking method is that they ignore the correlations between features. Hence the features selected may contain redundant information. Some of the methods suggested in literature for removing redundancy are Chernoff distance measure, Kullback divergence measure [14] and linear regression [16].

In order to obtain a quantitative measure of how separable are two classes, a distance measure can be easily extracted from some parameters of the data. A very important aspect of probabilistic distance measures is that a number of these criteria can be analytically simplified in the case when the class conditional p.d.f.s $p(X_i \mid C_i)$ follows multivariate normal distribution. In literature, for multivariate normal distribution for two classes, KD and CD measures are given as follows [14]:

$$J_k^d = \frac{1}{2}(\mu_k^2 - \mu_k^1)^T \left((\Sigma_k^1)^{-1} + (\Sigma_k^2)^{-1}\right)(\mu_k^2 - \mu_k^1) + \frac{1}{2} tr\left((\Sigma_k^1)^{-1}\Sigma_k^2 + (\Sigma_k^2)^{-1}\Sigma_k^1 - 2I_k\right) \tag{5}$$

$$J_k^c = \frac{1}{2}\beta(1-\beta)(\mu_k^2 - \mu_k^1)^T \left[(1-\beta)\Sigma_k^1 + \beta\Sigma_k^2\right]^{-1}(\mu_k^2 - \mu_k^1) + \frac{1}{2}\log\frac{|(1-\beta)\Sigma_k^1 + \beta\Sigma_k^2|}{|\Sigma_k^1|^{1-\beta}|\Sigma_k^2|^{\beta}} \tag{6}$$

where μ_k^i is a mean vector and Σ_k^i is a covariance matrix of k-dimensional data for class $C_i, i = 1, 2$. The regression analysis considers the relations between the selected features which minimizes redundancy. While using regression analysis for data a multiple regression model is considered because there can be many features which could affect the presence or absence of stegoimage. A multiple regression model with a target variable y and multiple variables X is given by [16]:

$$y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni} + \xi_i, \quad i = 1, 2, ..., n \tag{7}$$

Where $\beta_0, \beta_1, ..., \beta_n$ are constants estimated by observed values of X and class label y and is estimated by normal distribution having mean zero and a variance σ^2.

The error sum of squares SSE which is sum of the squared residuals is given by

$$SSE = \sum_{n=0}^{n}(y_i - predictedy_i) \tag{8}$$

A large value of SSE means that the regression is predicted poorly. The total sum of squares is given by

$$SSTO = \sum_{n=0}^{n}(y_i - \bar{y}) \tag{9}$$

Where $\bar{y}$ is the average of y_i. In a regression model the choice of features which best explains the class label depends on the value of $\mathcal{R}^2$ given by

$$\mathcal{R}^2 = 1 - \frac{SSE}{SSTO} \tag{10}$$

4 Experimental Setup and Results

We prepared a database of 1500 natural images taken from different sources i.e $www.1000pictures.com$, $www.1000wallpapers.com$ All the images were in JPEG format. The image resolutions were ranging from 800 × 600 to 1600 × 1200. We first resized each one of these images to 640 × 480 pixels images and embedded message images of six different resolutions 256 × 256, 128 × 128, 64 × 64, 32 × 32, 16 × 16, and 8 × 8 into cover image using OUTGUESS [19].

We have created 1000 non-stego images and 1000 stego images. Features are extracted from each one of the grey images using Haar wavelet. Each image is represented in terms of 72 statistics for four level wavelet decomposition. To remove redundancy from the selected pool of features three methods are compared: kullback divergence measure, chernoff distance measure and linear regression. For chernoff distance features are selected using 3 different values ranging from 0.1 to 0.9 with an increment of 0.4. We have used the following classifiers to evaluate the performance of the feature selection methods: naive Bayes Classifier (naivebc), Logistics linear classifier (loglc), Fisher linear classifier (fisherc), Nearest mean classifier (nmc), Normal Density based classifier-Independent features (udc), Normal Density based linear classifier (ldc), Normal Density based quadratic classifier (qdc). Classification error is computed using ten cross-validation. All the simulations are done using matlab. Tables 1-3 show the minimum classification error achieved with different classifiers along with the number of features for different measures. For chernoff distance measure the minimum classification error achieved for optimal value of β is shown in Tables 1-3. The best results in each category are indicated in bold. We observe the following from Tables 1-3:

1. The minimum classification error is achieved with linear regression for all classifiers and for different size of embedding.
2. The number of features required to obtain minimum classification error is significantly smaller using linear regression in comparison to baseline, kullback divergence measure and chernoff distance measure using all classifiers and different size of embedding.
3. The performances of ldc, fisherc and loglc are comparable and better than other classifiers in terms of classification error for all sizes of embedding used in experiments.

Table 1. Comparative results of classification error and minimum number of features for Size 8 × 8 and 16 × 16

	Size=8 × 8							Size=16 × 16						
	Baseline	JD		JC		Linear		Baseline	JD		JC		Linear	
	ErrAll	fea	ErrJD	beta	ErrJC	fea	ErrLin	ErrAll	fea	ErrJD	beta	ErrJC	fea	ErrLin
Ldc	0.3170	67	0.3045	38,.9	0.2985	23	**0.2905**	0.3210	39	0.3040	38,.5	0.2985	24	**0.2680**
Fisherc	0.3240	67	0.3045	46,.9	0.3000	24	**0.2905**	0.3160	34	0.3045	38,.5	0.2955	20	**0.2910**
Naivebc	0.5605	64	0.4705	16,.9	0.4430	2	**0.4360**	0.5390	27	0.4660	21,.9	0.4350	4	**0.3860**
Udc	0.4900	65	0.4860	22,.5	0.4875	9	**0.4260**	0.5025	55	0.4945	37,.9	0.4930	9	**0.3855**
Qdc	0.5000	37	0.4735	35,.5	0.4670	7	**0.3445**	0.5090	31	0.4760	24,.5	0.4675	7	**0.3085**
Nmc	0.4990	47	0.4970	52,.1	0.4950	4	**0.4165**	0.5020	3	0.4975	1,.1	0.4975	1	**0.4090**
Loglc	0.3090	63	0.3005	44,.1	0.2960	24	**0.2935**	0.3065	34	0.2995	44,.1	0.2985	18	**0.2670**

Table 2. Comparative results of classification error and minimum number of features for sizes 32×32 and 64×64

	Size=32×32							Size=64×64						
	Baseline	JD		JC		Linear		Baseline	JD		JC		Linear	
	ErrAll	fea	ErrJD	beta	ErrJC	fea	ErrLin	ErrAll	fea	ErrJD	beta	ErrJC	fea	ErrLin
Ldc	0.3155	54	0.3075	44,.1	0.3005	23	**0.2920**	0.3095	60	0.3090	43,.1	0.2990	37	**0.2925**
Fisherc	0.3210	51	0.3085	36,.9	0.3015	21	**0.2930**	0.3260	25	0.3130	37,.1	0.3020	19	**0.2925**
Naivebc	0.5385	19	0.4675	13,.9	0.4375	2	**0.4275**	0.5325	6	0.4725	15,.5	0.4420	2	**0.4330**
Udc	0.4985	38	0.4865	29,.5	0.4850	9	**0.4230**	0.5035	43	0.4925	47,.9	0.4895	11	**0.4225**
Qdc	0.5035	23	0.4650	22,.1	0.4595	2	**0.3400**	0.5025	26	0.4705	17,.5	0.4620	11	**0.3405**
Nmc	0.5030	72	0.4910	5,.5	0.4925	4	**0.4260**	0.5015	46	0.4960	2,.1	0.4940	4	**0.4135**
Loglc	0.3025	60	0.2965	66,.5	0.2975	17	**0.2925**	0.3135	66	0.3010	39,.9	0.2965	25	**0.2920**

Table 3. Comparative results of classification error and minimum number of features for Size 128×128 and 256×256

	Size=128×128							Size=256×256						
	Baseline	JD		JC		Linear		Baseline	JD		JC		Linear	
	ErrAll	fea	ErrJD	beta	ErrJC	fea	ErrLin	ErrAll	fea	ErrJD	beta	ErrJC	fea	ErrLin
Ldc	0.3195	64	0.3170	42,.1	0.3015	27	**0.2925**	0.2900	67	0.2820	55,.9	0.2675	24	**0.2680**
Fisherc	0.3355	69	0.3130	45,.5	0.3075	22	**0.2960**	0.3360	67	0.2770	46,.9	0.2730	20	**0.2965**
Naivebc	0.5085	25	0.4900	29,.5	0.4430	2	**0.4360**	0.4360	64	0.4330	22,.5	0.4125	4	**0.3860**
Udc	0.4955	69	0.4915	35,.5	0.4910	11	**0.4170**	0.4780	65	0.4765	71,.5	0.4710	9	**0.3855**
Qdc	0.5015	56	0.4875	34,.1	0.4730	12	**0.3415**	0.4745	37	0.4605	25,.5	0.4555	7	**0.3085**
Nmc	0.4985	22	0.4945	25,.5	0.4950	2	**0.3930**	0.4970	47	0.4865	45,.5	0.4930	1	**0.4090**
Loglc	0.3190	70	0.3190	41,.1	0.3020	22	**0.2905**	0.2835	63	0.2795	43,.5	0.2695	18	**0.2670**

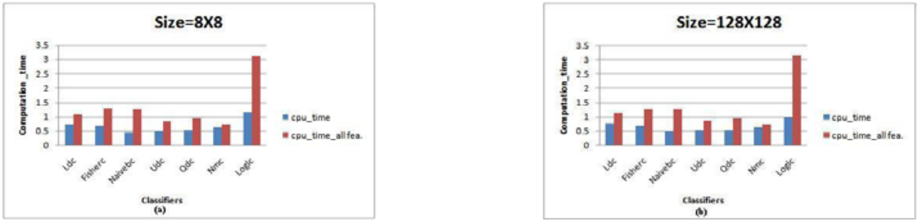


Fig. 1. Variation of Computation time (a) 8×8 (b) 128×128

The training time of different classifiers using features selected to achieve minimum classification error using linear regression measure and using baseline (all features) for two (8×8 and 128×128) sizes of embedding are shown in Figure 1. It can be observed from Figure 1 that the training time required using features selected by linear regression measure is very small in comparison to training time of classifier using baseline (all features) for all sizes of embedding used in experiments. The time difference observed is more in case of logc classifier.

5 Conclusions

The performance of a classifier depends on the choice of features and classifier for any pattern recognition system. Features based on higher order statistics are extracted from stego and non-stego images using wavelet decompositions. Feature extracted from image may contain irrelevant and redundant features which makes them inefficient for machine learning. Hence, relevant features which provide minimum classification error are selected using kullback divergence measure, chernoff distance measure and linear regression. The performance of steganalysis using different measures used for feature selection is compared and evaluated in terms of classification error and computation time of training classifier. Experimental results show that classification error with features selected using linear regression is less in comparison to other measures used for feature selection. Also the relevant features selected using linear regression is much smaller in size in comparison to other measures with all classifiers. Hence linear regression measure used for feature selection outperforms other measures in terms of training computation time of a classifier. To improve the performance of the Steganalysis, future work could include the following: (i) identifying the type of Steganography algorithm used for embedding the secret message, (ii) To locate, retrieve and analyze the embedded message to infer the information hidden in image.

References

1. Harmsen, J., Pearlman, W.: Steganalysis of additive noise modelable information hiding. In: Proc. SPIE Security and Watermarking of Multimedia Contents V, Santa Clara, CA, January 2003, vol. 5022 (2003)
2. Farid, H.: Detecting hidden messages using higher-order statistical models. In: Proc. of the IEEE int'l Conf. on Image processing 2002, vol. 2, pp. 905–908. IEEE, New York (2002)
3. Lyu, S., Farid, H.: Steganalysis using higher-order image statistics. IEEE Trans. Inf. Forensics Security 1(1), 111–119 (2006)
4. Mehrabi, M.A., Faez, K., Bayesteh, A.R.: Image Steganalysis on Statistical moments of Wavelet Subband histograms in Different Frequencies and Support Vector Machine. In: Proc. of Third int'l Conf. on Natural Computation, August 24-27 (2007)
5. Guyon, I., Elisseeff, A.: An Introduction to Variable and feature Selection. Journal of Machine Learning Research (3), 1157–1182 (2003)
6. Lie, W.-n., Lin, G.-S.: A Feature-Based Classification Technique for Blind Image Steganalysis. Proc. IEEE Transaction on Multimedia 7(6) (December 2005)
7. Kohavi, R., John, G.: Wrapper for feature subset selection. Artificial Intelligence (1-2), 273–324 (1997)
8. Ruiz, R., et al.: Incremental wrapper based gene selection from microarray data for cancer classification. Pattern Recognition 39(12), 2383–2392 (2006)
9. Fridrich, J., Goljan, m.: Practical steganalysis: State of the art. In: SPIE photonics West, Electronic Imaging, San Jose, CA (2002)
10. Johanson, N., Jajodia, S.: Exploring steganography: seeing the unseen. IEEE Computer, 26–34 (1998)
11. Westfeld, A., Pfitzmann, A.: Attacks on steganographic systems. In: Pfitzmann, A. (ed.) IH 1999. LNCS, vol. 1768, pp. 61–75. Springer, Heidelberg (2000)
12. Guyon, I., et al.: Gene Selection for cancer classification using support vector machine. Machine Learning (46), 263–268 (2003)
13. Duda Richard, O., Hart Peter, E., Stork David, G.: Pattern Classification, 2nd edn. Wiley India (P) Ltd., Chichester

14. Pierre, A.D., Kittler, J.: Pattern Recognition: A Statistical Approach. PHI (1982)
15. Tibsrani, R., et al.: Diagnosis of multiple cancer types by shrunken centriods of gene expression. Proc. Natl. Acad. Sci. USA (99), 6567–6572 (2002)
16. Han-Saem, P., et al.: Forward selection Method with regression analysis for optimal gene selection in cancer classification. International Journal of Computer Mathematics 84(5), 653–668 (2007)
17. Kahn, D.: The history of steganography. In: Anderson, R. (ed.) IH 1996. LNCS, vol. 1174. Springer, Heidelberg (1996)
18. Pevný, T., Fridrich, J.: Determining stego algorithm for JPEG images. IEEE Proc. -Iinf. Security 153(3) (September 2006)
19. Provos, N., Honeyman, P.: Hide and Seek: An Introduction to Steganography. In: IEEE Security and Privacy. The IEEE Computer Society, Los Alamitos (2003)
20. Wang, Y.: Optimized Feature Extraction for Learning-Based Image Steganalysis. IEEE Transactions on Information Forensics and Security @(1) (March2007)

Mutual Neighborhood Based Discriminant Projection for Face Recognition

Ben Niu[1], Simon Chi-Keung Shiu[1], and Sankar Pal[2]

[1] Department of Computing, Hong Kong Polytechnic University, Hong Kong, China
{csckshiu,csniuben}@comp.polyu.edu.hk
[2] Indian Statistical Institute, Kolkata, India
sankar@isical.ac.in

Abstract. Linear Discriminant Analysis is optimal under the assumption that the covariance matrices of the conditional densities are normal and all identical. However, this doesn't hold for many real world applications, such as Facial Image Recognition, in which data are typically under-sampled and non-Gaussian. To address this deficiency the Non-Parametric Discriminant method has been developed, but it requires model selection to be carried out for selecting the free control parameters, making it not easy for use in practice. We proposed a method, Mutual Neighborhood based Discriminant Projection, to overcome this problem. MNDP identifies the samples that contribute most to the Baysesian errors and highlights them for optimization. It is more convenient for use than NDA and avoids the singularity problem of LDA. On facial image datasets MNDP is shown to outperform Eigenfaces and Fisherfaces under various experimental conditions.

Keywords: *k*-nearest neighbors, mutual neighborhood, discriminant projection, face recognition.

1 Introduction

Linear Discriminant Analysis [1] is optimal for feature extraction assuming that the conditional densities are normal and all identical. However, for face recognition problems, datasets are often under-sampled and non-Gaussian. The Fisher-Raleigh criterion that defines LDA may lose its effectiveness for feature extraction and classification.

To address this problem, Fukunaga has developed the Nonparametric Discriminant Analysis method (NDA) [2] based on the idea of k-nearest neighborhood construction [3],[4]. Samples are assigned weights for training to increase the discriminant power of the extracted features. However, the weight function of NDA contains free parameters that have to be configured through model selections. They affect the weight function exponentially and range from 0 to the infinity. Users have to be careful in choosing them for maximal performance, making it not easy for use as LDA in practice.

We proposed an alternative approach, Mutual Neighborhood based Discriminant Projection (MNDP) to address this problem. The idea is that a majority of the Bayesian errors in classification arise from the samples that stay very close but belong to different

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 440–445, 2009.

classes. To improve classification, we should find out and highlight them for discriminant feature extraction. This can be achieved by incorporating the idea of mutual k nearest neighborhoods. That is, if 2 samples in the data space are within the k nearest neighborhood of each other then we pair up and pick out them for optimization to reduce the recognition errors. Differing from NDA, MNDA doesn't require the extra costs for choosing the weight parameters in model selection. Also, we can learn more projection vectors in NDA than in LDA to improve the recognition rate. It doesn't have the singularity problem of LDA. We have done experiments to evaluate MNDP on the AR databases. Our results indicate that MNDP outperforms Eigenfaces and Fisherfaces significantly and consistently for face recognition.

2 Related Works

For 2 class problem, NDA is formulated by Eq. (1)

$$W_{opt} = \arg\max_{W} \frac{\left|W^{T}\tilde{S}_{B}W\right|}{\left|W^{T}S_{W}W\right|}, \tag{1}$$

where $\tilde{S}_B = \frac{1}{N_1}\sum_{u=1}^{N_1} w_u (x_u^1 - m_2(x_u^1))(x_u^1 - m_2(x_u^1))^T + \frac{1}{N_2}\sum_{v=1}^{N_2} w_v (x_v^2 - m_1(x_v^2))(x_v^2 - m_1(x_v^2))^T$.

N_1 and N_2 are the number of samples in class 1 and class 2. x_u^1 and x_v^2 are 2 samples from class 1 and class 2. $m_2(x_u^1)$ and $m_1(x_v^2)$ are the mean vector of the k nearest neighbors of x_u^1 in class 2 and x_v^2 in class 1. w_u is the weight assigned to x_u^1,

$$w_u = \frac{\min\{d^{\alpha}(x_u, x_{kNN}^1), d^{\alpha}(x_u, x_{kNN}^2)\}}{d^{\alpha}(x_u, x_{kNN}^1) + d^{\alpha}(x_u, x_{kNN}^2)}. \tag{2}$$

In Eq. 2, x_{kNN}^1 and x_{kNN}^2 are the *k*-th nearest neighbor of x_u in class 1 and class 2, $d(\cdot,\cdot)$ is a distance measure, $d(x_u, x_{kNN}^1)$ and $d(x_u, x_{kNN}^2)$ are the radii of the kNN neighborhoods, α is a free control parameter that ranges from zero to the infinity. The function has the property that near the classification boundary the weight takes large values and drops off to zero as going faraway. The parameters, α, determine how fast w_u falls to zero. They range from 0 to the infinity and affect the weight function exponentially. Users have to be careful in choosing these parameters for performance, making it not easy for use as LDA in practice.

3 Concept of MNDP

Let X denote the space of observations, $X \subseteq R^n$. $X_1, X_2 \subset X$, $X_1 \cap X_2 = \phi$ are 2 classes of training samples. The between-class mutual *k*-nearest neighborhood is defined as the set of the sample pairs with mutual neighbor relations,

$$P = \{(x_i^1, x_j^2) \mid x_i^1 \in X_1, x_j^2 \in X_2, x_i^1 \in \mathrm{kNN}(x_j^2), x_j^2 \in \mathrm{kNN}(x_i^1)\}, \tag{3}$$

x_i^1 and x_j^2 is the i-th and the j-th sample in classes X_1 and X_2, respectively. $\mathrm{kNN}(x_i^1)$ is the set of the *k*-nearest neighbors of x_i^1 in X_2. $\mathrm{kNN}(x_j^2)$ is the set of the *k*-nearest neighbors of x_j^2 in X_1. By building the mutual *k*-nearest neighborhoods, we can readily identify the pairs of samples who are the mutual neighbors, as shown in Fig. 1.

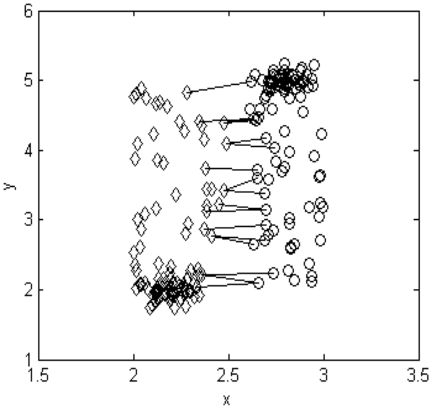


Fig. 1. Mutual neighbors identified for 2 data classes (represented as circles and diamonds).

4 Formulation of MNDP

The objective function of MNDP is defined by,

$$W_{opt} = \arg\max_{W} \frac{\left|W^T S_N W\right|}{\left|W^T S_F W\right|} \tag{4}$$

S_N is the between-class mutual kNN scatter matrix, $S_N = \frac{1}{M}\sum_{i=1}^{M}\left(x_1^i - x_2^i\right)\left(x_1^i - x_2^i\right)^T$,

M is the number of the sample pairs in the mutual kNN neighborhood, $\left(x_1^i, x_2^i\right) \in P$. S_F

is the within-class kFN scatter matrix, $S_F = \frac{1}{L}\sum_{j=1}^{L}\left(x_1^j - x_2^j\right)\left(x_1^j - x_2^j\right)^T$, L is the number of the sample pairs in the kFN neighborhoods, $\left(x_1^j, x_2^j\right) \in Q$. The MNDP projections, W_{opt}, are obtained by solving the eigenvectors for the matrix $S_F^{-1}S_N$.

We demonstrated the effectiveness of MNDP over LDA using simulated data, as shown in Fig. 2. LDA and MNDP both work well for normal distributions (Left). But LDA fails to learn the optimal projection when the densities are not typically Gaussian (Right). However, MNDP can work well under both conditions.

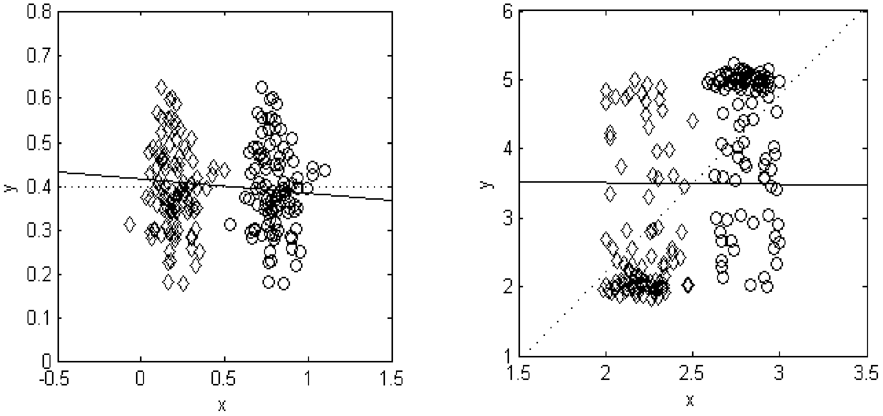


Fig. 2. First component of LDA (dot line) and MNDP (solid line)

5 Experimental Results

We evaluated MNDP for Face Recognition and compared it with Eigenfaces and Fisherfaces. The results on AR databases [5] indicate that MNDP achieves better recognition rate under various experiment conditions than others. We found that intriguingly it is most effective when the Cosine distance measure was utilized for classification on AR database. The accuracy rate was improved about 10 percent.

5.1 AR Database

We selected the sample images from 126 individuals, manually cropped and normalized them to 50×40 pixels. The neighborhood size was set to 3. For each person 7 samples out of the 14, as shown in Fig. 3, are chosen randomly for training, the remainders for testing. The L_1, L_2 norms and the Cosine distances are utilized for *1*-NN classification. We repeated the experiment 20 times. The recognition rates, their standard deviations, and the number of components employed for classification are shown in Table 1.

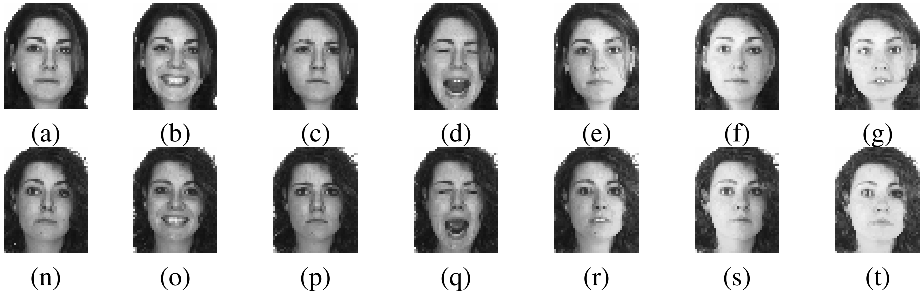


Fig. 3. Sample images of one subject from AR database

Table 1. Top average recognition rate (%) on AR database

Method	Distance measures		
	L1	L2	Cosine
Eigenfaces	92.08 ± 1.23(80)	80.92 ± 1.23(80)	85.48 ± 1.23(80)
Fisherfaces	84.61 ± 1.32(79)	82.74 ± 1.21(80)	87.59 ± 1.28(55)
MNDP	95.14 ± 1.12(40)	90.81 ± 0.81(41)	95.57 ± 0.81(40)

6 Conclusion

Linear Discriminant Analysis is optimal under the assumption that the covariances of the conditional densities are normal and all identical. This however doesn't hold for many real world applications, such as Face Recognition, in which data are typically under-sampled and non-Gaussian. Fukunaga has developed the Non-Parametric Discriminant method to address this deficiency but it was not quite easy for use as model selections have to be carried out to choose the optimal weight parameters. We proposed an alternative approach, Mutual Neighborhood based Discriminant Projection, to overcome this problem. MNDP highlights the mutually neighboring samples, which are considered to contribute most to the Baysesian errors for discriminant feature analysis. This was achieved by utilizing the idea of mutual k nearest neighborhood. MNDP is easier for use than NDA and avoids the singularity problem one of the limitations of LDA. On facial image datasets, MNDP achieved better recognition result than Eigenfaces and Fisherfaces under various experiment conditions.

References

1. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, Boston (1990)
2. Fukunaga, K., Mantock, J.: Nonparametric Discriminant Analysis. IEEE Trans. Pattern Anal. Machine Intell. 5, 671–678 (1983)

3. He, X., Cai, D., Yan, S., Zhang, H.J.: Neighborhood Preserving Embedding. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 1208–1213 (2005)
4. Hu, D., Feng, G., Zhou, Z.: Rapid and brief communication: Two-dimensional Locality Preserving Projections (2DLPP) with its Application to Palmprint Recognition. Pattern Recognition 40(1), 339–342 (2007)
5. Martinez, A., Benavente, R.: The AR Face Database. CVC Technical Report, no. 24 (June 1998)

Biometric Gait Recognition with Carrying and Clothing Variants

Shamsher Singh and K.K. Biswas

Dept. of CSE, IIT Delhi-110016, India
sskarki@gmail.com, kkb@cse.iitd.ac.in

Abstract. Compact spatio temporal representation of human gait in form of gait enery image (GEI) has attracted lot of attention in recent years for biometric gait recognition. Researchers have reported very high recognition rates for normal walk sequences. However, the rates come down when the subjects are wearing a jacket or coat, or are carrying a bag. This paper shows that the performance for the variant situations can be improved upon considerably by constructing the GEI with sway alignment instead of upper body alignment, and selecting just the required number of rows from the bottom of the silhouette as inputs for an unsupervised feature selection approach. The improvement in recognition rates are established by comparing performances with existing results on a large gait database.

1 Introduction

Biometric gait recognition refers to identifying a person from his normal walking style (gait). This is a marker less unobtrusive biometric which can be done at a distance and does not need user interaction or cooperation. [1][2][3][4][5]

Recently there has been a lot of interest in using gait energy as a feature for gait based recognition[6][7][8]. The silhouettes of a person are extracted from each frame of a walking cycle and superimposed on the first frame after aligning the upper body of the person. This single image is referred to as Gait Energy Image (GEI). This represents the walking dynamics of the person in a compact manner. It has been shown that this representation is less sensitive to noise and recognition rates are very promising for normal walk sequences[6][7][8]. However the performance of the recognition system goes down if the person is wearing a coat or is carrying a shoulder bag. This paper shows that by aligning the silhouettes in a slightly different manner and with a proper selection of energy points from the GEI, the gait recognition performance can be improved for clothing and carrying bag variants of walking. The performance has been measured with the same data set as used in Bashir[8].

2 Generating the GEI

We assume that the silhouettes of a person have already been extracted from the walk sequence video clips. Many data bases which are available online, provide

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 446–451, 2009.

such silhouettes. We have chosen the CASIA database[9] for our study. We define a walk cycle to consist of frames between two identical key frames consisting of legs together stances of a person. Using the white pixel coverage of lower portion of the body one can determine these key frames. We then normalize the silhouette height to a window of height 100 pixels.

The conventional way of projecting the silhouettes is to align the head and upper torso of the subject in each frame with the corresponding torso of the first frame. In this approach apart from dynamic hand motion information the upper body part appears static in the final GEI. However, as a person walks, his whole body sways forward and backward periodically and this information is lost in the simple alignment technique. This alignment also influences the shape of the the dynamic area of the lower body part as well, which is critical in defining a persons gait characteristics.

We propose a method of alignment that takes into consideration the whole body motion. For each frame we use the maximum and minimum horizontal points of the silhouette to compute the horizontal centroid and use this to align the silhouettes. These points do not necessarily belong to upper body part and hence depicts true horizontal centroid . We refer to this as "sway alignment" in our work.

3 Selection of Feature Points

To reduce the dimensionality of feature points , PCA is employed both for unsupervised or supervised selection approaches[7][8]. Supervised feature selection is tried on the bottom half of the GEI in[8]. It is obvious that more useful gait information is contained in the lower part of the body, particularly the legs. Also it should be noted that this part is usually unaffected, if the person wears a coat or a jacket or is carrying a bag on his shoulders. Thus we concentrate on the dynamic lower part by selecting a suitable area from the bottom of the silhouette, The motivation behind this is to use the pure gait information without the influence of shape. Also, it drastically reduces the number of feature points needed for training.

To get an approximate position of the knee, we first locate the lowest position of the palm using the golden ratio (1.618) of the human body[10]. As the silhouette height is 100 pixels in our case, we get palm position at row 62 from the top of the head. The knee would thus be located within 38 rows from the bottom of the silhouette. Our desired dynamic area now consists of silhouette rows between the ground level and a chosen limiting row. We name the limiting row as R_{lim} (which is less than 38). By carrying out exhaustive experimentation we found that R_{lim} can be set comfortably between 25 and 30. We keep the R_{lim} fixed for all the gallery and probe sequences of all the persons.

4 Similarity Analysis

4.1 Dataset

We are using the CASIA dataset created by the center for biometric and security research[9]. This database involves 124 subjects and was created with uniform background and controlled lighting condition and with different viewing angles. The Database consists of three variants which are carrying bag, different clothing and view angle. For first two variants 2 walk per person are available while 11 cameras at different angles are used for each walk. The starting stance of the walks is not fixed.

Our approach is aimed for single camera based systems and applicable in scenarios where restrictive entry is allowed. In GEI based approaches good results are reported when view angle is 90^0. We have used the silhouettes of 90^0 view angle from the dataset. The data is partitioned in three sets.
set A: Six normal walks for the same person
set B: Two walks by the same person but now carrying a bag
set C: Two walks by the same person with clothing change
To do the similarity analysis the whole data is divided into probe and gallery. The gallery set(Ag) is made of first four normal walks for each person. We chose this data set to compare our results with existing methods where the same very walks are used as the gallery set[8]. As our interest is to show that recognition can be carried out even when a person is carrying a bag, the Probe data (for recognizing a person) consists of 3 sets(Ap,B,C). Ap happens to be the last two walks of 6 normal walk data of set A. For both the probe and gallery sequence, we have generated two GEIs per walk, belonging to alternate gait cycles.

4.2 Analysis

As mentioned earlier, we are only using the feature points from dynamic area of GEI,bounded by R_{lim}. Similarity between probe and gallery walks is estimated by calculating the normalized sum of absolute differences between the Gray values of two GEI as shown in equation 1. Nearest neighbor classifier is then used for recognition. Figure 1 depicts the process of similarity analysis between the gallery sequence of $subject_i$ and the k^{th} probe sequence $subject_j$. As mentioned earlier for each subject four walk sequence of gallery and two walk sequence of probe are used where each walk sequence consists of two GEIs.

$$\begin{aligned} Diff_i &= Abs(g_i(p) - g_i(g)) \\ Diff &= \Sigma_i^n Diff_i/n \end{aligned} \tag{1}$$

$g_i(p)$: Gray value at location i for probe p
$g_i(g)$: Gray value at location i for gallery g
Diff : Normalized difference between probe(p) and gallery(g)
n : Number of points used, depends on R_{lim}

For each of the two probe sequences we calculate average difference with the gallery as shown in equation 2

$$AvgDiff(pr_k) = MIN1 + MIN2/2 \tag{2}$$

k : 1,2 denote two probe sequence
MIN1, MIN2: Least 2 values among Min(a1,b1), Min(a2,b2), Min(a3,b3) and Min(a4,b4)

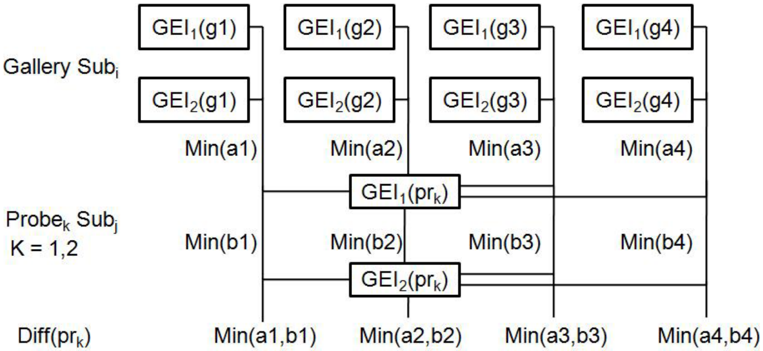


Fig. 1. Similarity analysis

Final similarity between Gallery Sub_i and Probe Sub_j is estimated by calculating the minimum average difference among the two probe sequences as show in equation 3.

$$sim(Sub_i, Sub_j) = min(AvgDiff(pr_k)) \tag{3}$$

5 Performance Evaluation

We have evaluated the performances of performed the gait analysis using upper body alignment and using our "sway alignment". For the CASIA dataset database we have calculated recognition rates for various rank values. Rank N recognition indicates that the person in the probe set is present among the top N similarity scores from the gallery.

We have performed the gait analysis using both the upper body alignment as well as sway alignment. We have calculated rank 1, rank 2 and rank 5 identification rates for each of the variant. Table 1 presents the results for R_{lim} =30 and table 2 shows the results for R_{lim} = 25. It is observed that in each case the sway alignment performs better in case of normal walk(set Ap) and clothing(set C) variant but upper body alignment performs better in case of carrying bag(set B) condition.

Comparison with existing results. We have compared the results of our experiments with CAS(Chinese Academy of Sciences): GEI shape match approach [6],], UCR(Univ. of California Riverside): PCA and MDA base feature learning method[7] and (Queen Marys, Univ of London): Feature selection based approach[8]. We have used the result of sway alignment for normal walk and clothing variant while upper body alignment based results have been presented for carrying bag variant. We have used Rlim as 30 for our results.

Both the CAS and UCR methods use the whole of the GEI for analysis. UCR employs PCA and MDA for feature learning. Table 3 presents the comparison

Table 1. Results: Static feature selection($R_{lim} = 30$)

Method	Identification rate	Set Ap	Set B	Set C
Upper body alignment	**Rank 1**	**88.52**	**74.59**	**58.19**
	Rank 2	92.80	79.50	69.66
	Rank 5	96.72	82.79	73.77
Sway alignment	**Rank 1**	**93.44**	**47.54**	**77.04**
	Rank 2	95.9	59.83	86.06
	Rank 5	96.72	68.85	87.70

Table 2. Results: Static feature selection($R_{lim} = 25$)

Method	Identification rate	Set Ap	Set B	Set C
Upper body alignment	**Rank 1**	**85.25**	**77.05**	**54.92**
	Rank 2	90.16	81.15	67.21
	Rank 5	94.26	89.34	78.69
Sway alignment	**Rank 1**	**90.16**	**50.81**	**76.23**
	Rank 2	94.26	59.83	82.79
	Rank 5	95.9	70.49	86.88

Table 3. Comparison with existing GEI based approaches

	Existing Method			Proposed Method
	CAS	UCR	QMUL(unsup)	
Set Ap	97.6	99.4	99.4	**93.44**
Set B	32.7	60.2	79.9	**74.58**
Set C	52.0	22.0	31.3	**77.04**

with these techniques. In case of normal walk(set Ap) our results are comparable to these approaches. When we analyze the difficult cases of carrying bag(set B) and clothing(set C) variant our approach shows significant improvement. This comparison supports our contention that features from lower leg part are sufficient for recognition and are less sensitive to variations from normal walk sequences. We have also shown the comparison with QMUL:Feature selection based approach of Bashir et al.[8]. Table 3 shows that while the performance of our approach is comparable to the performance of their unsupervised approach for the normal walk and carrying bags, there is a significant improvement for the variants of clothing. Our unsupervised approach also takes less time in comparison.

6 Conclusion

In this paper we have shown that the dynamic area of GEI near the foot can be effectively used to improve the recognition rate. This area is comparatively less sensitive to clothing and other changes. We have proposed construction of the GEI using our proposed "sway alignment" instead of the conventional "upper body alignment", and shown that selection of the area near the foot considerably improves the results. We have shown considerable improvement in recognition rates for variants over other existing GEI based approaches for a large gait dataset.

References

1. Liu, Z., Sarkar, S.: Improved gait recognition by gait dynamics normalization. IEEE Trans. Pattern Anal. Mach. Intell. 28(6), 863–876 (2006)
2. Ioannidis, D., Tzovaras, D., Damousis, I.G., Argyropoulos, S., Moustakas, K.: Gait recognition using compact feature extraction transforms and depth information. IEEE Transactions on Information Forensics and Security 2(3-2), 623–630 (2007)
3. Ekinci, M.: Gait recognition using multiple projections. In: FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, pp. 517–522. IEEE Computer Society, Los Alamitos (2006)
4. Nandini, C., Kumar, C.R.: Comprehensive framework to gait recognition. Int. J. Biometrics 1(1), 129–137 (2008)
5. Sundaresan, A., Chowdhury, A.R., Chellappa, R.: A hidden markov model based framework for recognition of humans. In: Proc. ICIP, pp. 93–96 (2003)
6. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: ICPR 2006: Proceedings of the 18th International Conference on Pattern Recognition, Washington, DC, USA, pp. 441–444. IEEE Computer Society, Los Alamitos (2006)
7. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. 28(2), 316–322 (2006)
8. Bashir, K., Xiang, T., Gong, S.: Feature selection on gait energy image for human identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, XXX (April 2008)
9. Center for Biometrics and Security Research, CASIA, http://www.cbsr.ia.ac.cn/
10. http://goldennumber.net/body.htm

Context Switching Algorithm for Selective Multibiometric Fusion

Mayank Vatsa[1], Richa Singh[1], and Afzel Noore[2]

[1] IIIT Delhi, India
mayank@iiitd.ac.in, rsingh@iiitd.ac.in
[2] West Virginia University, USA
afzel.noore@mail.wvu.edu

Abstract. This paper presents a multimodal biometric fusion algorithm that supports biometric image quality and case-based context switching approach for selecting appropriate constituent unimodal traits and fusion algorithms. Depending on the quality of input samples, the proposed algorithm intelligently selects appropriate fusion algorithm for optimal performance. Experiments and correlation analysis on a multimodal database of 320 subjects show that the context switching algorithm improves the verification performance both in terms of accuracy and time.

1 Introduction

A biometric system operates by acquiring biometric data from an individual, extracting a feature set from the acquired data, and comparing this feature set against known templates stored in the database. Further, the use of multiple biometric samples or evidences to verify the identity of an individual is often referred to as multibiometrics. Multimodal systems have several advantages over unimodal biometric systems such as resiliency to noise and malfunction, universality, and improved accuracy. Researchers have also shown that fusion of multiple biometric evidences, in general, enhances the recognition performance. Over the years, several fusion algorithms have been proposed and a comprehensive review of existing algorithms are presented in [1]. Currently, many law enforcement agencies use face, fingerprint and iris to authenticate the identity of an individual. Existing commercial systems such as HIIDE, capture these modalities and process them individually. However, the system does not fuse individual match scores or decisions to further obtain reliable and improved performance. Very limited research exists in making a paradigm shift toward an approach that can dynamically perform selective fusion and intelligently use image quality metric in the fusion framework.

This paper focuses on the need to enhance the capability to recognize individuals operating in uncontrolled environment that is common in many real world situations. In this research we develop an adaptive biometric fusion algorithm to efficiently match individuals using multiple modalities even when the biometric samples are non-optimal. Specifically, we design a context switching tool that

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 452–457, 2009.

can dynamically select the most appropriate constituent unimodal classifier or the most appropriate fusion algorithm for the given set of probe images using image quality scores. The concept of context switching can be stated as follows:

"When dealing with gallery-probe pairs of good quality, any efficient unimodal classifier can verify the identity without the need for fusion. When the quality of image falls in the range of good to average, biometric classifiers yield some conflicting results. For such cases, simple fusion rules such as sum rule with min-max normalization [1] can successfully fuse the match scores and yield correct results with very less time complexity. When dealing with non-optimal gallery probe samples due to poor image quality or availability of partial images, complex fusion rules using support vector machines are required to perform fusion. The proposed context switching tool reconciles constituent biometric classifiers (e.g. face, fingerprint, and iris recognition algorithms) with a set of fusion algorithms that contains both simple and complex schemes to optimize both verification accuracy and computational time."

On the multimodal biometric database of 320 subjects, the proposed context switching fusion algorithm improves the verification accuracy and reduces the computational cost of the system compare to existing fusion algorithms.

2 Proposed Context Switching Algorithm

Fig. 1 illustrates the steps involved in the proposed dynamic context switching algorithm. For a biometric system with two classes (genuine, impostor) and three modalities, the algorithm uses image quality scores and three classifiers (e.g. decision tree or support vector machine (SVM)) for context switching. Classifier-1 is used to choose between the unimodal and multimodal approach based on the input evidences. If the quality of probe image is above a non-linear threshold, then unimodal approach is selected otherwise multimodal approach is selected. Next, if the unimodal approach is selected, then Classifier-2 is used to select one of the three unimodal options: (1) only face, (2) only fingerprint, and (3) only iris. If Classifier-1 selects the multimodal approach, then Classifier-3 is used to select the optimal fusion rule for a given probe case. Classifier-3 selects a complex fusion algorithm only when there is uncertainty or imperfection in the image quality scores otherwise it selects a simple fusion algorithm for combining information obtained from multimodal biometric images.

2.1 Design of Algorithm

In the context switching algorithm, three SVMs are used as the three classifiers to reconcile unimodal algorithms and fusion algorithms. Input to the first SVM, denoted as SVM_1, is used to select unimodal algorithms (face, fingerprint, and iris) or fusion rules. If unimodal algorithms are selected then the second SVM, denoted as SVM_2, is used to choose among face, fingerprint, and iris. If the option pertaining to fusion rules is selected then the third SVM, denoted as SVM_3, is used to select the optimal fusion algorithm among a collection of

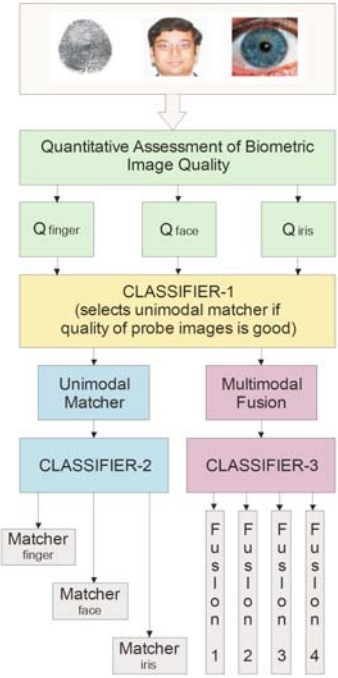


Fig. 1. Illustrating the concept of the proposed context switching algorithm

fusion rules. The context switching algorithm is divided into two stages: training SVMs for reconciliation and dynamic selection for every query instance.

Training SVMs for Reconciliation: Three SVMs are independently trained using the labeled training database. The training procedure is explained below.

1. SVM_1 is two-class classifier that is trained using the labeled training data $\{x_{1i}, y_{1i}\}$. Here, x_{1i} is the quality vector belonging to the i^{th} training gallery-probe pair. $y_{1i} \in (+1, -1)$ is the respective label such that $+1$ is assigned when the gallery-probe pair is of high quality and can be correctly matched using unimodal algorithm and -1 is assigned to the data that requires fusion.
2. SVM_2 is a multiclass classifier that is trained using the labeled training data $\{x_{2i}, y_{2i}\}$ where, x_{2i} is the quality vector belonging to the i^{th} training gallery-probe pair and y_{2i} is the multiclass soft label. In the soft labeling, range of [-1, - 0.66] belongs to gallery-probe pair that can be matched using face recognition, range of [- 0.66, 0.33] belongs to gallery-probe pair that can be fused using fingerprint recognition, and [0.33, 1] is assigned to the data that requires matching with iris recognition.
3. SVM_3 is also a multiclass classifier that is trained using the labeled training data $\{x_{3i}, y_{3i}\}$. Here, x_{3i} is the i^{th} training data vector that contains quality scores, match scores and verification accuracy priors pertaining to the three

unimodal recognition algorithms, and y_{3i} is the soft label such that SVM_3 classifies the constituent collection of fusion rules.

Dynamic Context Switching at Probe level: For probe verification, the trained SVMs are used to dynamically select the most appropriate algorithm depending on the quality scores.

1. The quality scores pertaining to both the gallery-probe images are provided as input to the trained SVMs. SVM_1 classifier selects between unimodal algorithms and fusion rules.
2. To improve the performance, the classification result of SVM_1, SVM_2 or SVM_3 are used to select one of the four options: (1) only face, (2) only fingerprint, (3) only iris, and (4) optimal fusion rule of the given probe images. The selected algorithm is then used for final decision-making.

2.2 Details of Implementation

In implementing the algorithm, we use existing algorithms for computing quality scores for face, fingerprint and iris images, and for extracting biometric features. Specifically, for computing face image quality score we use quality assessment algorithm describe in [2], RDWT based algorithm [3] is used for computing fingerprint quality score and Dempster Shafer theory based algorithm [4] for iris image quality assessment. Further, neural network architecture based Gabor transformation [5] is used to extract facial features, state-of-the-art commercial fingerprint and iris feature extraction and matching tools are used for fingerprint and iris recognition [6]. Furthermore, we use sum rule with min-max normalization [1] and likelihood ration based SVM fusion (referred as LR-SVM) [7] as two constituent fusion rules. Note that, in this paper, we use two score level fusion rules, however it can include other levels of fusion such as image or feature fusion. Finally, to train SVMs, we use radial basis kernel with kernel parameter of 4 (it our experiments we observe that it yields the best accuracy) and to compute soft labels, we use standard density estimation approach [8].

3 Experimental Evaluation and Discussion

Evaluation is performed on West Virginia University (WVU) multimodal database. This database contains face, fingerprint, and iris images from 320 subjects. The database is divided into two partitions: train and test. Training database is composed of face, fingerprint, and iris images pertaining to 128 subjects (40% of total population) and rest of the images pertaining to 192 subjects are used as test (gallery-probe) data. We also perform 20 times cross validation and receiver operating characteristics (ROC) curves are generated across these trials. Verification accuracies are reported at 0.01% false accept rate (FAR). Fig. 2 shows the ROC curves along with verification performance (accuracy and time) of the proposed algorithm and comparison with constituent unimodal and match score fusion algorithms. The key analysis of the results are delineated as:

Algorithm	Verification Accuracy At 0.01% FAR	Time (sec.)
Face	84.9	2.2
Fingerprint	76.1	1.0
Iris	92.0	0.4
Sum Rule + Min-Max Normalization	94.4	4.1
LR-SVM Fusion	97.1	5.4
Proposed	97.2	3.3

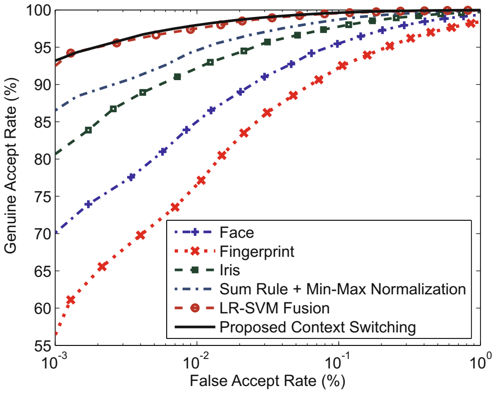


Fig. 2. Performance of the proposed context switching algorithm and comparison with unimodal and existing match score fusion algorithms

– We evaluate the correlation among biometric sources by estimating the Pearsons correlation coefficient and Spearmans rank correlation measure to model dependencies in match scores. We observe that fusing negatively correlated information sources results in pronounced improvement in matching performance which is in accordance to results of other researchers [9].
– The results show that among the unimodal traits, iris is among the best and fingerprint provides an accuracy of 76.1%. Lower performance of fingerprint is due to large variation in image quality and sensor noise.
– Both the fusion rules improve the verification accuracy by 2.4-5.1%. More importantly, correlation analysis between two fusion rules suggests that there are cases when sum rule yields better accuracy compare to LR-SVM fusion and vice-versa.
– The correlation analysis of unimodal and multimodal scores also supports the underlying concept of context switching. Experimental results show that the proposed context switching algorithm not only improves the verification accuracy, though slightly on WVU-multimodal database, but also decreases the average verification time. On 2.4 Duo Core processor with 2GB RAM under Matlab environment, the proposed algorithm require 3.3 seconds whereas existing fusion rules require 4.1-5.4 seconds.
– By design, another important aspect of the proposed context switching algorithm is that it can be easily modified to include other biometric modalities and fusion rules.

4 Conclusion

The paradigm of information fusion entails in processing evidence presented by multiple sources to enhance the recognition performance of biometric systems. Although extensive research has been done, there is a need to enhance the recognition capability when operating in uncontrolled environment. This paper presents a context switching algorithm that can fill the gap in the current state-of-art. The proposed algorithm analyzes the input biometric samples obtained from diverse, disparate sensors and characterizes the samples based on the quality and amount of information present. It further performs context switching by adaptively assessing if a unimodal biometric classifier can reliably identify an individual with high accuracy or it is required to choose the most appropriate multimodal biometric fusion algorithm based on the degree of uncertainty, incompleteness, and distortion present in the biometric samples. The proposed algorithm optimizes the underlying algorithmic and computational challenges in the decision making process such that the performance with respect to both accuracy and response time guarantees the success in real world operational scenario.

References

1. Ross, A., Nandakumar, K., Jain, A.: Handbook of multibiometrics. Springer, Heidelberg (2006)
2. Vatsa, M., Singh, R., Noore, A.: SVM based adaptive biometric image enhancement using quality assessment. In: Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, vol. 83, pp. 351–371. Springer, Heidelberg (2008)
3. Vatsa, M., Singh, R., Noore, A., Houck, M.: Quality-augmented fusion of level-2 and level-3 fingerprint information using DSm theory. International Journal of Approximate Reasoning 50(1), 51–61 (2009)
4. Kalka, N.D., Zuo, J., Dorairaj, V., Schmid, N.A., Cukic, B.: Image quality assessment for iris biometric. In: Proceedings of SPIE Conference on Biometric Technology for Human Identification III, vol. 6202, pp. 61020D-1–62020D-11 (2006)
5. Singh, R., Vatsa, M., Noore, A.: Face recognition with disguise and single gallery images. Image and Vision Computing 27(3), 245–257 (2009)
6. http://www.neurotechnology.com
7. Vatsa, M., Singh, R., Ross, A., Noore, A.: Likelihood ratio in a SVM framework: fusing linear and non-linear classifiers. In: Proceedings of IEEE Computer Society Workshop on Biometrics at Computer Vision and Pattern Recognition Conference, pp. 1–6 (2008)
8. Tao, Q., Wu, G., Wang, F., Wang, J.: Posterior probability support vector machines for unbalanced data. IEEE Transaction on Neural Network 16(6), 1561–1573 (2005)
9. Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., Duin, R.P.W.: Is independence good for combining classifiers? In: Proceedings of International Conference on Pattern Recognition, vol. 2, pp. 168–171 (2000)

Biometric Based Unique Key Generation for Authentic Audio Watermarking

Malay Kishore Dutta[1], Phalguni Gupta[2], and Vinay K. Pathak[3]

[1] Department of Electronics and Communication Engineering, Galgotias College of Engineering and Technology, Greater NOIDA, India
[2] Department of CSE, IIT-Kanpur, India
[3] Department of CSE, HBTI – Kanpur, India
malay_kishore@rediffmail.com, pg@cse.iitk.ac.in,
vinaypathak.hbti@gmail.com

Abstract. This paper proposes a method of generating pseudorandom number sequences based on biometric templates of iris image. These sequences are found to be unique in nature. Such sequences can be stored in a database for distinct identification of the extracted keys and they can act as secret keys for audio watermarking with a stamp of ownership unlike arbitrary pseudorandom number sequences and chaotic sequences. Correlation scores achieved under signal processing attacks is more than 0.9 that is significant for identification.

Keywords: Audio watermarking, Biometric based keys, Perceptual transparency, Digital rights management.

1 Introduction

In Digital audio watermarking, embedding of watermark in audio signals is to be made in such a way that it does not degrade the audibility of the signal. Applications of watermarking involve copyright protection to resolve piracy disputes, proof of ownership, broadcast monitoring and secret communication. Several schemes of audio watermarking are proposed in different domains. [1], [2]. A popular method of audio watermarking employs spread spectrum techniques. In a number of the developed algorithms [3], [4] the watermark embedding and extraction is carried out using spread-spectrum technique. Pseudorandom sequences are used in spread spectrum audio watermarking techniques to spread the watermark data across the entire audible spectrum. These pseudorandom sequences are generally generated using methods like random number generator and chaotic maps. Sometime a logo or a symbol is used as a seed to generate the watermark. However if there is a piracy dispute on the ownership the symbol or the logo may not be considered as an adequate proof of ownership. In addition to that a malicious attacker may embed a watermark of a rival counterpart in an audio signal in pirated media files to mislead. As a general perspective a normal random number sequence or a pseudorandom sequence cannot be claimed for ownership until that sequence can be uniquely mapped to an entity that is logically or physically owned by the claimant. These limitations of existing watermarking systems have been a cause of concern and there is a need for more secure and unique authentication methods.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 458–463, 2009.

One way to overcome the above-mentioned limitations is to explore the possibility of mapping a digital watermark to an entity that can be physically or logically owned. Keeping these issues in mind this paper proposes to incorporate biometric data as the seed of the watermark. Biometric features are used for the generation of the watermark key (bio-key*)* for efficient use for identification and authentication. If biometric features are used as a key in a watermarking system then the authentication and ownership issues can automatically be addressed, as the biometric features are unique for any individual and can be mapped in a database.

In order to establish the proposed method of generating bio-key that can be incorporated in the audio signal, iris images are considered. From the iris images iris features are extracted and stored in a database. The method of iris feature extraction has been discussed in Section 2. The correlation between these feature vectors are discussed in section 3. The method of bio-key generation has been discussed in Section 4. The method proposed in [8] has been used to embed and recover the bio key from the audio signal is discussed in Section 5 along with robustness test results. Next section discusses the identification of the extracted bio-keys. Finally concluding remarks are given in the last section.

2 Iris Feature Extraction

Haar wavelet technique is used to extract features from the iris image. The inner iris boundary is localized on the iris image using circular Hough transformation [5], [6]. Once the inner iris boundary (which is also the boundary of the pupil) is obtained, outer iris is determined using intensity variation approach [7]. The annular portion of iris after localization is transformed into rectangular block to take into consideration the possibility of pupil dilation. This transformed block is used for feature extraction using Discrete Haar Wavelet Transform (DHWT). Haar wavelet operates on data by calculating the sums and differences of adjacent values. It operates first on adjacent horizontal values and then on adjacent vertical values. The decomposition is applied up to four levels on transformed rectangular iris block as shown in Fig. 1. A *d*-dimensional real feature vector A_1 is obtained from the fourth level decomposition and is given as

$$A_1 = [\ i_1, i_2, ..i_d] \qquad (1)$$

Plot of a feature vector and its power spectral density (PSD) is shown in Fig 1 and Fig. 2 respectively. The PSD of the feature vector reveals that the power of the signal is approximately equally distributed in the entire frequency spectrum.

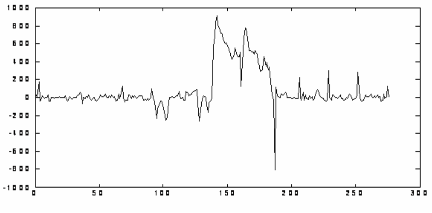


Fig. 1. Plot of an Iris Feature Vector

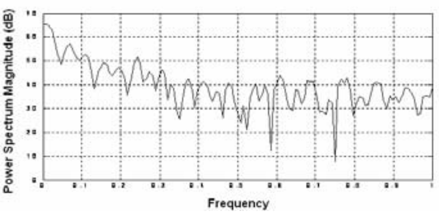


Fig. 2. PSD of the Sample Iris Feature Vector

It is clearly seen from the PSD curve that the power is approximately distributed over the entire frequency range. This property is attractive for spread spectrum techniques [3], [4] where the watermark is needed to be spread across the entire spectrum.

3 Correlation of Feature Vectors

Fig 3 shows the normalized correlation (NC) of the 100th sample feature vector with all other feature vectors in a database of 150 samples. The high spike indicates the autocorrelation of the feature vector. Subsequent to the highest spike in the figure the next highest spike is 0.79 that is the best correlation with some other feature vector in the database. The lowest correlation is found to be 0.61.

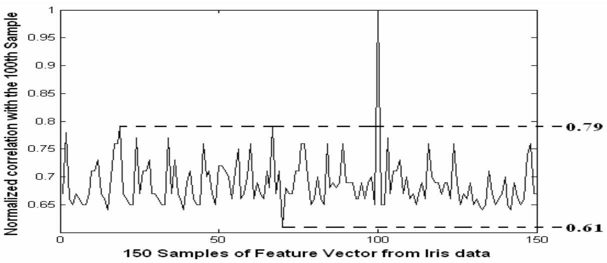


Fig. 3. NC of the 100th Feature Vector with All Others

4 Bio-key Generation From Iris Data

This section presents a novel approach to generate a bio-key from the feature vector of the iris data. The gray scale iris image is resized with 450 × 350 pixel resolution. Using the method described in Section 2 a feature vector *A* is generated from the iris image. The feature vector is then normalized taking the absolute value of the elements. The median element of the vector *A* is used to define a vector *B* such that the element *B(i)* is +1 if *A(i)* is larger or equal to the median element; otherwise it is set to –1. This bio-key that is generated form the feature vector becomes unique. The mean of these bio-keys are approximately equal to zero. Fig. 4 shows the power spectral density (PSD) of a bio-key generated by the method described above. It is clearly evident from the PSD of the bio-key that the power is evenly distributed throughout the spectrum.

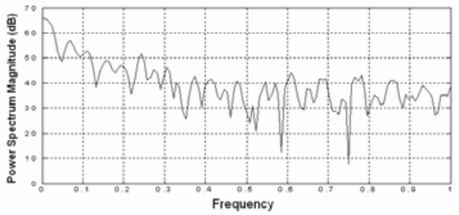


Fig. 4. PSD of a sample Bio-key

The NC between the 100th bio-key with all other samples is shown in Fig 5. The highest spike in the figure is the autocorrelation of the sample that is equal to one. It can be seen from the figure that the maximum correlation coefficient of the bio-key of the 100th sample with rest of the bio-keys is 0.35 while the minimum correlation is 0.1. This correlation is much lesser than the correlation of the same feature vector with other feature vectors. This reduction in the correlation allows having optimal values of threshold for unique detection of the watermark. Fig 6 shows the correlation of the 50th Feature vector with all the other feature vectors of the database (plot A , dotted line) and the correlation of the bio-key generated from the 50th feature vector with all other bio-keys in the database (plot B, solid line). It can be seen that the correlation of the bio-key is comparatively much lesser than the corresponding correlation of the feature vector. These bio-keys with less correlation allow deciding a judicious value of threshold for detection of watermark.

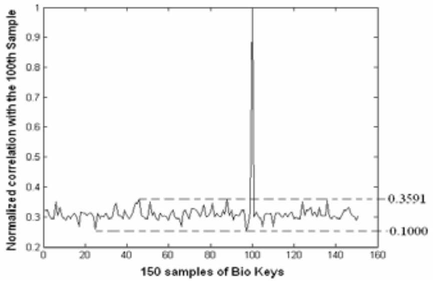


Fig. 5. The NC of the 100th Bio-key with all others

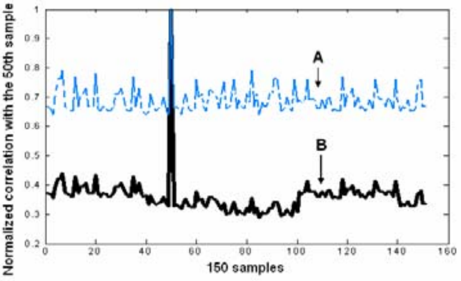


Fig. 6. NC of the 50th Bio-key and 50th Feature Vector

5 Watermark Embedding and Detection

Following the method proposed in [8] an attempt has been made to embed and to recover the bio-keys from the given audio signal. The method selects embedding regions on the original audio waveform based on a threshold in the time domain. The length of the bio-key used in the experiments is 276. A database of 150 bio-keys used for experiments to embed and then recover them from audio signals under signal processing attacks. The robustness test for survival of the bio-keys is shown in Table 1.

6 Identification

Bio keys were picked from a database of 150 samples for embedding in audio signals. These audio signals were then subjected to signal processing attacks and then the bio-key was detected and recovered from the audio signal. For the mapping of the extracted bio- key normalized correlation (NC) is used. The NC of the extracted bio-key is determined with all the keys available in the database. If there is a large difference between the highest NC coefficient and the next highest NC coefficient then we can conclude that the bio key is uniquely mapped in the database.

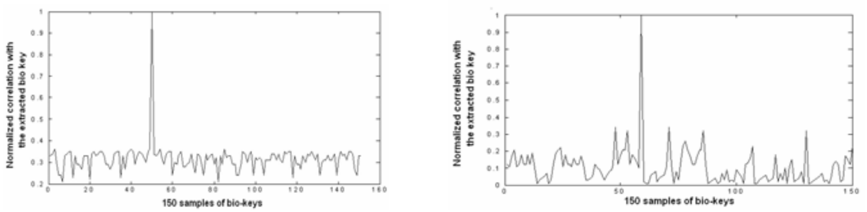


Fig. 7. NC of the extracted bio-key with a database of 150 for two samples (under attack free condition)

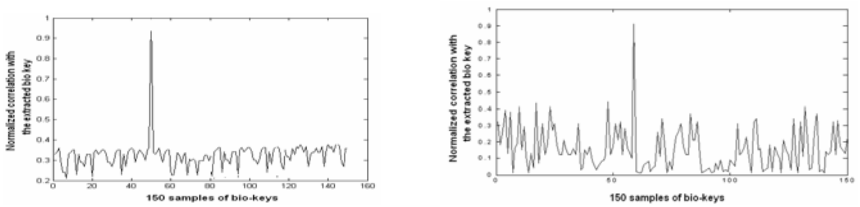


Fig. 8. NC of the extracted bio-key with a database of 150 for two samples. (under TSM attack of + 10%).

Table 1. Robustness Tests against Signal Processing Attacks

Audio File	Type of attack	NC	BER %	Type of attack	NC	BER %
Sample 1 **Tabla** (Indian musical instrument) 18 Sec, 44.1 KHz Sampling, 16 bits/ sample	Attack free Low pass (4 KHz) Low pass (8 KHz) Re-sampling(22 KHz) Re-sampling (11. 5KHz) Add Gaussian Noise	1 0.988 0.991 0.962 0.956 1	0 0 0 0 8 0	TSM (+5%) TSM (+10%) TSM (-5%) TSM(-10%) MP3 (32(kbps) Cropping	0.922 0.911 0.927 0.914 0.997 1	13 16 21 17 1 0
Sample2 **Multiple musical Instruments** 18 Sec, 44.1 KHz Sampling, 16 bits/ sample	Attack free Low pass (4 KHz) Low pass (8 KHz) Re-sampling(22 KHz) Re-sampling (11. 5KHz) Add Gaussian Noise	1 0.981 0.992 0.967 0.942 1	0 0 0 0 11 0	TSM (+5%) TSM (+10%) TSM (-5%) TSM(-10%) MP3 (32(kbps) Cropping	0.922 0.901 0.921 0.904 0.990 1	14 21 16 21 2 0
Sample 3 Classical1 18 Sec, 44.1 KHz Sampling, 16 bits/ sample	Attack free Low pass (4 KHz) Low pass (8 KHz) Re-sampling(22 KHz) Re-sampling (11. 5KHz) Add Gaussian Noise	1 0.978 0.989 0.987 0.949 1	0 0 0 2 10 0	TSM (+5%) TSM (+10%) TSM (-5%) TSM(-10%) MP3 (32(kbps) Cropping	0.922 0.90 0.923 0.901 0.98 1	13 24 13 24 4 0
Sample 4 Piano 18 Sec, 44.1 KHz Sampling, 16 bits/ sample	Attack free Low pass (4 KHz) Low pass (8 KHz) Re-sampling(22 KHz) Re-sampling (11. 5KHz) Add Gaussian Noise	1 0.976 0.932 0.983 0.961 1	0 0 0 2 9 0	TSM (+5%) TSM (+10%) TSM (-5%) TSM (-10%) MP3 (32(kbps) Cropping	0.932 0.913 0.931 0.914 0.991 1	16 19 16 19 1 0
Sample 5 Country 18 Sec, 44.1 KHz Sampling, 16 bits/ sample	Attack free Low pass (4 KHz) Low pass (8 KHz) Re-sampling(22 KHz) Re-sampling (11. 5KHz) Add Gaussian Noise	1 0.943 0.912 0.978 0.954 1	0 0 0 6 0 0	TSM (+5%) TSM (+10%) TSM (-5%) TSM (-10%) MP3 (32(kbps) Cropping	0.95 0.901 0.962 0.909 0.990 1	6 21 8 20 2 0

Results of two such experiments are presented in Fig 7. The high spike with NC of 1 suggests that the key is mapped to one of the sample. The next highest correlation is below 0.35. Hence the mapping of the extracted bio-key is very distinct for identification purpose. In Table 1 it is mentioned that the most serious signal processing attack is Time scale modification (TSM). To test the identification of the bio-keys under this attack the watermarked signal was subjected to a TSM attack of 10%. Then the NC of the extracted bio-key was done with all samples in the database as shown in Fig 8. It can be seen that in the TSM attack of 10% the extracted watermark has a correlation of 0.9 with a sample in the database and the next best correlation is less than 0.5 which is good enough for identification.

7 Conclusion

This paper has proposed a method to generate the watermark (bio-key) from biometric data for audio signals. The proposed method addresses an important limitation in ownership of digital watermarks for identification and authentication. The bio-keys generated have a correlation less than 0.3 among themselves and could survive signal-processing attacks. The extracted bio-keys from watermarked audio signals were subjected to identification from the database from which it was picked. Under no-attack condition the correlation score of 1 was achieved which clearly maps the extracted key to the database. Under serious synchronization attacks like time scale modification (10%) the bio-key a correlation score of 0.9 was achieved which is good enough for identification purpose. The results indicate distinct identification of an extracted bio-key from a database. If the owner is not an individual but a legal entity then such keys has to be generated by a combination of biometric data from multiple subjects. This is left for future work in this area.

References

1. Fridrich, J., Goljan, M., Du, R.: Distortion-Free Data Embedding. LNCS, vol. 2173, pp. 27–41. Springer, Heidelberg (2001)
2. Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for Data Hiding. IBM Systems Journal 35(3-4), 313–336 (1996)
3. Kirovski, D., Malvar, H.S.: Spread-spectrum Watermarking of Audio Signals. IEEE Transactions on Signal Processing 51(4), 1020–1033 (2003)
4. Bassia, P., Pitas, I., Nikolaidis, N.: Robust Audio Watermarking in the Time Domain. IEEE Transactions on Multimedia 3(2), 35–41 (2001)
5. Chen, T.C., Chung, K.L.: An Efficient Randomized Algorithm for Detecting Circles. Computer Vision and Image Understanding 83(2), 172–191 (2001)
6. He, X., Shi, P.: A Novel Iris Segmentation Method for Hand-held Capture Device. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 479–485. Springer, Heidelberg (2005)
7. Ma, L., Tan, T., Zhang, D., Wang, Y.: Local Intensity Variation Analysis for Iris Recognition. Pattern Recognition 37(6), 1287–1298 (2004)
8. Li, W., Xue, X., Lu, P.: Localized Audio Watermarking Technique Robust against Time Scale Modification. IEEE Transactions on Multimedia 8(1), 61–69 (2006)

Annular Iris Recognition Using SURF

Hunny Mehrotra[1], Banshidhar Majhi[1], and Phalguni Gupta[2]

[1] National Institute of Technology Rourkela, Rourkela 769008
[2] Indian Institute of Technology Kanpur, Kanpur 208016
{hunny,bmajhi}@nitrkl.ac.in, pg@iitk.ac.in

Abstract. This paper proposes an iris recognition system which can handle efficiently the problem of rotation, scaling, change in gaze of individual and partial occlusions that are inherent to non-restrictive iris imaging system. In addition to this, traditional iris normalisation approach deforms texture features linearly due to change in camera to eye distance or non-uniform illumination. To overcome the effect of aliasing features are extracted directly from annular region of iris using Speeded Up Robust Features (SURF). These features are invariant to transformations and occlusion. The system is tested on BATH, CASIA and IITK databases and is showing an accuracy of more than 97%. From the results it is inferred that local features from annular iris gives much better accuracy for poor quality images in comparison to normalised iris.

Keywords: Annular Iris Region, SURF, Local Features, Occlusion, Transformation.

1 Introduction

Iris is gaining added attention since last few decades due to accuracy, reliability and speed. Iris image acquisition is a highly restrictive process that requires cooperative and well trained audience. Acquired image is localised using pupil and iris boundary. Further, segmented iris region is normalised to form a rectangular image for matching. However there are several issues to be taken into consideration prior to feature extraction and matching. Some of these issues are worth to mention. During image acquisition, there may be some tilt in head or change in gaze of an individual. Thus the features are transformed circularly in Cartesian plane. Again iris image may be occluded by lower and upper eyelids that makes images inappropriate for matching. Another issue is that texture pattern in iris undergoes linear deformation due to expansion and contraction of pupil under non-uniform illumination. Further mapping from Cartesian to polar plane creates the effect of aliasing that loses significant texture details that are relevant from recognition point of view.

There exists several global feature extraction techniques in iris. In [1] Gaussian filter at multiple scales is used to extract features. Iris coding method based on differences of Discrete Cosine Transform (DCT) coefficients of overlapped angular patches from normalised iris images is presented in [2]. In [3] wavelet transform is applied on circular bands of iris and zero crossing representation is used

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 464–469, 2009.

for coding. These approaches extract global features and works on normalised iris image. Global feature extraction techniques generally fail under change in illumination, rotation and occlusion. To make a robust iris recognition system, local features can be extracted around the keypoints detected from annular iris image. These features are based on object representation using interest points and possess invariance to rotation, scaling, illumination and performs with certain degree of occlusions as well. There has been considerable amount of work done for object localisation and biometrics recognition using keypoint descriptors [4][5]. These local descriptors have been applied on iris recognition as well. A novel corner point descriptor is developed for iris that uses Harris corner detector to extract keypoints and entropy information of window around the corner as descriptor [6]. However, Harris corners are not scale invariant and fails to achieve property of repeatability when iris deforms due to illumination. Further, region based Scale Invariant Feature Transform (SIFT) has been applied on annular portion of iris. The system is designed for non-cooperative iris database [7].

This paper proposes an iris recognition system which makes an attempt to handle the above mentioned issues. To avoid loss of information due to mapping, segmented iris image (annular region) is considered directly for feature extraction without normalisation as given in Section 2. Further local features are extracted from segmented annular region to overcome the effect of occlusion, rotation and illumination. In this paper, Speeded Up Robust Features (SURF) [8] are used for feature extraction as given in Section 3. SURF uses same matching approach as SIFT but with few variations. Firstly, SURF uses sign of Laplacian to have sharp distinction between background and foreground features. Secondly, SURF uses only 64 dimensions compared to SIFT using 128 dimensional vector. This reduces feature computation time and allows quick matching with increased robustness simultaneously. After feature extraction, two iris images are paired using matching strategy given in Section 4. The results of the proposed system has been analysed in Section 5. Conclusion is given in the last section.

2 Iris Preprocessing

The acquired iris image contains some irrelevant information that has to be removed prior to feature extraction. Preprocessing involves detection of inner and outer iris boundary using Circular Hough Transform (CHT) [9] as shown in Fig. 1(a). Light spots on pupil region adds noise and impediments the localization process. These spots are detected and filled using morphological region filling algorithm [10]. Due to variation in illumination and change in camera to eye distance the annular region lying between pupil and iris boundary is highly variable. In order to overcome these variations traditional iris recognition approaches transforms annular region into polar coordinate. However, Hugo et. al in [11] have studied the problem of aliasing that occurs during polar transformation. The authors have studied the relationship between size of captured iris image and recognition accuracy. It has been inferred that due to change in area the recognition accuracy reduces considerably. In the proposed paper the problem of aliasing is removed by considering the annular region of iris without normalisation. After localisation of inner and outer iris boundary iris ring is segmented from the rest of the image as shown in Fig. 1(b).

3 Feature Extraction Using SURF

Global feature extraction techniques fail to work directly on annular iris images. The performance of feature extractor lies in its ability find the same image despite of change in scale, rotation and illumination. Local features are extracted by finding the keypoints in an image and forming descriptor vector around each detected keypoint. These features renders small changes in the descriptor with significant change in relative position of keypoints. Hence descriptors are paired even with affine distortions and occlusions of keypoints. SURF is a descriptor that extracts distinctive and stable features with relatively less computational requirements [8]. Steps involved to extract features using SURF are

3.1 Keypoint Detector

For fast detection of keypoints Gaussian second order derivative is used with box filters, in contrast to SIFT that approximates Laplacian of Gaussian (LOG) with Difference of Gaussian (DOG). For each pixel p in an image, Hessian matrix for p at scale σ is obtained. The determinant of Hessian matrix is used for selecting location and scale. The local maxima found of the approximated Hessian matrix determinant are interpolated in scale and image space. The detected keypoints on iris is shown in Fig. 1(c).

3.2 SURF Descriptor

A circular window is constructed around every detected keypoint and orientation is estimated using Haar Wavelet responses to have invariance to rotation. Further, SURF descriptors are obtained by taking a rectangular window around every detected keypoint in the direction of orientation. The windows are split into 4×4 sub regions. For each sub region Haar wavelet responses are extracted for equally spaced sample points [12]. Finally the wavelet response in horizontal (d_x) and vertical (d_y) direction are summed up for each sub region. The absolute values of wavelet responses ($|d_x|$ and $|d_y|$) are summed up to find the polarity of image intensity changes. Hence feature vector for each sub region is given by

$$\mathrm{fv} = \left(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|\right) \tag{1}$$

Thus summing up the descriptor vectors from all 4×4 sub-regions results in feature descriptor of length 64. The descriptor vector of length 64 for each interest point forms the feature vector.

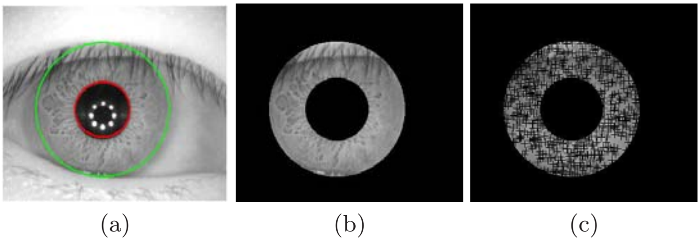


Fig. 1. Steps involved in proposed iris recognition system (a) Iris localisation using Hough transform (b) Annular segmentation of iris (c) Detection of keypoints on annular region

4 Iris Pairing

After detection of keypoints in gallery image (A) and probe image (B), matching is done using interest point pairing approach. The best candidate match for each keypoint in A is found by identifying the closest pair from the set of keypoints in B. The nearest neighbor is defined as the keypoint with minimum Euclidean distance for the invariant descriptor vector. Let $L = \{l_1, l_2, l_3.....l_m\}$ and $E = \{e_1, e_2, e_3.....e_n\}$ be vector arrays of keypoints of A and B respectively. The descriptor array l_i of keypoint i in L and descriptor array e_j of keypoint j in E are paired if the Euclidean distance $||l_i - e_j||$ between them is less than a specified threshold τ. Threshold based pairing results in several number of matching points. To avoid multiple matches, the keypoints with minimum descriptor distance and less than threshold are paired. This results in a single matching pair, and is called as nearest neighbourhood matching method. The paired points (l_i, e_j) are removed from L and E respectively. The matching process is continued until there are no more keypoints in L. The number of paired points between sample images A and B is shown in Fig. 2.

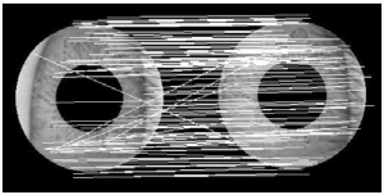


Fig. 2. Pairing between gallary and probe iris images

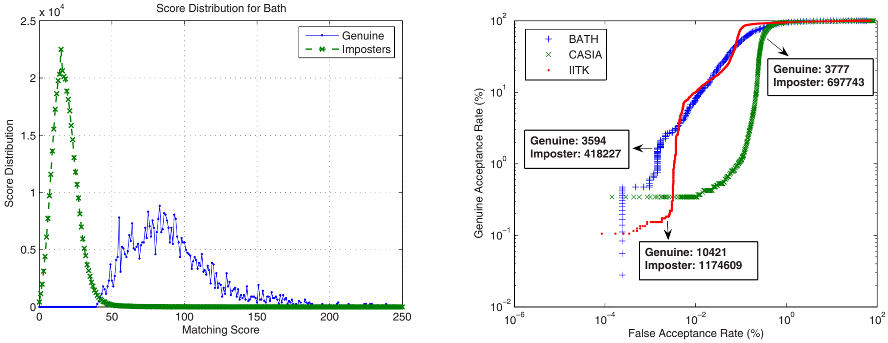
5 Experimental Results

The proposed system has used BATH [13], CASIA [14] and Indian Institute of Technology Kanpur (IITK) [15] databases to determine its performance. Database available from BATH University comprises of 20 images from both

Table 1. Results generated using SURF for normalized and annular iris image (values in %)

Databases →	**BATH**			**CASIAV3**			**IITK**		
Testcases ↓	FAR	FRR	Accuracy	FAR	FRR	Accuracy	FAR	FRR	Accuracy
Normalized Iris	10.35	21.11	84.26	3.31	5.13	95.77	0.86	5.52	96.80
Annular Iris	2.37	1.97	97.84	1.44	4.07	97.23	4.65	1.41	97.15

the eyes for 50 subjects (50 × 20). CASIA-IrisV3 database comprises of 249 subjects with total of 2655 images from both the eyes. The database collected at IITK consists of over 1800 iris images taken from 600 subjects (600 × 3) only from left eye. To measure the performance of the proposed system, genuine and imposter scores are generated using three databases. Distribution of genuine and imposter scores for BATH is given in Fig. 3(a). Similar curves are obtained for CASIA and IITK databases. To measure the robustness of the system, ROC curves are obtained for all the three databases as shown in Fig. 3(b). Table 1 provides comparative performance of SURF using two different testcases i.e., (a) normalized and (b) annular iris images. From the results it is evident that the system performs well in both the testcases for IITK database where the acquired image is of sufficiently large size and covers most part of the eye. Thus the problem of aliasing does not arise and hence traditional normalisation technique gives equally good results. However for BATH database the images are distant from the camera with less iris details. Traditional normalisation in case of BATH creates aliasing effect and shows poor performance with very low accuracy of 84.26%. However, using annular iris image the accuracy increases significantly to 97.84%. Similarly, the accuracy of CASIA database increases by 2%.



(a) (b)

Fig. 3. Performance results of annular iris recognition: (a) Distribution of genuine and imposter scores for BATH database (b) ROC curves for BATH, CASIA and IITK databases

6 Conclusion

This paper has proposed an iris recognition system that works under unconstrained imagining conditions. In this approach the problem of aliasing that occurs using traditional normalisation approach has been addressed by extracting features directly from annular iris image. Feature extraction using iris circle is a challenging task. Thus, SURF is used to extract local features having invariance to basic transformations. The system has been tested against standard available databases with and without normalisation. It has been found that the system performs better for all kinds of iris imagery with robustness to rotation and linear deformations that are inherent to iris.

References

1. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(11), 1148–1161 (1993)
2. Monro, D.M., Rakshit, S., Zhang, D.: Dct-based iris recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(4), 586–595 (2007)
3. Boles, W.W., Boashash, B.: A human identification technique using images of the iris and wavelet transform. IEEE Transactions on Signal Processing 46(4), 1185–1188 (1998)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
5. Bicego, M., Lagorio, A., Grosso, E., Tistarelli, M.: On the use of sift features for face authentication. In: Conference on Computer Vision and Pattern Recognition Workshop, June 2006, pp. 35–35 (2006)
6. Mehrotra, H., Badrinath, G.S., Majhi, B., Gupta, P.: An efficient dual stage approach for iris feature extraction using interest point pairing. In: IEEE Workshop on Computational Intelligence in Biometrics: Theory, Algorithms, and Applications, April 2009, pp. 59–62 (2009)
7. Belcher, C., Du, Y.: Region-based sift approach to iris recognition. Optics and Lasers in Engineering 47(1), 139–147 (2009)
8. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision and Image Understanding (CVIU) 110(3), 346–359 (2008)
9. Kerbyson, D.J., Atherton, T.J.: Circle detection using hough transform filters. In: Fifth International Conference on Image Processing and its Applications, July 1995, pp. 370–374 (1995)
10. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice-Hall, Englewood Cliffs (2007)
11. Proenca, H., Alexandre, L.A.: Iris recognition: An analysis of the aliasing problem in the iris normalization stage. In: International Conference on Computational Intelligence and Security, vol. 2, pp. 1771–1774 (2006)
12. Bay, H., Fasel, B., Gool, L.V.: Interactive museum guide: Fast and robust recognition of museum objects (May 2006)
13. Bath University Database, http://www.bath.ac.uk/elec-eng/research/sipg/irisweb
14. Casia Database, http://www.cbsr.ia.ac.cn/english/Databases.asp
15. Database of Indian Institute of Technology Kanpur, http://www.cse.iitk.ac.in/users/biometrics

Face Recognition Using Posterior Distance Model Based Radial Basis Function Neural Networks

S. Thakur[1], J.K. Sing[2,*], D.K. Basu[2], and M. Nasipuri[2]

[1] Department of Information Technology, Netaji Subhas Engineering College, Kolkata, India
[2] Department of Computer Science & Engineering, Jadavpur University, Kolkata, India
jksing@ieee.org

Abstract. The success rate of a face recognition system heavily depends on two issues, mainly, i) feature extraction method and ii) choosing/designing of a classifier to classify a new face image based on the extracted features. In this paper, we have addressed both the above issues by proposing a new feature extraction technique and a posterior distance model based radial basis function neural networks (RBFNN). First, the dimension of the face images is reduced by a new direct kernel principal component analysis (DKPCA) method. Then, the resulting face vectors are further reduced by the Fisher's discriminant analysis (FDA) technique to acquire lower dimensional discriminant features. During classification, when the RBFNN is not so confident to classify a test image, we have introduced a statistical method called the posterior distance model (PDM) to resolve the conflict. The PDM is an approach, which takes a decision by integrating the outputs of the RBFNN and a distance measure. We call the new classifier the posterior distance model based radial basis function neural networks (PDM-RBFNN). The proposed method has been evaluated on the AT&T database. The simulation results in terms of recognition rates are found to better than some of the existing related approaches.

Keywords: Face recognition, Radial basis function neural networks, Direct kernel principal component analysis, Fisher's discriminant analysis.

1 Introduction

Since the last decade, human face recognition is an active research area in the field of pattern recognition and computer vision due to its wide range of applications, such as identity authentication, access control, surveillance systems, security, etc. As a result, a number of methods have been proposed in the past and surveys in this area can be found in [1]-[4]. To achieve higher performance in face recognition, two issues are needed to be addressed: i) feature extraction process from a face image and ii) choosing or designing of a classifier to classify a new face image based on the extracted features. Till now, most of the efforts have been made to address the first issue by proposing various feature extraction methods [3]-[7]. Very few efforts have been

[*] Corresponding author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 470–475, 2009.

made to address both the above issues [8]-[10]. The present study addresses both the above issues to improve the performance of a face recognition system.

In this paper, we have addressed the feature extraction process and designing of a classifier to achieve higher success rate. A new kernel PCA, referred to as the direct kernel PCA (DKPCA) [11], is used for dimension reduction of the image data. The DKPCA explicitly maps an input image nonlinearly into a feature space spanned by the number of training samples and then computes the principal components directly in the mapped space. This method considers the structural information of the input images in the feature space for computation of principal components, leading to have higher discriminating power. It has been found that the DKPCA has higher discriminating power than the KPCA for face recognition [11]. After reducing the dimension of image data, the reduced feature vectors are further reduced using the Fisher's discriminant analysis (FDA) technique to acquire lower-dimension discriminant features. Since the DKPCA algorithm reduces the dimension of the image data considerably, the so-called 'small-sample-size' (SSS) problem does not arises while performing the FDA in the mapped space. We call the above feature extraction process using the DKPCA and FDA algorithms as the Direct Kernel Fisher Discriminant Analysis (DKFDA).

We have also designed a new classifier based on RBF neural networks (RBFNN) and a posterior distance model for face recognition. When the RBFNN is not so confident to recognize a test image, we have introduced a new statistical method called the posterior distance model (PDM) to resolve the conflict. The PDM is an approach, which takes a decision by integrating the top three outputs of the RBFNN and the distance measures between the test image and centres of the classes associated with the top three outputs of the RBFNN. The PDM improves the success rate for face recognition by reducing the numbers of false positives and false negatives. We call the new classifier the posterior distance model based radial basis function neural networks (PDM-RBFNN).

2 Feature Extraction

The face image features are extracted using a new direct kernel Fisher's discriminant analysis (DKFDA) algorithm. The DKFDA algorithm is implemented in two stages. In the first stage, high dimensional image data are reduced using direct kernel PCA method (DKPCA) [11]. The DKPCA method explicitly maps an input image nonlinearly into a feature space spanned by the rank of the training samples and then computes the principal components directly in the mapped space. Since the DKPCA considers the nonlinear structures of the input samples in the mapped feature space, the computed principal components have higher discriminating power. In the second stage, reduced image data are further reduced using Fisher's discriminant analysis (FDA) method. The FDA searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). It finds an optimal subspace for classification that maximizes the ratio of the between-class scatter matrix and the within-class scatter matrix. We call the above technique, which uses the DKPCA and FDA, the direct kernel Fisher's discriminant analysis (DKFDA) for feature extraction.

3 Posterior Distance Model (PDM)

The schematic diagram of the PDM-RBFNN is shown in Fig. 1. Let the output vector of the RBFNN for a test pattern $\mathbf{Z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})$ is $\mathbf{C}_i = (c_{i1}, c_{i2}, \ldots, c_{iC})$, where C is the total number of individuals in the database. Let top three outputs are o_{i1}, o_{i2} and o_{i3}, respectively. Let the classes associated with the top three outputs are l_1, l_2 and l_3, respectively. We define a confidence factor (*CF*) for classifying a test pattern as the standard deviation (*SD*) of the top three outputs of the RBFNN. If the *CF* is higher than a predefined threshold ε, the class of the test pattern is determined as the index of the output neuron, which produces maximum value. Otherwise, the PDM model determines its class by integrating the top three outputs of the RBFNN and distance measures as follows:

$$\boldsymbol{class}(\boldsymbol{Z}_i) = \arg\min_{l_1,l_2,l_3} \{\boldsymbol{d}(\boldsymbol{Z}_i, \boldsymbol{t}_{i_k}) * (1 - \boldsymbol{o}_{ik})\}, \boldsymbol{k} = 1,2,3 \qquad (1)$$

where $d(\mathbf{Z}_i,\mathbf{t}_{ik})$ is a distance measure between the test pattern $\mathbf{Z}_i$ and the mean training pattern $\mathbf{t}_{ik}$ of the class associated with the top k^{th} output of the RBFNN. In our experiments, we have used the Euclidean and Manhattan distance as the distance measure. The Manhattan distance is found to be better than the Euclidean distance. The above model, called the posterior distance model (PDM), is used to resolve the conflict arises when the RBFNN fails to recognize a test pattern confidently. The PDM improves the success rate by reducing the numbers of false positives and false negatives.

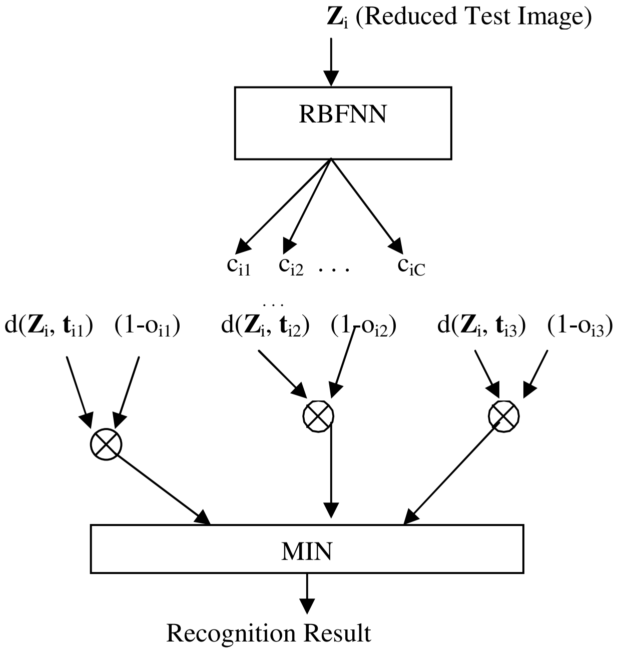


Fig. 1. Schematic diagram of the PDM-RBFNN

4 Experimental Results

The performance of the proposed method has been evaluated on the AT&T Laboratories Cambridge database (formerly ORL database) [12]. Several experiments have been carried out using three different methodologies; namely, i) randomly partitioning the database, ii) N-fold cross validation test and iii) leave-one-out strategy.

4.1 Randomly Partitioning the Database

In these experiments, we have selected randomly s images per person from the database to form the training set. Remaining images form the corresponding test set. It should be noted that there is no overlap between the training and test images. In this way 10 different training and test sets have been generated for each value of s.

We have conducted several experiments by varying DKFDA features. Table 1 shows the best performances of the proposed method for s = 4, 5, and 6. The best average recognition rates of 92.29%, 95.05% and 95.38% are achieved for s = 4, 5 and 6, respectively using Manhattan distance as the distance measure.

Table 1. Average recognition rates of the proposed methods by randomly partitioning the database

# samples/person, s	4	5	6
# training samples	160	200	240
# test samples	240	200	160
# hidden layer neurons	120	120	160
# DKFDA features	50	55	55
Average Recog. (%)	**92.29**	**95.05**	**95.38**
Standard Deviation	1.84	1.59	1.07

4.2 N-Fold Cross Validation Test

In this study, the AT&T database is randomly divided into ten-folds, taking one image of a person into a fold. Therefore, each fold consists of 40 images, each one corresponding to a different person. For ten-fold cross validation test, in each experimental run, nine folds are used for training and remaining one fold is used for testing. Therefore, training and test sets consist of 360 and 40 images, respectively in a particular experimental run. Several experiments are conducted by varying DKFDA features from 10-50. The best average recognition rate using the Manhattan distance as the distance measure is achieved using 20, 30, 40 and 50 DKFDA features. Table 2 shows the performances of the proposed method and the RBFNN without PDM model. The average recognition rate is found to be 97.75%. It may be noted that performance of the proposed method is better than the RBFNN without the PDM model.

Table 2. Recognition rates using N-fold cross validation test on the ORL database

Method	**Proposed method**	**RBFNN without PDM**
# training samples	360	360
# hidden layer neurons	200	200
Average Recog. (%)	**97.75**	**97.50**
Standard Deviation	2.19	2.36

4.3 Leave-One-Out Strategy

In this study, the experiments are carried out using the "*leave-one-out*" strategy. To classify a test image of a subject, the image is removed from the database of *N* images and placed into a test set. Remaining *N*-1 images are used in the corresponding training set. In this way, experiments are performed *N* times, removing one image from the database at a time. For the AT&T database, we have performed 400 experimental runs for the database of 400 images. Table 3 shows the average recognition rate of the proposed system using 20 DKDFA features and 200 hidden layer neurons. We have achieved 97.50% average recognition rate.

Table 3. Experimental results using leave-one-out strategy

# of hidden layer neurons	# of DKFDA features	Average recog. (%)
200	20	97.50

5 Conclusion

Success rate of a face recognition system is heavily depends on the feature extraction process and the choosing or designing of a suitable classifier. In this paper, we have described a direct kernel Fisher's discriminant analysis (DKFDA) technique and proposed a posterior distance model-based radial basis function neural networks (PDM-RBFNN), to address the above two issues, respectively. The new system has been evaluated on the AT&T face database. Experiments are carried out in three different strategies; namely, i) randomly partitioning the database, ii) N-fold cross validation test and iii) leave-one-out strategy. The experimental results have demonstrated improved performance for the new system. The best average recognition rates of 92.29%, 95.05% and 95.38% have been achieved for $s = 4$, 5 and 6, respectively using Manhattan distance as the distance measure. In N-fold cross validation test we have achieved 97.75% average recognition rate. Whereas in leave-one-out strategy, we have achieved 97.50% average recognition rate.

Acknowledgments. This work was partially supported by the CMATER and the SRUVM projects of the Department of Computer Science & Engineering, Jadavpur University, Kolkata, India. The author, S. Thakur would like to thank Netaji Subhas

Engineering College, Kolkata for providing computing facilities and allowing time for conducting research works. The author, D. K. Basu would also like to thank the AICTE, New Delhi for providing him the Emeritus Fellowship (F.No.: 1-51/RID/EF(13)/2007-08, dated 28-02-2008).

References

1. Samal, A., Iyengar, P.: Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition 25, 65–77 (1992)
2. Chellapa, R., Wilson, C., Sirohey, S.: Human and machine recognition of faces: A survey. J. IEEE 83, 705–741 (1995)
3. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Surveys 35, 399–458 (2003)
4. Tolba, A.S., El-Baz, A.H., El-Harby, A.A.: Face Recognition: A Literature Review. International Journal of Signal Processing 2, 88–103 (2006)
5. Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N.: Face recognition using kernel direct discriminant analysis algorithms. IEEE Trans. Neural Networks 14, 117–126 (2003)
6. Yang, J., Frangi, A.F., Yang, J.-Y.: A new kernel Fisher discriminant algorithm with application to face recognition. Neurocomputing 56, 415–421 (2004)
7. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces versus Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. 23, 711–720 (1997)
8. Er, M.J., Wu, S., Lu, J., Toh, H.L.: Face recognition with radial basis function (RBF) neural networks. IEEE Trans. Neural Networks 13, 697–710 (2002)
9. Sing, J.K., Basu, D.K., Nasipuri, M., Kundu, M.: Face recognition using point symmetry distance-based RBF network. Applied Soft Computing 7, 58–70 (2007)
10. Sing, J.K., Basu, D.K., Nasipuri, M., Kundu, M.: High-speed face recognition using self-adaptive radial basis function neural networks. To appear in Neural Computing & Application, Springer, Heidelberg
11. Sing, J.K., Basu, D.K., Nasipuri, M., Kundu, M.: Direct kernel PCA with RBF neural networks for face recognition. In: IEEE TENCON 2008, Hyderabad, India (2008)
12. ORL face database. AT&T Laboratories, Cambridge, UK, `http://www.uk.research.att.com/facedatabase`.html

Segmentation for Iris Localisation: A Novel Approach Suitable for Fake Iris Detection

Bodade Rajesh M.[1] and Talbar Sanjay N.[2]

[1] Military College of Telecommunication Engineering, Mhow-453441, India
rajeshbodade@gmail.com
[2] S.G.G.S. Institute of Engineering and Technology, Vishnupuri, Nanded, India
sntalbar@yahoo.com

Abstract. In iris recognition system, accurate iris segmentation and localisation from eye image is the foremost important step. In this paper a robust and efficient method of iris segmentation is proposed. In the proposed method, the outer boundary of iris is calculated by tracing objects of various shape and structure. Based on the pupil size variation, the inner boundary of iris is detected. The variation in pupil size is also used for aliveness detection of iris. Thus, this approach is a very promising technique in making iris recognition systems more robust against fake-iris-based spoofing attempts. The algorithm is tested on UPOL database of 384 images both eyes of 64 subjects. The success rate of accurate iris localisation from eye image is 99.48% with minimal loss of iris texture features in spatial domain as compared to all existing techniques. The processing time required is also comparable with existing techniques.

Keywords: Iris Segmentation, Fake Iris Detection, Pupil Dynamics, Dynamic Iris Localisation.

1 Introduction

Identification of people is getting more and more importance in the increasing network society. Biometrics is the branch of science in which human beings are identified with their behavioral or physical characteristics. Among all biometrics, iris recognition is the gaining more attention because iris of every person is unique and it never changes during a person's lifetime [1-8].

The acquired image of eye does not contain only iris but it also contains pupil and data derived from the surrounding eye region like sclera, eyelid and eyelashes. Therefore, it is extremely important to segment and localize the iris from the acquired eye image, prior to feature extraction. Thus, the overall performance of iris recognition system is decided by the fact that how accurate iris is segmented and localized from an eye image[1].

This paper explains the novel approach of dynamic iris localisation based on comparison of two images of different intensities for complete and accurate localisation of iris without any loss of iris features.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 476–482, 2009.

The reminder of this paper presents Related work, Proposed system, Outer boundary detection, Inner boundary detection, Experimental results and Conclusions in Sections 2, 3, 4, 5, 6 and 7 respectively.

2 Related Work

A generalised iris recognition consists of image acquisition, iris segmentation and localization (preprocessing), feature extraction and feature comparison (matching). Biometric based personal identification using iris requires accurate iris localization from an eye image [2]. Several researchers have implemented various methods for segmentation and localising the iris. John Daugman [3] has proposed one of the most practical and robust methodologies, constituting the basis of many functioning systems. He used integro-differential operator to find both the iris inner and outer boundaries for iris segmentation. Wildes [2] proposed a gradient-based binary edge map construction followed by circular Hough transform for iris segmentation. Several researchers have proposed several variants of these methods with minor variations in the research paperes [5-8].

Almost all methods stated are based on the assumption that centre of iris (Outer Boundary) and Pupil (Inner boundary) is same and iris is perfectly circular in shape which are seldom true. Therefore, the iris segmentation and localization from an acquired image leads to loss of texture data near to pupil and/or outer boundary.

3 Proposed System of Iris Localisation

Iris possess high degree of randomness with high variation in eye colour contrast and texture, to obtain maximum information and efficiency for recognition, high contrast images are required. Fig.1 shows sample images with low, medium and high contrast from CASIA, UBIRIS and UPOL iris databases respectively[9-11].

Fig. 2 shows the flow chart of proposed method of iris localisation using two iris images at different light levels (Intensities). It mainly consist of three steps, viz., Outer (Iris) Boundary Detection, Inner (Pupil) Boundary Detection and Normalisation.

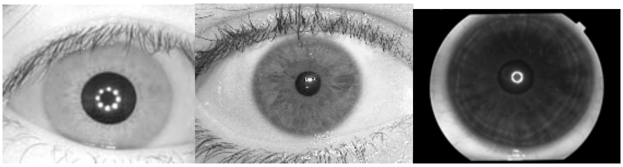


Fig. 1. Low, Medium and High contrast eye images

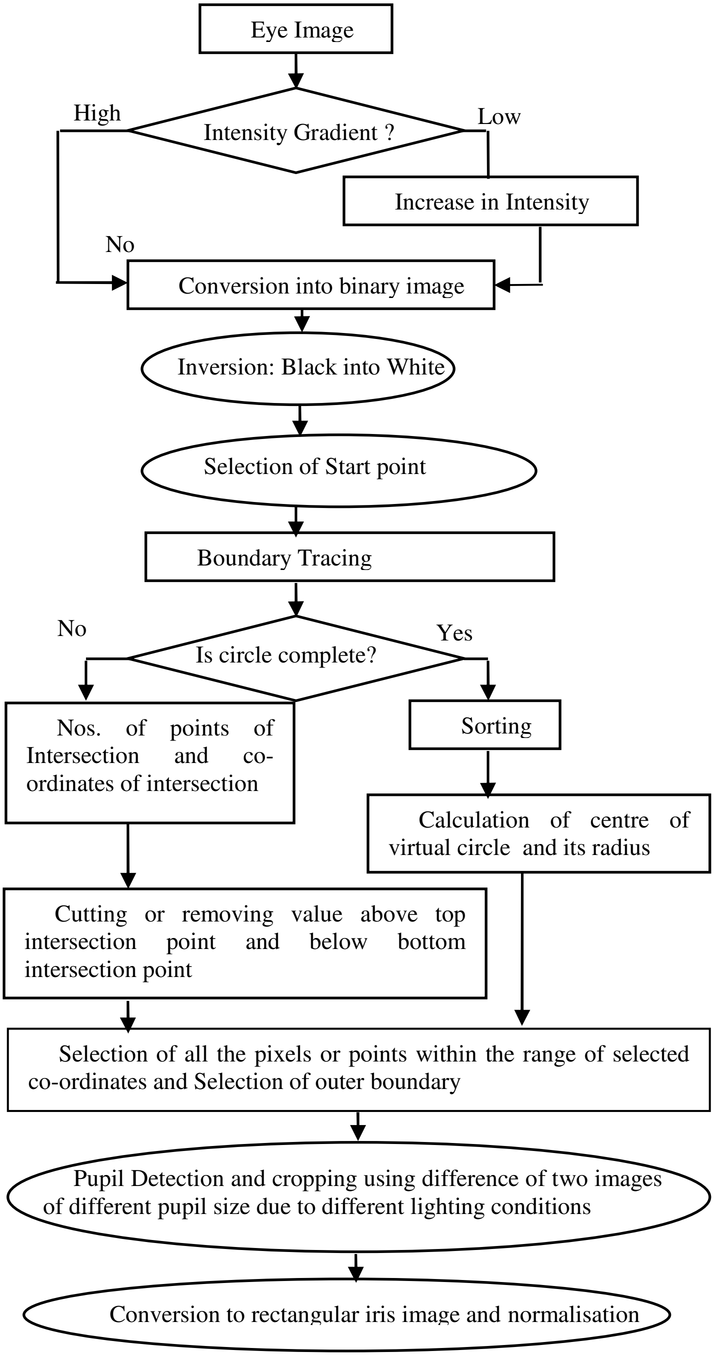

Eye Image
High
Intensity Gradient ?
Low
Increase in Intensity
No
Conversion into binary image
Inversion: Black into White
Selection of Start point
Boundary Tracing
No
Is circle complete?
Yes
Nos. of points of Intersection and co-ordinates of intersection
Sorting
Calculation of centre of virtual circle and its radius
Cutting or removing value above top intersection point and below bottom intersection point
Selection of all the pixels or points within the range of selected co-ordinates and Selection of outer boundary
Pupil Detection and cropping using difference of two images of different pupil size due to different lighting conditions
Conversion to rectangular iris image and normalisation

Fig. 2. Flow chart of a proposed system

4 Outer Boundary Detection

Since the high contrast images provides prominent texture information of iris as well as other components of eye, these other components creates problem in iris localization while calculating the threshold value, which is used for converting image into binary images. So to overcome this problem it is required to trace irregular objects with of different shapes and structure. Empirical study of various shows that iris is located generally on central area of image. Tracing can be started from any one corner of image.

Boundary is being traced of all points with binary value as '1' in all direction starting from selected point that is the first point that has value as '0' coming from top to bottom in any one quarter of image. Thus complete boundary is traced for a complete iris without any intersection. For images with intersection with upper or lower eyelids, as shown in Fig 3(b) may not result into a complete one object, for such cases point of intersection is calculated and all points above point of intersection in case of intersection with upper eyelid and points below in case of intersection with lower eyelids are removed. The traced boundary of iris is shown in Fig 3(a) with green colour and a virtual circle is drawn using all these traced points with blue colour. In case of complete iris or complete circle is being traced then area under the green colour boundary is trace else in case of intersection objects within virtual circle with blue colour boundary is selected. This selected area is segmented from rest of the image.

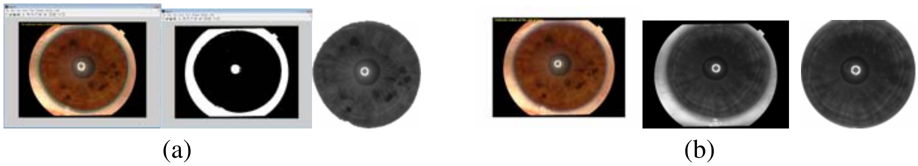


Fig. 3. Outer Boundary detection in case of (a) complete circle and (b) incomplete circle

5 Inner Boundary (Pupil) Detection

Selection of inner boundary of iris is one of the most complicated part of iris localisation. Thus in this report a novel approach is used, in which, two images of iris are taken at different light intensities and the variation in intensity of light will produce variation in size of pupil[3-4] thus taking images of iris at these two light intensities will give similar images but with variation in size of pupil.

Now comparing or subtracting these two images. As iris part of two images is same, result of subtraction will give 0 value and only place where non zero values are obtained is the region of pupil due to variation in size of pupil. The resultant image will also have number of visible spots as noise which are required to be remove to calculate exact size of pupil.

Size of pupil is larger than the size of obtained noise, converting this image into binary image gives complete iris of smaller pupil along with noise, on removing all the objects of smaller size than the predefined value results into exact inner boundary of pupil as shown in Fig 4.

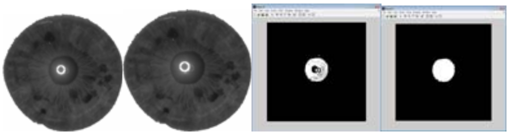


Fig. 4. Detection of Inner (Pupil) Boundary

Tracing this inner boundary and selecting region outside inner boundary and below outer boundary will give exact iris with minimum losses as shown in Fig 5(a). Finally, completely detected iris is converted to rectangular image using normalization Equation (1) and (2) as shown in Fig 5(b).

$$x_1 = x + r * \cos(\Phi) \tag{1}$$

$$y_1 = y + r * \sin(\Phi) \tag{2}$$

where, (x, y) are the coordinates of center of the ring,(x_1, y_1) are the coordinates of pixel of rectangular image, r is a radius of iris ring that varies form inner to outer boundary of iris image and Φ is an angle of that varies from 0 to 360 degree.

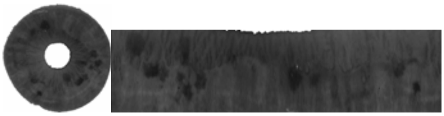


Fig. 5. (a) Segmented Iris (b) Normalised rectangular Iris

6 Results

The database contains 3 x 128 iris images (i.e. 3 x 64 left and 3 x 64 right). The images are: 24 bit - RGB, 576 x 768 pixels, file format: PNG. The irises were scanned by TOPCON TRC50IA optical device connected with SONY DXC-950P 3CCD camera[17]. The proposed algorithm is implemented in MATLAB7.0, on PIV-3Ghz, Intel processor with 512MB RAM and tested on UPOL database[11].

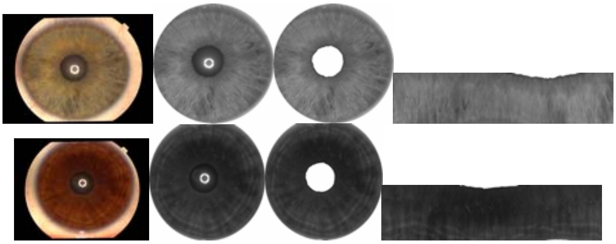


Fig. 6. Iris localisation output for a few eye images of the database

Fig 6 shows the output of various stages of algorithm for sample images of database. The segmentation accuracy and timing analysis of the algorithm and its comparison with existing algorithms is given in Table 1.

Table 1. Result of segmentation accuracy and timing analysis

Methodology	Accuracy	Time in Seconds
Proposed	99.48 %	1.43
Daugman[3]	67.23%	1.03
Wildes[2]	88.49%	1.3
Masek[7]	83.97%	7.8
Narote[8]	91.33%	1.21

The experimental results have shown that the proposed algorithm gives better results. The algorithm accurately extracted irises of 382 images out of 384 images of 64 subjects giving a success rate of 99.48 % that very minimal loss of iris texture features as compared to existing methods, specially for high contract iris images.

7 Conclusions

The proposed method showed the very high accuracy rate of iris segmentation and at comparable timing cost.

The strength of the method is that it does ont assume the centre of pupil and centre of iris as same as against other method and hence it is more practical one.

Moreover, the method is based on the comparison of two iris images at different light levels to detect the change in the sixe of pupil. Thus, this is a very promising technique for making iris recognition systems more robust against fake-iris-based spoofing attempts as well. This makes this method more useful than any other methods.

We are extending the use of this method for fake iris detection / aliveness detection of iris for full-proof iris recognition system with inherent anti-spoofing mechanism[4].

References

1. Kong, W., Zhang, D.: Accurate iris segmentation based on novel reflection and eyelash detection model. In: Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, May 2001, pp. 263–266 (2001)
2. Wildes, R.: Iris Recognition: An Emerging Biometric Technology. Proc. IEEE 85, 1348–1363 (1997)
3. Daugman, J.: How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology 14, 21–30 (2004)
4. Daugman, J.: Anti-spoofing Liveness Detection, `http://www.cl.cam.ac.uk/users/igdl000/countermeasures.pdf`
5. Boles, W., Boashash, B.: A Human Identification Technique Using Images of the Iris and Wavelet Transform. IEEE Trans. Signal Processing 46, 1185–1188 (1998)
6. Ma, L., Wang, Y., Tan, T.: Personal Identification Based on Iris Texture Analysis. IEEE Trans. on PAMI 25(12), 414–417 (2003)

7. Masek, L., Kovesi, P.: MATLAB source code for a Biometric Identification System Based on Iris Paterns. The school of Computer Science and Software Engineering, The University of Western Austrilia (2003)
8. Narote, S.P., Narote, A.S., Waghmare, L.M.: An automated Segmentation Method for Iris Recognition. In: Proceedings of TENCON 2006. 2006 IEEE Region 10th Conf., November 14-17 (2006)
9. Chinese Academy of Sciences Institute of Automation, Database of 756 Greyscale Eye Images, http://www.sinobiometrics.com
10. Proença, H., Alexandre, L.A.: UBIRIS: iris image database (2004), http://iris.di.ubi.pt
11. High contrast Iris image database downloaded from, http://phoenix.inf.upol.cz/iris/download/

Key Independent Retrieval of Chaotic Encrypted Images

Ram Ratan

Defence Research and Development Organisation
Scientific Analysis Group, Metcalfe House Complex
Delhi-110054, India
ramratan_sag@hotmail.com

Abstract. A chaotic image encryption algorithm based on circular shift functions proposed for high security is analysed in this paper for retrieving encrypted images. Proposed retrieval scheme is key independent and based on divide and conquer attack where neighbourhood similarity characteristic of images is applied. The simulation results show that retrieved images have very good visual perception quality and are as similar as original images. The analysis indicates that above algorithm in present form is insecure and encrypted images can be retrieved efficiently.

Keywords: Image Secrecy, Chaotic Image Encryption, Circular Shift Function, Image Decryption, Divide and Conquer Attack, Neighbourhood Similarity.

1 Introduction

Security is an important issue in communication and storage of digital images because of rapid use of such media in the digital world nowadays and encryption is one of the ways to achieve security. Encryption is the process which tranforms the information with the help of encryption key into encrypted form which is unintelligible and looks like a random mesh. Image encryption has wide applications in strategic communication, telemedicine, medical imaging, multimedia systems, etc. Images are different from text and it is not a wise idea to use traditional encryption schemes to encrypt them because of much encryption time of large image size. Moreover, retrieved text must be same as original text but this is not necessary for images because of visual characteristics of human perception which tolerate small errors in retrieved images,i.e., small errors in retrieved images are acceptable. Decryption is the process by which original information is retrieved from encrypted infromation with the help of decryption key.

In order to achieve security, a variety of encryption schemes have been proposed which can be classified in three types: position permutation [1]-[5], value transformation [6]-[9] and visual transformation [5]. The present paper is concerned with position permutation where circular shift functions are used [1-2,8] to encrypt images. An image encryption algorithm given in [1] is based on bit circulation and called as Bit Recirculation Image Encryption (BRIE) and an

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 483–488, 2009.

algorithm proposed in [2] is based on pixel circulation and called as Chaotic Image Encryption (CIE). Both the algorithms are also known as two dimentional circulation encryption algorithms (TDCEA) which consist of two dimentional circular shift functions.

The analysis of such methods for retrieval of information from encrypted images becomes important and necessary in some applications and is useful also in evaluation of such schemes for security. There are following attacks which can be applied in analysis of encryption techniques depending on various situations: (1) Known cipher image attack (2) Chosen cipher image attack (3) Known plain image attack and (4) Chosen plain image attack. The analysis of above methods is carried out for security [10-12] by applying known or chosen plain image and known cipher image attacks to obtain encryption key.

In this paper, a CIE algorithm based on pixel circulation using shift functions [1-2] is analyzed and an efficient retrieval scheme is proposed for retrieving chaotic encrypted images in which rows and columns shifts are applied. The retrieval scheme is independent of key and is applicable in a situation when only an encrypted image is known. The retrieval scheme is based on divide and conquer attack where neighbourhood similarity is considered as the correlation between adjacent columns and adjacent rows to correct the effect of circulation of pixels.

The paper is organized as follows. We firstly give a brief introduction of CIE in Section 2. Proposed image retrieval scheme is presented in Section 3. Simulation results and discussions on security issues of CIE are presented in Section 4. Finally, the paper is concluded in Section 5 followed by references.

2 CIE Algorithm

Idea of CIE is the pixel circulation of images which is controlled by chaotic pseudo random sequence. The algorithm has following four steps:

`Step-1` determines a chaotic system and its initial point $x(0)$, rowsize M and columnsize N of an image, iteration number $n0$, and constants α, β, and γ used to determine the rotation number.
`Step-2` generates the chaotic sequence from the chaotic system.
`Step-3` geterates the binary sequence.
`Step-4` includes special functions to rearrange image pixels.

Let f be an image of size $M \times N$ pixels, $f(x,y), 0 \leq x \leq M-1, 0 \leq y \leq N-1$, be the pixel value of pixel in f at position (x,y), the transformation image f' for given image f is obtained by following circular shift functions:

(i) $ROLR_l^{i,p}$: $f \rightarrow f'$ is defined to rotate each pixel in the i^{th} row in f, $0 \leq i \leq M-1$, in the left direction p pixels if l equals 0 or in the right direction p pixels if l equals 1.
(ii) $ROUD_l^{j,p}$: $f \rightarrow f'$ is defined to rotate each pixel in the j^{th} column in f, $0 \leq j \leq N-1$, in the up direction p pixels if l equals 0 or in the down direction p pixels if l equals 1.

(iii) $ROUR_l^{k,p}$ *:* $f \rightarrow f'$ is defined to rotate each pixel at position (x, y) in the image such that $x + y = k$, $0 \leq k \leq M + N - 2$, in the upper right direction p pixels if l equals 1 or in the lower left direction p pixels if l equals 0.
(iv) $ROUL_l^{k,p}$ *:* $f \rightarrow f'$ is defined to rotate each pixel at position (x, y) in the image such that $x - y = k$, $-(N-1) \leq k \leq M - 1$, in the upper left direction p pixels if l equals 0 or in the lower right direction p pixels if l equals 1.

BRIE consists of first two functions and CIE consists of all four functions. The solution to CIE is obtained for first two circular shift functions which are used in two dimentional circulation of pixels to get an encrypted image.

3 Image Retrieval Scheme

We see in the images that the value of pixels is normally varying smoothly in the neighbourhood regions which can help us in retrieval process. Divide and conquer attack can make the solution efficient to given complex problem by decomposing it into simple problems. We use these concepts in developing proposed image retrieval scheme. For retrieval of an encrypted image, we normally require decryption key but our proposed scheme is independent of keys and we do not require any information of such keys. As per divide and conquer attack we divide given image into columns and rows which make the retrieval of an encrypted image easy and efficient. The neighbourhood similarity is considered here as the correlation which is measured as the correlation between two adjacent rows (columns). These correlation values are used to get correct shift for each row and column and rearrange the pixels accordingly in obtaining retrieved image. The scheme is described as follows:

3.1 Removal of Columnwise Circulation Effect

It is achieved by computing correlation between all adjacent columns of given encrypted image f'. The correlation between two adjacent columns, j and $j+1$, with shift t is computed as

$$corr_t(j, j+1) = f'(i, j) \times f'((i+t) mod M, j+1), 0 \leq i \leq M - 1$$

This correlation is computed for all possible shifts, $0 \leq t \leq M - 1$ and the correct shift t for column $j + 1$ is obtained for which the value of correlation is maximum. In this manner the correct shift t_j is obtained for all the columns, $1 \leq j \leq N - 1$.

3.2 Removal of Rowwise Circulation Effect

It is achieved by computing correlation between all adjacent rows of given encrypted image f'. The correlation between two adjacent rows, i and $i+1$, with shift t is computed as

$$corr_t(i, i+1) = f'(i,j) \times f'((i+1, (j+t) modN), 0 \leq j \leq N-1$$

This correlation is computed for all possible shifts, $0 \leq t \leq N-1$ and the correct shift t for row $i+1$ is obtained for which the value of correlation is maximum. In this manner the correct shift t_i is obtained for all the rows, $1 \leq i \leq M-1$.

As per above procedures, the effect of columnwise and rowwise circulation of pixels is removed by rearranging pixels in columns and rows of f'. As the shifts for first column and first row are not taken into consideration while rearranging pixels in columns and rows, these leave shifting effects in the retrieved image.

For correcting the effects due to first column and first row, we compute correlation between adjacent columns, $corr(j, (j+1) modN), 0 \leq j \leq N-1$ and correlation between adjacent rows, $corr(i, (i+1) modM), 0 \leq i \leq M-1$. The shifts t_c for column and t_r for row are obtained as $j+1$ and $i+1$ respectively for which correlation value is minimum. All the columns and rows of f' are rearranged according to t_c and t_r. Finally, the retrieved image f" is obtained.

4 Results and Discussions

CIE algorithm and image retrieval scheme given in this paper have been implemented on MATLAB Platform using MATLAB programming. The visual perception quality of retrieved images obtained with our scheme is very good and is same as that of original images. The error in retrieved images is measured as mean square error (MSE) which is computed as

$$MSR = \frac{1}{M \times N}[f(i,j) - f"(i,j)]^2, 0 \leq i \leq M-1 \text{ and } 0 \leq j \leq N-1.$$

The error in retrieved image f" is tolerable because it does not leave objectionable distortion in retrieved images. Simulation results for image encryption and image retrieval are respectively shown in *figure 1* and *figure 2*.

In *figure 1*, (a) and (e) are original images; (b) and (f) are encrypted images obtained by applying only rowwise circulations; (c) and (g) are encrypted images obtained by applying only columnwise circulations, and (d) and (h) are final encrypted images obtained by applying both rowwise and columnwise circulations. In *figure 2*, (a) and (b) are given encrypted images which have MSE as 7717 and 13312 and (c) and (d) are retrieved images which have MSE as 80 and 630. We see in *figure 1* and *figure 2* that retrieved images have very good visual perception quality with unnoticeable distortion and are as similar as original images.

The security of CIE is mentioned as $2^{(3M+3N-2) \times n(0)}$ and claimed it as very high [2]. In exhaustive approach there are $(M \times N)!$ trials for normal pixels permutation and $(M^N \times N^M)$ trials for circular shifts (ROLR and ROUD) used in encryption. In proposed retrieval scheme we require only $(M+N)$ trials to retrieve an image encrypted with above two circular shift functions. This shows that the CIE algorithm is insecure in present form and encrypted images of CIE can be retrieved efficiently. The CIE can be made secure against above attacks by incorporating masking or substitutions [6,13] in addition to circular shifts.

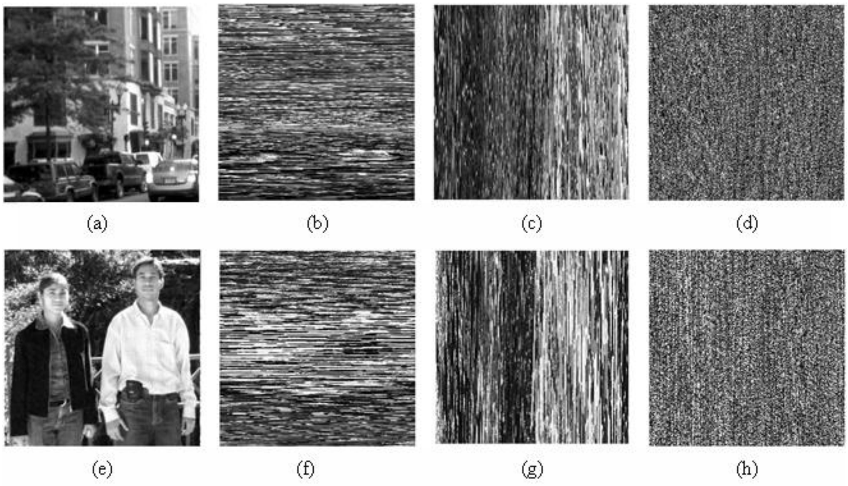


Fig. 1. Original and Encrypted Images

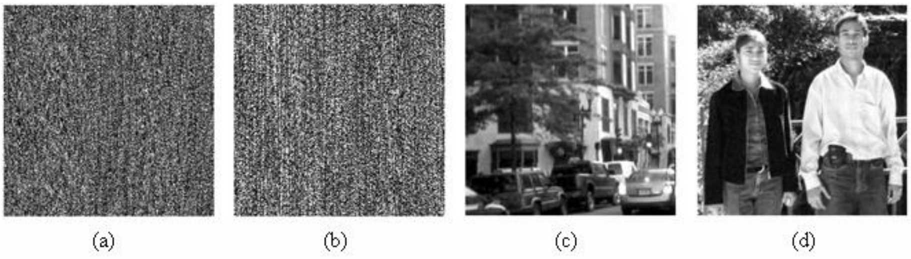


Fig. 2. Encrypted and Retrieved Images

Above retrieval scheme is developed for gray level images but can also be considered for binary images to retrieve chaotic encrypted images. For binary images, the correlation between two adjacent columns (rows) with shift t required in image retrieval is to be computed as given below:

$corr_t(j, j+1) = \frac{1}{N}[$ no. of matches $(f'(i,j) = f'((i+t)modM, (j+1)modN))-$ no. of mismatches $(f'(i,j) \neq f'((i+t)modM, j+1)), 0 \leq i \leq M-1]$.

The pixels in f' are rearranged as per above retrieval scheme to retrieve binary encrypted images.

5 Conclusion

An image retrieval scheme has been presented in this paper for retrieving encrypted images of chaotic image encryption algorithm in which circular shift functions are applied for columnwise as well as rowwise circulation of pixels.

Image retrieval is automatic, efficient and independent of keys. Image retrieval scheme is based on divide and conquer attack in which neighbourhood similarity characteristic between adjacent columns and rows is used in image retrieval. It has been shown in simulation results that the chaotic image encryption algorithm which consists of circular shift functions is insecure and the encrypted images can be retrieved with very good visual perception quality as similar to original images.

References

1. Yen, J.-C., Guo, J.-I.: A new chaotic image encryption algorithm and its VLSI architecture. In: Proc. IEEE Workshop Signal Processing Systems, pp. 430–437 (1999)
2. Yen, J.-C., Guo, J.-I.: A new chaotic image encryption algorithm. Department of Electronics Engineering National Lien-Ho College of Technology and Commerce, Miaoli, Taiwan, Republic of China, E-mail: jcyen@mail.lctc.edu.tw
3. Yen, J.-C., Guo, J.-I.: Design of a new signal security system. In: Proc. IEEE Intl. Symposium on Circuits and Systems (ISCAS 2002), vol. 4, pp. 121–124 (2002)
4. Furht, B., Kirovski, D.: Multimedia Security Handbook. CRC Press, Boca Raton (2004)
5. Guo, J.-I., Yen, J.-C.: A new mirror-like image encryption algorithm and its VLSI architecture. In: Proc. 10th VLSI Design/CAD Symposium, Taiwan, pp. 319–322 (1999)
6. Maniccam, S.S., Bourbakis, N.G.: Lossless image compression and encryption using SCAN. Pattern Recognition 34, 1229–1245 (2001)
7. Maniccam, S.S., Bourbakis, N.G.: Image and video encryption using SCAN patterns. Pattern Recognition 37, 725–737 (2004)
8. Ozturk, I., Sogukpinar, I.: Analysis and comparision of image encryption algorithms. Proc. of World Academy of Science, Engineering and Technology 3 (2005)
9. Chang, C.-C., Hwang, M.-S., Chen, T.-S.: A new encription algorithm for image cryptosystems. Journal of Systems and Software 58, 83–91 (2001)
10. Li, S., Zheng, X.: On the security of an image encryption method. In: Proc. IEEE Intl. Conf. on Image Processing (ICIP 2002), vol. 2, pp. 925–928 (2002)
11. Li, C., Li, S., Chen, G., Chen, G., Hu, L.: Cryptanalysis of new signal security system for multimedia data transmission. EURASIP Journal on Applied Signal Processing 8, 1277–1288 (2005)
12. Li, C.: Cryptanalysis of some multimedia encryption schemes. M.S. Thesis (2005)
13. Menezes, A., Van Oorschot, P., Vanstone, S.: Handbook of applied cryptography. CRC Press, Boca Raton (1996)

A Bayesian Approach to Hybrid Image Retrieval

Pradhee Tandon and C.V. Jawahar

Center for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500032, India
{pradhee@research.,jawahar@}iiit.ac.in

Abstract. Content based image retrieval (CBIR) has been well studied in the computer vision and multimedia community. Content free image retrieval (CFIR) methods, and their complementary characteristics to CBIR has not received enough attention in the literature. Performance of CBIR is constrained by the semantic gap between the feature representations and user expectations, while CFIR suffers with sparse logs and cold starts. We fuse both of them in a Bayesian framework to design a hybrid image retrieval system by overcoming their shortcomings. We validate our ideas and report experimental results, both qualitatively and quantitatively. We use our indexing scheme to efficiently represent both features and logs, thereby enabling scalability to millions of images.

1 Introduction

Retrieval of similar images and videos from large databases, has received significant attention in recent years [1]. There are two prominent approaches to solve this problem: (i) Content-based image retrieval (CBIR), popular in the computer vision community (ii) Content-free image retrieval (CFIR), which has received some amount of attention in the database community. A CBIR method typically converts an image into a feature vector representation, and matches with the images in the database to find out the most similar images. On contrary, CFIR methods exploit the co-occurrence information (for example in a collaborative filtering framework) in the logs of image-access to model the similarity across images [3,5]. If a user accesses/accepts two images together, then these images are treated as semantically related. There are also attempts which tried to combine both these approaches [6,8,10].

We are interested in designing a practical image retrieval system which (i) naturally scales to large number of images (ii) allows the simultaneous use of ideas from visual similarities as well as user behavior patterns (iii) allows overcoming the limitations (Section 3) of CBIR and CFIR by exploiting their complementary nature. We meet these objectives by reasoning in a Bayesian framework, where the *a priori* information comes from the logs, and visual similarity acts as the evidence. We conduct extensive experiments and report results to validate the superiority of the hybrid solution both qualitatively and quantitatively. We also demonstrate the scalability and efficiency of the solution. We can successfully

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 489–494, 2009.

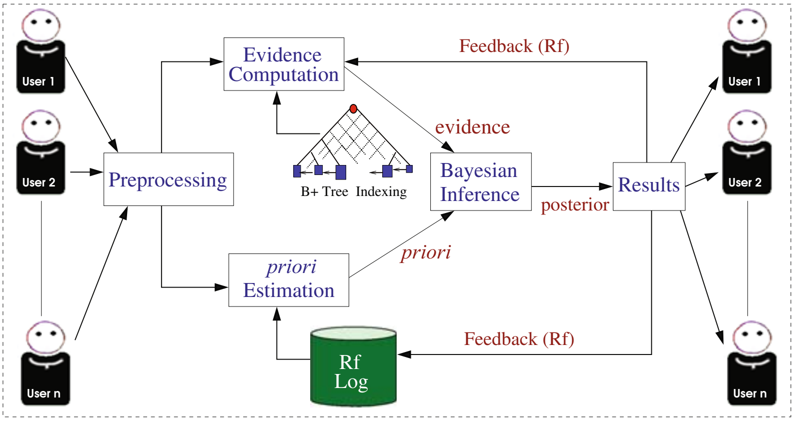


Fig. 1. Architecture of our scalable Bayesian Image Retrieval system

retrieve from millions of images in interactive (sub-second) time as demonstrated in Section 4.

2 A Scalable Indexing Scheme

Our indexing scheme is an extension of [7] and [2], which were primarily derived for CBIR in presence of a changing similarity metric. [7] demonstrated the utility of a B+-tree based indexing scheme for efficient approximate nearest neighbor (ANN) computation in high dimensional vector spaces. Later on, [2] exploited it based on the fact that most concepts get clustered in the feature spaces.

The indexing scheme used in this work (briefly sketched in Figure 1) allows simultaneous indexing of both visual clues (raw features) and user interaction patterns in the form of logs (unlike in [2,7]). Logs are represented as relationships across images in a MySQL database. Images are represented as fixed-length feature vectors in a B+ tree index structure as in [7]. A computationally efficient reasoning (Section 3), which combines these two factors with the help of a set of *learned* weights, is carried out for the interactive search. We use the logs of retrieval process to provide co-occurrence information for pairs of images. When processing the query, we use these co-occurrence relations, as explained in the next section. At the end of the retrieval, the database of logs is refined based on user acceptability of images. Note that this log contains significant amount of subjectivity (and therefore uncertainty).

Efficiency of our indexing scheme can be attributed to: (a) indexing the image feature-by-feature makes the indexing scheme consistent even when relative importance of features changes. (b) feature vectors are bulky (due to large number of dimensions) and they are represented in B+ trees. Logs are compact relationships and they are represented in MySQL. This makes our indexing scheme space efficient. (c) Since ANN (feature-based retrieval) and log based retrieval (from a standard database) are individually fast and our fusion scheme is based on very few multiplications, our scheme is overall efficient. This meets our requirement of interactive retrieval from large databases.

3 Bayesian Image Retrieval

CBIR Vs CFIR: CBIR methods use low level features for representing and retrieving images. They, typically, are unable to represent human perception of visual content. The semantic gap is the primary bottleneck of CBIR methods. Many of the previous methods extensively explored the use of feedback based learning for improved performance [4,9]. However, they are either critically dependent on features or computationally infeasible. CFIR on the other hand uses only feedback based co-occurrence among images. Therefore they are able to capture semantic relations among images and predict accurately their relevance to other seen images. User feedback and logs are difficult to obtain. Unless the system is functional, users do not provide any feedback. This creates a deadlock and cold start. In addition, CFIR has no clue about previously unseen images. Prediction accuracy is also critically dependent on the availability of logs.

Bayesian Integration: CBIR and CFIR thus provide two complimentary estimates of similarity among images. Their effective integration can overcome the critical dependencies of both of them and provide improved accuracy. We use a Bayesian framework for fusing the two approaches. Bayes theorem provides a method to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself. We formulate the image retrieval problem as one of estimating the probability of retrieving an image, as a posterior estimation problem. We model the *a priori* on the co-occurrence information from the history logs. The visual similarity, between the query and the database images, is used as the evidence in favor of the match. The two are combined in the Bayesian inference to estimate the *a posteriori* of the image being relevant to the query. If $R(\mathbf{q}, \mathbf{a})$ denotes the event of retrieving the image $\mathbf{a}$ given $\mathbf{q}$ as the query. The *a priori* probability of this event $P(R)$ can be computed from the co-occurrence as

$$P(R) = \frac{n(\mathbf{a}, \mathbf{q})}{n(\mathbf{q})} \tag{1}$$

where $\mathbf{q}$ has been found relevant by users, $n(\mathbf{q})$ times and it has been relevant with $\mathbf{a}$, $n(\mathbf{a}, \mathbf{q})$ times. It is 1 when $\mathbf{a}$ and $\mathbf{q}$ were always retrieved together. It is zero, when they never co-occur as acceptable images together. $n(\mathbf{a}, \mathbf{q})$ and $n(\mathbf{q})$ are initially assumed to be 1 to avoid inconsistencies.

Let $S(\mathbf{q}, \mathbf{a})$ be the feature-level similarity of the images $\mathbf{q}$ and $\mathbf{a}$. We learn the query concept as weights for the features $\mathbf{w}$ as discussed later. Using these weights we also learn feature weights $\mathbf{c}$, for the popular concept in individual images. $S(\mathbf{q}, \mathbf{a})$ can then be estimated as

$$p(S|R) = f(\mathbf{c}, \mathbf{w}, \mathbf{q}, \mathbf{a}) \tag{2}$$

We can now estimate the posterior using Bayes Rule as in

$$p(R|S) = p(S|R)P(R) \tag{3}$$

We do not consider the denominator of the Bayes rule, since it does not modify the relative ranking of the database images, given the query. In practice, one could use alternate 'definitions' of the probabilities, as long as they satisfy the basic axioms of probability. The top N images with the maximum *a posteriori* probability can be returned to the user.

Bayesian Image Retrieval Process: In a query-by-example framework like ours, each image in the system is represented as a vector of numeric feature values $[X_1, \ldots, X_d]^T$, constituting a multi-dimensional space where each image is a point. The database is pre-processed while the query is processed online to extract the set of features. The query is then compared for similarity with a subset of the dataset, using the Bayesian integration scheme discussed above and top N results are returned for interaction.

Given a query vector, $\mathbf{x_q}$, we retrieve the p ($p >> N$) nearest data points from each B+-tree. The trees are enumerated in order of decreasing relevance ($\mathbf{w}$) making it likely to retrieve the closest points earlier. The scheme is analyzed in detail in [2,7]. Both relevant and irrelevant images, as marked by the user (relevance feedback), are used for incrementally learning $\mathbf{w}$. $\mathbf{w}$ is learnt by iteratively estimating the relevance of a feature, s_j, based on the dispersion of the feature over relevant and irrelevant sets. At the end of the query session the relevance of features, $\mathbf{c}$, for every relevant image, is updated using $\mathbf{w}$. The expressions for estimating and updating relevances have been discussed in detail in [2]. Visual similarity S_i between x_q and image x_i is computed using a weighted Mahalanobi's metric as in

$$S_i = \left[\left(\mathbf{W}^T \left[\mathbf{x}_i - \mathbf{x_q} \right] \right)^T \mathbf{M} \left(\mathbf{W}^T \left[\mathbf{x}_i - \mathbf{x_q} \right] \right) \right]^{\frac{1}{2}}$$

where $\mathbf{W}$ is the *diagonal* matrix of $\mathbf{w}$. Co-variance matrix M is computed initially.

Co-occurrence information is summarized into $\mathbf{V}$. In the first iteration retrieval happens only on visual similarity but next one onwards the feedback pattern is used for estimating the closest concept and only these samples are used for posterior computations.

Co-occurrence information is updated either at the end of every query session or deferred by a few. We try to discover the implicit concepts in the co-occurrence information using an incremental k-means clustering on the vectors for images. This results in $\mathbf{V}$ concepts, $\mathbf{V}_i, ..., \mathbf{V}_k$ which represent the relationships existing in the database as of now. This helps us learn higher level concepts and also prunes the search space. Incremental clustering, though repetitive, allows self discovery of concepts as logs improve. Being modeled as an off-line process it does not effect retrieval. Thus together the index and the off-line summarizing allows sub-second retrieval times.

4 Experiments and Discussions

Datasets: For our accuracy experiments, we have used two datasets with different characteristics. The first, $\mathbf{D_1}$, is a completely annotated, 58 category set of $12,000$ real natural images, collected from Flickr, COREL and cartoon videos. We represent it using MPEG-7's Color Structure and Edge Histogram descriptors. The second set, $\mathbf{D_2}$, is also completely annotated and comprises of the Caltech-256 dataset. We use the state-of-art *bag of words* approach to represent these images using a $2,000$ word *SIFT* vocabulary. For the scalability study we use a dataset, $\mathbf{D_3}$ of *1 million* points generated from a uniform distribution. For collecting logs, users were asked to select a query and provide feedback to a randomly selected set of 20 images from the dataset, supposedly similar to the query. For $\mathbf{D_2}$ we used the available annotations, instead of users, for automatically creating logs. Note that logs do not provide the complete similarity distribution (users do not see all the examples in the database). Typically only 10% of the valid co-occurrences are available while testing the system.

Precision improvement: In this experiment we use human logs to show how our approach achieves better accuracy compared to a pure CBIR. We used 5 random queries from each of the 58 categories in $\mathbf{D_1}$. We computed average precision for both the approaches and compare it for a few categories in Table 1(a). A similar comparison for $\mathbf{D_2}$ using annotation based logs can be found in Table 1(b).

Learning in BSIR: Our Bayesian systems learns across users and is able to retrieve with better accuracy by using improved co-relevance information or the *a priori* and feature relevance in images, **c**.

Qualitative Comparison: Next we visually compare with CBIR using the top few results for some queries in Figure 2. As can be seen, our Bayesian retrieves with better accuracy in both the examples. The leftmost image is also the query.

Efficiency and Scalability: In Figure 2 we show how our approach retrieves in sub-second times both with increase in database size and the number of dimensions using $\mathbf{D_3}$ and 5 randomly selected queries. We have optimized on the retrieval time by designing *a priori* updates as an off-line process and storing the co-occurrence matrix in MySQL.

Table 1. Tables shows the improved precision for 3 categories with our approach over CBIR. (a) uses real user logs while (b) uses annotation based logs.

Approach	**Category**		
	1	2	3
BAYES	59.91%	75.69%	63.94%
CBIR	31.38%	39.38%	41.38%

(a)

Approach	**Category**		
	1	2	3
BAYES	72.08%	67.22%	59.35%
CBIR	42.00%	47.08%	28.84%

(b)

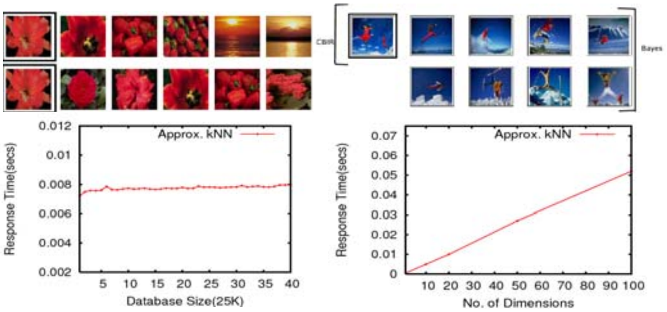


Fig. 2. (Clockwise from top-left) Top 6 results from CBIR (first row) and our Bayesian (second row); More semantically similar images got added to the CBIR result(first row) set with our Bayesian(first and second rows); Avg. retrieval time with increasing number of dimensions; Avg. retrieval time with increasing database size

5 Conclusions

We have proposed a Bayesian inference based hybrid image retrieval system which fuses complimentary techniques of CBIR and CFIR and overcomes many of their shortcomings. We have presented extensive experiments to validate the advantage in terms of accuracy, interactive retrieval times and efficient learning. We would like to further extend concept discovery in our future work.

References

1. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Survey, 1–60 (2008)
2. Tandon, P., Nigam, P., Pudi, V., Jawahar, C.V.: FISH: a practical system for fast interactive image search in huge databases. In: ACM CIVR, pp. 369–378 (2008)
3. Su, Z., Zhang, H., Li, S.Z., Ma, S.: Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning. IEEE Transactions on Image Processing, 924–937 (2003)
4. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 6, 536–544 (2003)
5. Zhou, X., Zhang, Q., Zhang, L., Liu, L., Shi, B.: An Image Retrieval Method Based on Collaborative Filtering. In: Liu, J., Cheung, Y.-m., Yin, H. (eds.) IDEAL 2003. LNCS, vol. 2690. Springer, Heidelberg (2003)
6. Melville, P., Mooney, R.J., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: NCAI, pp. 187–192 (2002)
7. Jammalamadaka, N., Pudi, V., Jawahar, C.V.: Efficient Search with Changing Similarity Measures on Large Multimedia Datasets. In: Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T. (eds.) MMM 2007. LNCS, vol. 4352, pp. 206–215. Springer, Heidelberg (2006)
8. Yu, K., Schwaighofer, A., Tresp, V., Ma, W.-Y., Zhang, H.: Collaborative Ensemble Learning: Combining Collaborative and Content-Based Information Filtering via Hierarchical Bayes. In: UAI, pp. 616–623 (2003)
9. Heisterkamp, D.R.: Building a latent semantic index of an image database from patterns of relevance feedback. In: ICPR, pp. 134–137 (2002)
10. Han, J., Ngan, K., Li, M., Zhang, H.: A Memory Learning Framework for Effective Image Retrieval. IEEE Trans. on Image Processing, 511–524 (2005)

Hierarchical System for Content Based Categorization and Orientation of Consumer Images

Gaurav Sharma[1], Abhinav Dhall[1], Santanu Chaudhury[2], and Rajen Bhatt[1]

[1] Samsung Delhi R&D, D5 Sec. 59, Noida
grvsharma@gmail.com, {abhinav.d,rajen.bhatt}@samsung.com
[2] Multimedia Lab, Dept. EE, Indian Institute of Technology Delhi
santanuc@ee.iitd.ac.in

Abstract. A hierarchical framework to perform automatic categorization and reorientation of consumer images based on their content is presented. Sometimes the consumer rotates the camera while taking the photographs but the user has to later correct the orientation manually. The present system works in such cases; it first categorizes consumer images in a rotation invariant fashion and then detects their correct orientation. It is designed to be fast, using only low level color and edge features. A recently proposed information theoretic feature selection method is used to find most discriminant subset of features and also to reduce the dimension of feature space. Learning methods are used to categorize and detect the correct orientation of consumer images. Results are presented on a collection of about 7000 consumer images, collected by an independent testing team, from the internet and personal image collections.

1 Introduction

In this paper we present a digital content management (DCM) solution which (a) automatically categorizes consumer images into four broad categories and (b) detects their correct orientation, based on their content. We first categorize the images into four categories namely Mountains, Monuments, Water bodies and Portraits. As the rotation of the input image is unknown (among multiples of 90 degrees), we do the categorization in a rotation invariant way. Then within each category we detect the correct orientation of the image by methods tuned to the statistics of that category.

We are interested in fast solutions suitable for implementation in a resource limited target. Hence, we use simple and inexpensive features based on color and edge information. To further speed up we use a recently proposed feature selection method [11], based on information theory concepts, to reduce the dimension of the feature space by mining out a small subset of most discriminant features. Support vector machine (SVM), Gaussian mixture models and variant of the boosting classifiers are used as the learning methods for the various tasks.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 495–500, 2009.

Fig. 1. Example images from the database; Mountains, Monuments, Water bodies and Portraits

1.1 Related Work

Image categorization is an area of much recent research. However, most of the work uses high time complexity point detection e.g. Scale invariant feature transform (SIFT) [7]. A recent representation of images, bag of words [14], has been shown to be very good for categorization tasks e.g. [14,6,4]. However, resource constraints prevent us from using point detection.

Orientation detection is also a well researched field. A Bayesian learning framework was presented in [10] for estimating the orientation of the images. [12,13] did content based image orientation detection using SVMs with spatial color moments (CM) and edge direction histograms (EDH) features. [15] used AdaBoost algorithm with the CM and EDH features. They trained a indoors versus outdoors classifier using similar AdaBoost algorithm. [1] proposed a scalable boosting approach for image reorientation. The features used were statistics of different sized image blocks from RGB and YUV channels and vertical and horizontal edge images with comparisons as weak classifiers.

2 Proposed Framework

2.1 Categorization Method

We categorize the images into four classes namely Mountains, Monuments, Water bodies and Portraits. Fig. 1 shows examples of the images from the database. We solve the problem with a learning based framework. First we extract color based rotation invariant features (color correlograms) to represent the image and then use a statistical learning method (support vector machines, SVM) to do the categorization.

Low level features. The color correlogram (CC) [5] feature is used for the categorization task. For any pixel, the CC gives the probability of a pixel at a distance k away to be of certain color, and is defined as $\gamma_{c_i,c_j}^{(k)} = \Pr_{p_1 \in I_{c_i}, p_2 \in I}[p_2 \in I_{c_j} ||p_1 - p_2| = k]$. The choice of feature was motivated by two factors: (a) the input images may be rotated by multiples of 90 degrees and (b) images are expected to be color images and color distribution is a good discriminant for the different classes. CC [5] feature captures the spatial correlation of colors. It has been proposed and used for image indexing and retrieval. We show that the feature along with a suitable classifier can be used for image categorization giving good results.

Support Vector Machines. Once the images are represented with CC vectors we use support vector machines (SVM) [9] for categorization of the vectors into 4 classes. SVM finds a separating hyperplane, in the $\phi(.)$ induced high dimensional space, having the maximum margin [9] and has been found useful in many machine learning applications. We train one-vs-one SVM classifier on the training data, the type of SVM used was C-SVC with an radial basis function (RBF) kernel, given by $K(x_i, x_j) = \phi(x_i)^T \phi(x_j) = \exp(-\gamma||x_i - x_j||^2), \quad \gamma > 0$. The cost parameter C and the kernel parameter γ were optimized using cross validation. We used libsvm [3] for the experiments.

2.2 Image Orientation Detection

Once the images are categorized into four categories, we proceed to detect the correct orientation of the images. Each category is expected to have (a) different distributions of the low level features for the different orientations and (b) different features which are more discriminant for the task. Keeping these two points in mind we design feature extraction, feature selection and orientation detection modules which are tuned to the particular category.

Low level features. We use low level color and edge features for the current task due to the resource limitations on embedded platform on which the system is expected to run. We extract the color moments (mean and variance) of the normalized R and G planes (which lends some robustness against illumination differences) i.e. $R_{norm} = \frac{R}{R+G+B}, \; G_{norm} = \frac{G}{R+G+B}$ with the planes divided into $k \times k$ blocks. We use this feature for estimating the orientation of the Mountains and Monuments class. Further, for Water bodies class color is not enough because of similar colored sky/water. We use edge direction histograms as the texture of the water surface is a good cue for the orientation of the image. The features are calculated for $k \times k$ image blocks. The horizontal and vertical edges images are calculated using Sobel operators. The pixels with small edge responses are discarded and edge direction θ at each edge pixel (x, y) is calculated as, $\theta(x, y) = \tan^{-1} \frac{G_y(x,y)}{G_x(x,y)}$. The edge directions obtained are then used to construct b bin histograms for each block of the image, which are concatenated to form the feature for the full image.

Information theory based feature selection. To reduce the dimensionality and to mine out the most discriminant features we use a recently proposed information theory based feature selection method [11]. The method maximizes the mutual information (MI) among the features and the class labels given by $I(X; Y) = \sum_i \int_{\chi} p_{X,Y}(x, i) \log \frac{p_{X,Y}(x,i)}{p_X(x) p_Y(i)} dx$where X is the random process generating features x and Y is the random process generating labels i. [11] decompose the mutual information (MI) into two components, $I(X; Y) = M(X; Y) + C(X; Y)$, marginal mutual information (MMI) given by $M(X; Y) = \sum_{k=1}^{b} I(X_k; Y)$ and the conjunctive component of mutual information (CCMI) given by $C(X; Y) = \sum_{k=2}^{b} [I(X_k; X_{1,k-1}, Y) - I(X_k; X_{1,k-1})]$ where $X_{i,j}$ is set of features from index i through j and X_k is the k^{th} feature. The MMI measures the discrimination power of individual feature while CCMI measures discrimination power of interdependence of features. They use theoretical tools and empirical validation to arrive at the important conclusion that, for common features, only pairwise dependence of features results in practical gains and modeling dependence of features beyond pairs complicates the model and reduces performance. We use the approximate Infomax algorithm, given in [11], to mine out the most discriminant features.

Gaussian mixture model classifier. We use gaussian mixture models (GMM) to capture the distribution of the features (after feature selection) for the 4 rotation classes of the images. GMM pdf given by,

$$f(x; M) = \sum_{i=1}^{k} \pi_i \frac{1}{(2\pi|\Sigma_i|^d)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right) \tag{1}$$

where $M = (\pi, \mu, \Sigma)$ are the GMM model parameters for the present GMM, is a light weight generative model for capturing the distribution of the feature vectors extracted from images with different rotations. We learn the GMM parameters $\{M_l = (\pi_l, \mu_l, \Sigma_l) | l = 1 \ldots k\}$ using the expectation maximization (EM) algorithm on the reduced feature space after feature selection.

We fit four GMMs, one for each rotation of 0, 90, 180 and 270 degrees and assign orientation to a new image based on the likelihood of the four models of having generated that vector i.e. $l^* = \arg\max_l f_l(x; M_l)$.

Scalable boosting. The Portrait class is more varied because of the different backgrounds. We use a recently proposed scalable boosting algorithm [1] for orientation detection of the Portrait class. First basic images with RGB, HSV normalized versions of RGB, HSV and horizontal and vertical edge maps are generated. Then the features are extracted from those images i.e. the mean and variances of multiple sized blocks and strips (both horizontal and vertical). Finally a modified boosting algorithm [1] is used to learn a strong classifier with the weak classifiers being simple comparison etc. between the feature values as used in [1]. We train the strong classifier for 0 degree vs other rotations. When a new image is presented, we extract only the required parts of the basic images and calculate the features which are required for evaluation of the strong classifier. We calculate the response of all four rotated version of the new image and decide the final orientation of the image by the maximum response value.

3 Experimental Results

We tested our system on a database of about 7000 images collected over the internet and from personal image collections, by an independent testing team. To test our system, we split the dataset randomly into 100 images per category for testing and the rest for training. We did 10 such random splits and report the average performance here.

3.1 Categorization Results

Table 1 gives the confusion matrix for the categorization task, where the column is the true class. The overall performance for the method was 83.6 %, with the most confusion classes being Monuments and Mountains with 10.5 % confusion.

Table 1. Confusion matrix for categorization task (column is the true class)

	Mountains	Monuments	Waterbodies	Portraits
Mountains	83.75	10.50	9.25	2.75
Monuments	5.25	82.50	3.75	9.75
Waterbodies	9.00	2.75	83.50	3.00
Portraits	2.00	4.25	3.50	84.50

3.2 Orientation Detection Result

Mountains and Monuments classes. We use color moment features with Infomax feature selection for the Mountains and Monuments classes. Table 2 shows the confusion matrix for the orientation detection task. The mirror image classes show relatively more confusion among themselves e.g. for Mountains 0 degree images are 8.5 % confused with 180 degrees class. The performance for all the rotations are almost same for both the classes showing that the classifier is not biased towards one rotation. The average accuracy for orientation detection achieved for Mountains and Monuments class was 86.1 % and 79.4 % respectively.

Water bodies class. Table 2 shows the confusion matrix for the water bodies class. Again there is high confusion between mirror image rotations and very less confusion between images rotated with complementary angles. The average performance reached is 74.0 % for water bodies class with uniform performance for all rotations.

Portrait class. Portrait class is the most varied class in terms of appearance due to the varied backgrounds against which the pictures have been taken. Table 2 shows the confusion matrix for the orientation detection of the Portrait classes. Again there is much confusion between the classes with opposite rotations. The performance reached for Portrait class is 80.4 %.

Table 2. Confusion matrices for orientation detection

Mount	0 deg.	90 deg.	180 deg.	270 deg.	Monu	0 deg.	90 deg.	180 deg.	270 deg.
0 deg.	86.2	3.0	9.4	2.9	0 deg.	81.6	6.2	11.0	4.5
90 deg.	3.4	85.2	1.8	8.4	90 deg.	5.0	79.2	5.3	10.4
180 deg.	8.5	2.8	86.6	2.2	180 deg.	8.0	4.8	78.9	7.1
270 deg.	1.9	9.0	2.2	86.5	270 deg.	5.4	9.8	4.8	78.0

Wbod	0 deg.	90 deg.	180 deg.	270 deg.	Port	0 deg.	90 deg.	180 deg.	270 deg.
0 deg.	76.8	1.6	24.7	0.8	0 deg.	81.4	1.1	17.9	1.6
90 deg.	0.6	72.8	0.7	25.3	90 deg.	1.9	79.5	1.1	15.8
180 deg.	21.9	1.1	73.5	1.1	180 deg.	15.2	2.1	79.1	1.1
270 deg.	0.7	24.5	1.1	72.8	270 deg.	1.5	17.3	1.9	81.5

3.3 Discussion and Future Work

When compared to scene categorization method of [8] we achieve a performance of 83.6 % on 4 categories while they achieve 89 % on 4 scene categories. However, we use simple features for reducing time complexity are able to achieve comparable results. Compared to the large scale performance evaluation in [2] we achieve better accuracies c.f. their 79.1 % (except in water bodies) by first categorizing the images into more homogeneous categories and then training the algorithm independently on each category.

To sum up, we have presented a system for categorization and orientation detection of consumer images using low level features and learning methods. The system is meant to be run on a low resource device and so only allows for the computation of simple features. This doesn't allow us to compute higher complexity sift features [7], which could have potentially resulted in better results. As a future task, we would like to try out more complicated features as much allowed by our constraints.

References

1. Baluja, S.: Automated image-orientation detection: a scalable boosting approach. Pattern Analysis and Applications 10, 247–263 (2007)
2. Baluja, S., Rowley, H.A.: Large scale performance measurement of content-based automated image-orientation detection. In: ICIP (2003)
3. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
5. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.-J., Zabih, R.: Image indexing using color correlograms. In: CVPR (1997)
6. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
8. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV 42 (2001)
9. Scholkopf, B., Smola, A.J.: Learning with kernels. Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge (2001)
10. Vailaya, A., Zhang, H.-J., Yang, C., Liu, F.-I., Jain, A.K.: Automatic image orientation detection. IEEE Trans. IP 11(7), 746–755 (2002)
11. Vasconcelos, M., Vasconcelos, N.: Natural image statistics and low-complexity feature selection. PAMI 31(2), 228–244 (2009)
12. Wang, Y., Zhang, H.: Content-based image orientation detection with support vector machines. In: CBAIVL (2001)
13. Wang, Y.M., Zhang, H.: Detecting image orientation based on low-level visual content. CVIU 93, 328–346 (2004)
14. Willamowski, J., Arregui, D., Csurka, G., Dance, C.R., Fan, L.: Categorizing nine visual classes using local appearance descriptors. In: IWLAVS (2004)
15. Zhang, L., Li, M., Zhang, H.-J.: Boosting image orientation detection with indoor vs. outdoor classification. In: WACV (2002)

Effective Visualization and Navigation in a Multimedia Document Collection Using Ontology

Surjeet Mishra and Hiranmay Ghosh

TCS Innovation Labs Delhi, Tata Consultancy Services
{surjeet.mishra,hiranmay.ghosh}@tcs.com

Abstract. We present a novel user interface for visualizing and navigating in a multimedia document collection. Domain ontology has been used to depict the background knowledge organization and map the multimedia information nodes on that knowledge map, thereby making the implicit knowledge organization in a collection explicit. The ontology is automatically created by analyzing the links in Wikipedia, and is delimited to tightly cover the information nodes in the collection. We present an abstraction of the knowledge map for creating a clear and concise view, which can be progressively 'zoomed in' or 'zoomed out' to navigate the knowledge space. We organize the graph based on mutual similarity scores between the nodes for aiding the cognitive process during navigation.

1 Introduction

The retrieval tools and classification hierarchy at digital libraries do not provide a collection overview, making exploration difficult for novice or casual users. A user is often lost in the document collection that can be thought as a network of large number of information nodes in different media forms (text, image, video, etc.). The user, interested in information pertaining to a set of related topics, may not have sufficient subject knowledge to locate it on the knowledge map. An efficient visualization tool allows the user to explore the collection by direct interacting with the view. We present a novel user interface that facilitates visualization of and navigation in large multimedia collection.

Earlier work on collection visualization [1,2,3] (re)-organize the document collection to create different views of the collection based on classical distance measures of information retrieval theories. These measures can be computed for text documents alone and cannot be extended predominantly to multimedia collections. Moreover, these visualization schemes cluster document collections based on inter-document distance measures. The semantic relations between the clusters remain unexplored. We consider the underlying knowledge organization structure is important for visualizing a collection. We propose construction of an ontological map of the collection and map the document nodes on it, thereby making the implicit knowledge organization in the collection explicit. The ontology is automatically created by analyzing the links in a public knowledge resource, namely the Wikipedia, and is delimited to tightly cover the information nodes in the collection. The resulting ontology contains several thousand nodes, which cannot be meaningfully visualized together. We have

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 501–506, 2009.

developed a new method to present an abstraction of the knowledge map, which can be progressively ‘zoomed in’ or ‘zoomed out’ to navigate in the knowledge space. Several approaches to visualize large ontology have been presented in [4] and [5]. Our approach is unique in that it presents abstract views of the knowledge structure at different levels, does not clutter the display with too many nodes and enables flexible zoom and navigation operations. Starting with the overall structure of the collection, a user can deep-dive into some broad areas of interest, progressively refine his information needs and contextually discover the documents in the collection. Our main claims in this paper are (1) automatic organization of the information nodes present in a collection in an ontological structure, and (2) intelligently creating and presenting suitable abstractions of information structure to facilitate the navigation process. A prototype has been developed with a collection of over 650 book and documentaries taken from a multimedia document collection[1].

The paper is organized as follows. Section 2 describes our work in detail. Section 3 has implementation details and illustrative examples. Section 4 concludes the paper.

2 Description of the Work

Our work has two major components, namely (a) creating ontology from Wikipedia, covering all information nodes in a collection, and (b) providing abstract views and navigation facility in the collection with the ontology.

2.1 Ontology Creation

For effective visualization and navigation, the documents in the collection need to be interrelated on a knowledge space. Ontology is a formal tool to represent a bounded knowledge domain. Research in informal ontology (or folksonomy) [6] has assumed significance because of the challenges in creating formal ontology for an ill-defined domain like the collection of a bookstore. The categorization structure of an online public encyclopedia, e.g. Wikipedia, represents an informal knowledge organization. Our use of Wikipedia to create an informal ontology is motivated by [7]. We have created a constrained ontology by restricting it to contain the category nodes, sufficient to cover the information nodes in the collection. We treat the category node “Main topic classifications” (MTC), which cover all Wikipedia topics (articles), as the root node of our ontology.

We assume sufficient metadata is associated with each document, to relate the document to one or more Wikipedia topic. Once the topics pertaining to a document have been identified, we create the ontology superstructure for that document in a bottom-up manner starting from a topic and going up the Wikipedia category hierarchy ladder using Depth First Traversal, till we reach the MTC node. This process is repeated for every topic in a document and for every document in the collection, merging the common sub-graph as soon as it is discovered. Thus, a document is associated with one or more leaf category nodes in the ontology (Fig.1). The incremental way of building helps in updating the ontology when new documents are added to a collection.

[1] We thank ‘Scholars Without Borders’, an online bookstore for scholarly publications, for making its collection available for this research.

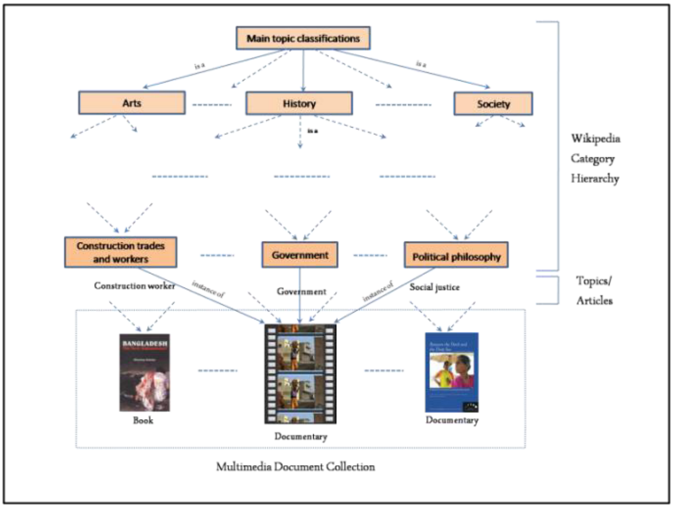


Fig. 1. Knowledge organization in Ontology

We represent the knowledge graph as a singly rooted Directed Acyclic Graph (DAG), with MTC node as the root, to restrict the user from looping in cycles during navigation. We have discovered and removed some cycles in Wikipedia categorization (e.g. Psychology → Behavior → Human behavior → Psychology) by backtracking when an edge completes a loop and by removing the Orphan Nodes[2].

Thus, the ontology created by us is represented by a singly rooted DAG <V, E>, where set of vertices comprise a set of documents D and set of categories C. The set E comprises two types of edges: (a) "is-a", that connects a category node to another, and (b) "instance-of", that connects a document node to a category node. We represent these relations by the symbols '$\subset$' and '$\langle$' respectively. Note that the document nodes appear as the leaf nodes. We observe the following transitive properties:

1. Let $c1, c2, c3 \in C$. If $c1 \subset c2$ and $c2 \subset c3$, then $c1 \subset c3$
2. Let $c1, c2 \in C$, $d \in D$. If $d \langle c1$ and $c1 \subset c2$, then $d \langle c2$

These properties has been used to abstract the knowledge map at multiple levels and to allow 'zoom-in' and 'zoom-out' operations, that is explained in next sub-section.

2.2 Navigation Interface

The knowledge map so created contains a few thousands of category nodes, posing a challenge to visualization, though several large graph visualization techniques have been developed. The folding and unfolding of subject taxonomies [4] provides a shallow overview of the knowledge organization and may pose a significant navigational challenge to a casual user. Hyperbolic distortion [4,8] present nodes of interest at the center of view with magnification, while pushing the other nodes to the periphery which prevents visualizing the overall knowledge structure. Another approach is clustering of nodes [3,4], where semantic labeling of the clustered nodes is a challenge.

[2] Category node with no parent category.

We present an abstract view of the ontology with a few selected category nodes and edges that depicts the overall structure of the graph. We compute Strahler Score [9], which is a measure of structural complexity of the sub-graph below the node, for all the category nodes in the graph and select a set of top-ranking nodes. The number of nodes selected depends on the area allocated for visualizing the graph. Let the set of these summary nodes be designated by V'. We construct the summary graph by drawing summary edges $v_1 \rightarrow v_2$ ($v_1, v_2 \in V'$) if there is a path from v_1 to v_2, since both the semantic relationships are transitive.

For a graph G = <V,E>, where V and E are set of vertices and edges respectively, and a given set of vertices V', the Summary Graph will be G' = <V',E'> where,

$$V' \subseteq V, \text{ and}$$

$$E' = \{(v_1 \rightarrow v_2) \mid (v_1, v_2 \in V') \quad \text{AND}$$

$$(\exists \text{ path p from } v_1 \text{ to } v_2 \text{ in G} \mid !\exists\ v_3 \in p, v_3 \in V'))\} \tag{1}$$

Wikipedia category nodes show high degree of interconnectivity. The large amount of edges present in the summary graph clutters the view. So, we create a minimum weighted DAG around the selected nodes by using Chu-Liu/ Edmond's algorithm [10,11] we get the abstract view graph G" = <V', E"> where E" $\subseteq$ E'. Visualization is further aided by placing the category nodes with higher semantic similarities closer to each other. The similarity of the nodes are measured with Jaccard's coefficient [12]

$$\text{Sim}(v_a, v_b) = P(S_a \cap S_b)/P(S_a \cup S_b) \tag{2}$$

where, S_a, S_b are the set of descendent nodes of selected nodes v_a, v_b respectively and P(S) is cardinality of set S. The user can click on any of the nodes depicted on the screen to navigate on the ontology graph. As a user zooms in, the topics below the selected node gets magnified, the topics in the other regions shrink and some of the dynamically created document clusters split. Thus, a contextual view of the collection is dynamically created as a user navigates the collection. Completeness of the summary view demands that all the category nodes in the original graph should be navigable. To guarantee completeness, we add the immediate parent and child nodes of the node currently being explored to the dynamic knowledge graph. Once the category graph is organized, the cluster of document nodes is attached to each category node signifying the number of documents related to that category.

3 Implementation Details and Illustrative Examples

We have created the knowledge graph for well over 650 documents from a multimedia collection using JWPL[3]. Fig.2 shows the increase in number of ontology nodes with increase in the number of documents in the collection. The number of ontology nodes tends to saturate at a finite value, after sufficient number of documents are inducted, thereby ascertaining tractability of the problem.

[3] Java Wikipedia Library, http://www.ukp.tu-darmstadt.de/software/jwpl/

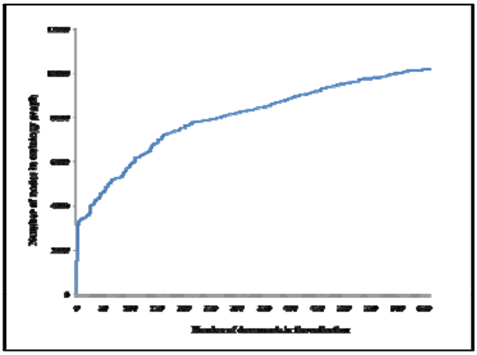


Fig. 2. Increase in number of nodes in ontology with number of documents in the collection

The graph visualization has been implemented with JUNG 2.0[4]. We illustrate the navigational steps with a few illustrative examples.

1. To discover the documents ("History and Ideology" and "India") related to tapestry, a user may traverse the path:
   MTC → Arts → Visual arts → Textile arts → Tapestries
2. To discover documents on Tourism in India, a user traverses the path:
   MTC → Geography → Geography by place → Geography by countries → Members of the Commonwealth of Nation → India → Tourism in India

Another interesting aspect of the traversal is user's shift of attention. For example, a user while studying the document "They who walks the mountain" (TWWTM), which speaks about the 'life-style of Himalayan people", may be tempted to find more documents on the religious aspect (Buddhism) and end up reaching a document "The Dalai Lama" (TDL). The corresponding navigation path can be:

TWWTM → Life in Himalaya → Religion → Buddhism → TDL

Some of the navigational steps are shown in Fig. 3.

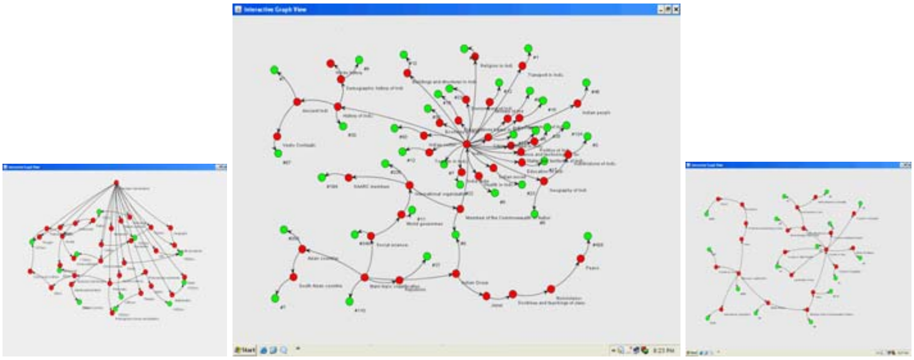


Fig. 3. Abstract views of: (a) overview of ontology (b) intermediate navigation stage (c) final view on "Tourism in India"

[4] Java Universal Network/Graph Framework, http://jung.sourceforge.net/index.html

The red bubbles indicate the category nodes and the green bubbles indicate the cluster of associated documents to that category, labeled by the number of documents in the cluster. The cluster expands on clicking and each document can then be accessed.

4 Conclusion

The work presented in this paper provides a novel interface for visualization and navigation in a multimedia collection using automatically created ontology. This work can be used to build various applications dealing with multimedia assets in libraries; online book stores, movie stores; video sharing portals; etc. The system can be extended to work with document segments (scenes of video and chapters from literature artifacts) as it can handle any form of information nodes.

References

1. Grobelnik, M., Mladenic, D.: Efficient visualization of large text corpora. In: Proceedings of the Seventh TELRI seminar, Dubrovnik, Croatia (2002)
2. Paulovich, F.V., Minghim, R.: HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections. IEEE Transactions on Visualization and Computer Graphics 14(6) (2008)
3. Berendonck, C.V., Jacobs, T.: Bubbleworld- A New Visual Information Retrieval Technique. In: Australasian Symposium on Information Visualisation (2003)
4. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology Visualization Methods—A Survey. ACM Computing Surveys 39(4) (2007)
5. Geroimenko, V., Chen, C. (eds.): Visualizing the Semantic Web-XML-Based Internet and Information Visualization, 2nd edn. Springer, Heidelberg (2006)
6. Hepp, M., Bachlechner, D., Siorpaes, K.: Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203–217 (2008)
7. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Large Ontology from Wikipedia and WordNet (2008)
8. Heymann, S., Tham, K., Kilian, A., Wegner, G., Rieger, P., Merkel, D., Freytag, J.C.: Viator - A Tool Family for Graphical Networking and Data View Creation. In: Proc. of the 28th VLDB Conference, Hong Kong, China (2002)
9. Delest, M., Herman, I., Melancon, G.: Tree Visualization and Navigation Clues for Information Visualization. Computer Graphics Forum 17(2) (1998)
10. Chu, Y.J., Liu, T.H.: On the shortest arborescence of a directed graph. Science Sinica 14, 1396–1400 (1965)
11. Edmonds, J.: Optimum branchings. Journal of Research of the National Bureau of Standards 71B, 233–240 (1967)
12. Doan, A., Madhavan, J., Domingos, P., Halevy, A.: Learning to map between ontologies on the semantic web. In: Proc. of the 11th International WWW Conference (2002)

Using Concept Recognition to Annotate a Video Collection

Anupama Mallik and Santanu Chaudhury

Electrical Engineering Department, IIT Delhi
ansimal@gmail.com, schaudhury@gmail.com

Abstract. In this paper, we propose a scheme based on an ontological framework, to recognize concepts in multimedia data, in order to provide effective content-based access to a closed, domain-specific multimedia collection. The ontology for the domain is constructed from high-level knowledge of the domain lying with the domain experts, and further fine-tuned and refined by learning from multimedia data annotated by them. MOWL, a multimedia extension to OWL, is used to encode the concept to media-feature associations in the ontology as well as the uncertainties linked with observation of the perceptual multimedia data. Media feature classifiers help recognize low-level concepts in the videos, but the novelty of our work lies in discovery of high-level concepts in video content using the power of ontological relations between the concepts. This framework is used to provide rich, conceptual annotations to the video database, which can further be used to create hyperlinks in the video collection, to provide an effective video browsing interface to the user.

1 Introduction

Meaningful access to the ever-increasing multimedia data in the pubic domain faces the crunch of available conceptual metadata and annotation text. This textual metadata is helpful in bridging the semantic gap between high-level semantic concepts and the low-level content-based media features. Video annotation is essential for successful content-based video search and retrieval, but done manually it is tedious and prone to inaccuracy. In [1], Zha et al propose to refine video annotation by leveraging the pairwise concurrent relation among video concepts. In [2], the authors have systematically studied the problem of event recognition in unconstrained news video sequences, by adopting the discriminative kernel-based method. Concept Recognition using an ontology for the purpose of enhancing content-based multimedia access as attempted in our work, is a relatively new approach.

In our work, we propose a scheme based on an ontological framework, to recognize concepts in multimedia data, in order to generate rich, conceptual annotations for the data. The annotations generated by this scheme provide associations between the concepts in the domain and the content in the multimedia files, forming a basis for effective content-based access to the multimedia data in a closed, domain-specific collection. The highly specialized knowledge that experts of a scholarly domain have, is encoded into an ontological representation of the domain, and is refined by learning from observables in the multimedia examples of the domain. This approach to concept learning has been detailed in our earlier work [3].

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 507–512, 2009.

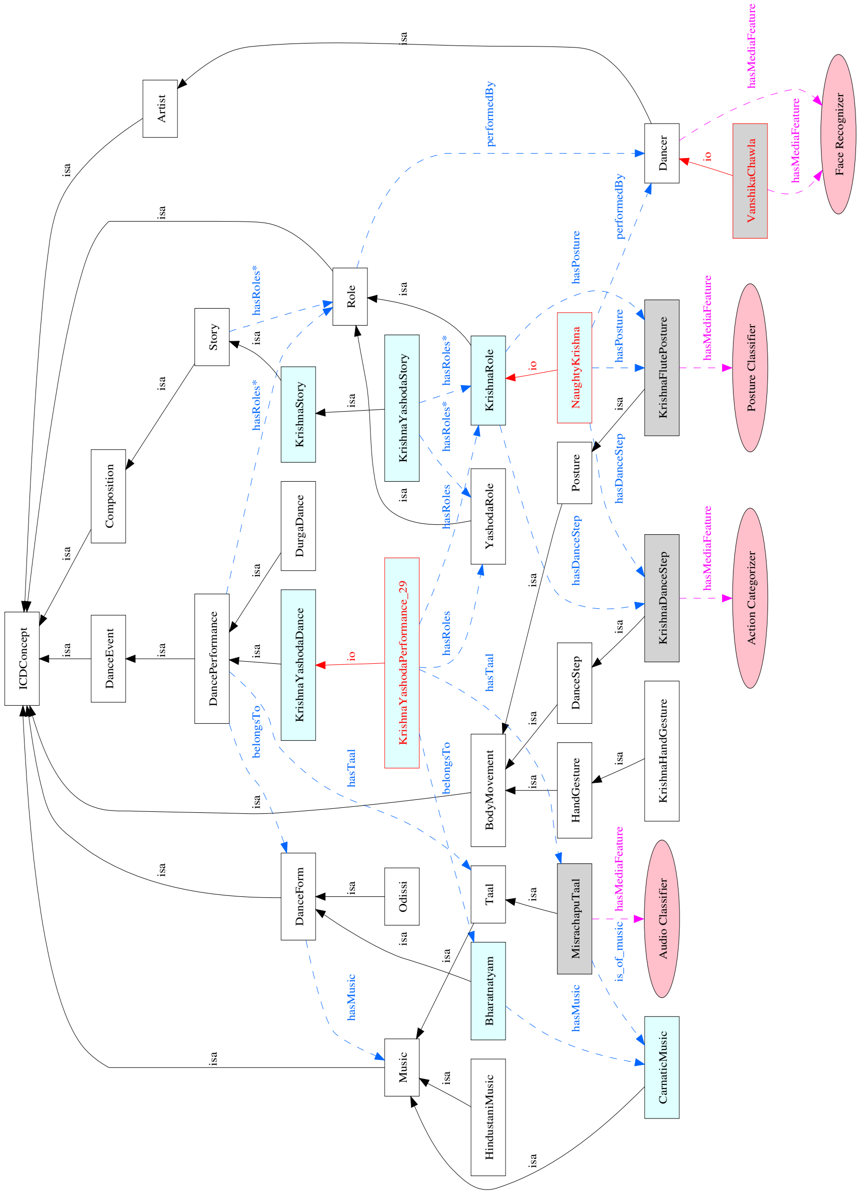


Fig. 1. MOWL Ontology of Indian Classical Dance

The key contribution of our current work is the discovery of high-level concepts in video content using the power of the MOWL encoded ontology to propagate media properties across concepts. We have shown the success of our technique by applying our work to a cultural heritage domain of Indian classical dance. The conceptual annotations generated can be used to create hyperlinks in the video collection, to provide an effective video browsing interface to the user.

2 Concept Recognition Using MOWL Ontology

We have extended the existing ontology representation based on OWL to include a perceptual description of the concepts and formalized it as Multimedia Web Ontology Language(MOWL). MOWL supports probabilistic reasoning with Bayesian Networks in contrast to crisp Description logic based reasoning with traditional ontology languages [4]. In this paper, we have used a robust evidential reasoning framework around MOWL, where a concept can be recognized in a multimedia entity on the evidential strength of physical observations of some its expected media properties.

Figure 1 shows a snippet of the Indian Classical Dance (ICD) ontology represented graphically. Root Node represents an ICD concept, some of which are shown as 'Music', 'DanceForm', 'Artist' and 'Composition'. This snippet focuses on an important mythological figure - an Indian God named Krishna. Stories about Krishna abound in folklore, and all classical dances of India have performances dedicated to events in his life. One of the events depicted here is enactment of a scene between Krishna and his mother, Yashoda through a performance in Bharatnatyam dance form. Linkages and dependencies between a 'Story', a 'Role', a 'DancePerformance', a 'DanceForm', a 'Dancer', various 'Body Movements' are encoded in the MOWL ontology. The leaf nodes in elliptical shape denote 'Media Feature' nodes and represent various media classifiers like posture recognizer, face recognizer. This graphical representation of the ICD ontology represents a Bayesian Network. The Conditional Probability values are not shown here in order to preserve the visual clarity of the diagram. Evidence is gathered at the leaf nodes, as different media features are recognized or classified by the media classifiers. If evidence is above a threshold, the media feature node is instantiated. These instantiations result in belief propagation in the Bayesian Network, and posterior probability at the associated concept nodes is computed. The algorithm for recognizing the concepts in this BN is as following :

```
Inputs : 1) Video segment $\mathcal{V}$ for which concepts are to be recognized
         2) Bayesian Network $\Omega$ of the relevant MOWL ontolgy segment
Output : Set $\mathcal{C}$ of Recognized Concepts
Algorithm :
1. For each leaf-node Concept $\mathcal{LC}_i$ in $\Omega$,
   i. Run the appropriate Media Feature classifier.
   ii. If Classification evidence for the concept $>$ threshold,
      a. Add $\mathcal{LC}_i$ to the set $\mathcal{I}$ of instantiated nodes
      b. Add $\mathcal{LC}_i$ to the result set $\mathcal{C}$ of recognized nodes.
2. Carry out Belief Propagation in the BN $\Omega$.
3. For each node $\mathcal{IC}_i$ in $\mathcal{I}$
   i. Compute the set $\mathcal{RC}$ of Related concepts at next higher level.
```

```
  ii. For each node $\mathcal{RC}_i$ in $\mathcal{RC}$
    a. Compute the posterior probability P($\mathcal{RC}_i$) at $\mathcal{RC}_i$
    b. If P($\mathcal{RC}_i$) $>$ threshold,
      •Add $\mathcal{RC}_i$ to the set $\mathcal{I}$ of instantiated nodes
      •Add $\mathcal{RC}_i$ to the to the result set $\mathcal{C}$ of recognized nodes.
4. Iterate steps 2 and 3 till Root node is reached.
```

3 Annotation Generation

The input to our concept-recognition scheme is an initial multimedia ontology of the domain constructed with the help of domain knowledge provided by a group of domain experts, and fine-tuned by learning from the training set of annotated videos [3]. A semi-automated annotation generation module provides an interface where domain concepts present in the video content are recognized automatically by the system, and presented to the annotator to verify and confirm their existence. The module consists of 5 functional components :

- **Object/Feature Extractor:** This module extracts the features/objects from the multimedia data. The extracted features are given to the XML generator, to store them in XML format.
- **MOWL Parser:** This module is responsible for generating the Bayesian network from the given MOWL ontology.
- **Concept Recognizer:** The task here is to recognize the high-level semantic concepts in multimedia data with the help of low-level media-based features. This module gets the feature values either by invoking the feature extractor or from the feature database. The concept recognizer either highlights or prompts the concept to the annotator, resulting in a kind of supervised annotation. It can also directly convey the concept/s recognized to the XML generator.
- **Classifiers:** Media Feature classifiers are trained with feature vectors extracted from the multimedia data. These are detailed in section 4.
- **XML based Annotation generator:** This module is responsible for generating the XML. The inputs to this module are the manual annotations, conceptual annotations and features of the multimedia data (in MPEG-7 format) and output is the XML file as per MPEG-7 standard, containing the video annotation as well as media based feature-vectors.

4 Experimental Results

We tested our ontology based annotation scheme on a captive collection of videos which belong to the scholarly domain of Indian Classical Dance(ICD). We compiled a heritage collection by gathering dance videos from different sources. We started work with a data set of approximately 200 videos of duration 10 to 15 minutes. These consist of dance performances of different Indian classical dance forms - Bharatnatyam, Odissi, Kuchipudi and Kathak; plus music performances of Indian classical music forms - Carnatic Music and Hindustani Music.

With reference to the snippet in Fig. 1, concept-recognition occurs with belief propagation in BN. The concept nodes highlighted in gray color are the low-level concepts which are recognized due to presence of the media features in data. These are 'Misrchapu Taal'(a taal/beat in Carnatic music), 'KrishnaDanceStep', 'KrishnaPosture' and 'VanshikaChawla' (a dancer). Due to further belief propagation in the BN, higher level concept nodes (in cyan color) are recognized to be present in the video. Conceptual annotations are generated and attached to the video through the Annotation Generation module detailed in section 3. Videos are hyperlinked if they are annotated with common concepts, or ontologically related concepts. This hyperlinking formed the basis of an ontology-based Browsing application for the video database(not detailed here due to space constraint). Some of the Media feature classifiers used by our concept-recognition scheme are detailed below :

4.1 Human Action Categorization Using pLSA

Our framework for detecting human action categories includes the following steps :

- Spatio-temporal interest points are extracted for frames of a video.
- The extracted spatio-temporal interest points are used in the bag of words approach to summarize the videos in the form of spatio-temporal words.
- The process automatically learns the probability distributions of the spatio-temporal words and intermediate topics for detecting action categories using pLSA technique [5].
- The topic-to-video probability distributions we get from pLSA training and testing, are fed to an SVM classification scheme for categorisation of actions.

For performing pLSA categorization, some of the recognizable dance actions selected from Bharatnatyam dance were - **Sitting and getting up**, **Side-stepping**, **Taking a circle**, **Krishna Step**, **Teermanam Step**. Approximately 30 video shots of each action were submitted to the pLSA for training. We performed 6-fold cross-validation tests on 77 videos to test the classification of the various dance actions by pLSA technique. The accuracy of classification was found to be approx. 76.8% on an average.

4.2 Dance Posture Recognition Using SIFT

We have used the SIFT approach [6] to recognize dance postures in still images taken from dance videos. Steps of our computation are :

- Collect labelled examples of Dance Posture images from different dance videos.
- Extract SIFT descriptors for all the images, and quantize the SIFT descriptors by K-means clustering algorithm to obtain a discrete set of local N_s SIFT words.
- A posture image P_i is represented by an Indicator vector $iv(P_i)$, which is a histogram of its constituent SIFT words,

$$iv(P_i) = \{n(P_i, s_1), ..., n(P_i, s_j), ..., n(P_i, s_(N_s))\} \tag{1}$$

where $n(P_i, s_j)$ is the number of local descriptors in image P_i, quantized into SIFT word s_j.

- Train an SVM classifier with the indicator vectors to classify the postures.

We extracted about 288 images of various Dance postures from our set of ICD videos. These were classified into 7 broadly similar dance postures, as shown in table 1. An average of 232 detected points (depending on image content) and K-means clustering with 50 cluster centers yielded indicator vectors for all the 288 images. A 10-fold

Table 1. SVM Classification Results for Dance Postures

$Classes$	$TPRate$	$FPRate$	$Precision$	$Recall$	$F-Measure$	$ROCArea$
RightAnjali	0.867	0	1	0.867	0.929	0.933
LeftAnjali	0.966	0	1	0.966	0.982	0.983
FrontPranam	0.931	0.004	0.964	0.931	0.947	0.964
ArmsUp	0.731	0.068	0.704	0.731	0.717	0.831
Squat	1	0	1	1	1	1
KrishnaPose	0.936	0.057	0.889	0.936	0.912	0.94
HandsOnWaist	0.552	0.039	0.615	0.552	0.582	0.757

cross-validation using SVM classifer on Weka machine learning framework, yielded an accuracy of 87.8%. The detailed results are shown in table 1.

5 Conclusion

In this paper, we have outlined a novel approach to recognize concepts in a closed collection of videos belonging to a scholarly domain, and use it to generate conceptual annotations for videos at different levels of granularity. The multimedia ontology is capable of incorporating uncertainties attached to media observables, and thus offers a probabilistic framework, which can be enhanced using ontology learning from annotated data. Thus the whole system is self-enhancing where ontology is refined from annotated data, and data annotation is improved based on fresh, refined knowledge from the ontology. This ontological framework can offer a robust ground for several multimedia search, retrieval and browsing applications.

References

1. Zha, Z.J., Mei, T., Hua, X.S., Qi, G.J., Wang, Z.: Refining video annotation by exploiting pairwise concurrent relation. In: MULTIMEDIA 2007: Proceedings of the 15th international conference on Multimedia, pp. 345–348. ACM, New York (2007)
2. Xu, D., Chang, S.F.: Video event recognition using kernel methods with multilevel temporal alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1985–1997 (2008)
3. Mallik, A., Pasumarthi, P., Chaudhury, S.: Multimedia ontology learning for automatic annotation and video browsing. In: MIR 2008: Proceeding of the 1st ACM international conference on Multimedia information retrieval, pp. 387–394. ACM, New York (2008)
4. Ghosh, H., Chaudhury, S., Kashyap, K., Maiti, B.: Ontology specification and integration for multimedia applications. Springer, Heidelberg (2006)
5. Hofmann, T.: Probabilistic latent semantic analysis. In: Proc. of Uncertainty in Artificial Intelligence, UAI 1999, pp. 289–296 (1999)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 20, 91–110 (2003)

Variational Bayes Adapted GMM Based Models for Audio Clip Classification

Ved Prakash Sahu, Harendra Kumar Mishra, and C. Chandra Sekhar

Speech and Vision Lab, Dept. of Computer Sc. & Engg.,
Indian Institute of Technology-Madras, India
{vedsahu,harendra,chandra}@cse.iitm.ac.in

Abstract. The most commonly used method for parameter estimation in the Gaussian mixture models (GMMs) is maximum likelihood (ML). However, it suffers from the overfitting when the model complexity is high. Adapted GMM is an extended version of GMMs and it helps to reduce the overfitting in the model. Variational Bayesian method helps in determining optimal complexity so that it avoids overfitting. In this paper we propose the variational Bayes learning method for training the adapted GMMs. The proposed approach is free from overfitting and singularity problems that arise in the other approaches. This approach is faster in training and allows a fast-scoring technique during testing to reduce the testing time. Studies on the classification of audio clips show that the proposed approach gives a better performance compared to GMMs, adapted GMMs, variational Bayes GMMs.

Keywords: GMM, variational learning, Bayesian adaptation.

1 Introduction

This paper aims to investigate the behavior of different types of Gaussian mixture based models such as conventional GMMs, adapted GMMs referred to as Gaussian mixture model-universal background model (GMM-UBM), and variational Bayesian GMMs (VB-GMMs) for audio clip classification task. We propose to use the variational Bayes adapted GMMs (VB-GMM-UBM) for audio clip classification and compare it with the GMMs, GMM-UBM, and VB-GMMs.

The GMM commonly uses the maximum likelihood (ML) method for parameter estimation. However, the ML method suffers from the problem of overfitting due to the presence of singularities, when the model complexity is high [1]. One way to handle this is by adapting the parameters of GMM using the training examples. The GMM-UBM [2] is one such technique, where the UBM is built as a large GMM from the data of all classes. The model for a class is obtained by updating the parameters of the UBM using the training examples of the respective class using an adaptation method [2]. This provides a tighter coupling between the class model and the UBM, and gives a better performance than the decoupled models like conventional GMMs. Training a GMM-UBM is much faster than the conventional GMMs and also allows a fast-scoring technique [2] during testing. But it

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 513–518, 2009.

suffers from the overfitting when the model complexity is high. To overcome this problem, the VB-GMMs [1] can be used. The VB approach for GMMs helps to determine the optimal number of components to build a model [1]. It prunes the mixture components which are not useful. Hence the variational approach does not suffer from overfitting and has good generalization capabilities [3]. Variational approach assumes that the parameters are not uniquely specified, but instead are modeled by probability density functions. This introduces an additional set of hyperparameters for the distributions of parameters [1].

In this work, we propose to use the VB-GMM-UBM. In the proposed approach, we build the UBM using the variational approach instead of the ML method. Training the UBM using the variational approach gives an optimal number of components in the UBM. The hyperparameters of each component in the UBM are used to derive the Gaussian parameters for the corresponding component in the UBM. The class model is obtained by Bayesian adaptation that adapts the Gaussian parameters of the UBM using the training examples of that class. The proposed approach is observed to be faster in training. It is free from singularity and overfitting problems that arise in the other approaches. This generalizes the model well and hence its performance is better compared to the other models.

Section 2 describes the proposed VB-GMM-UBM approach. Section 3 presents the experimental studies and the summary of the work is presented in section 4.

2 Variational Bayes Adapted GMMs

2.1 Variational Gaussian Mixture Models

The GMMs can be trained using the variational approach. The distribution of d-dimensional feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n, .., \mathbf{x}_N\}$ can be written as a linear superposition of Gaussians in the form $p(\mathbf{x}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_\mathbf{k}, \mathbf{\Lambda}_\mathbf{k}^{-\mathbf{1}})$, where π_k, $\mu_\mathbf{k}$ and $\mathbf{\Lambda_k}$ are the mixture weight, mean vector and precision matrix, respectively for k^{th} Gaussian. In Bayesian approach, we assume priors for the parameters. The conjugate priors are the Dirichlet distribution $Dir(\pi|\alpha)$ for mixture weights, Gaussian distribution $\mathcal{N}(\mu|\mathbf{m}, (\beta\mathbf{\Lambda})^{-\mathbf{1}})$ for mean, and Wishart distribution $\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, \nu)$ for precision matrix. Here $\{\alpha, \beta, \nu, \mathbf{m}, \mathbf{W}\}$ form a set of hyperparameters. This set of hyperparameters can be obtained using the EM approach [1] as follows:

(E)xpectation Step : Responsibility r_{nk} of the k^{th} component for the n^{th} example $\mathbf{x}_n$ is defined as $r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}$. Here,

$$\ln \rho_{nk} = d\beta_k^{-1} + \nu_k(\mathbf{x}_n - \mathbf{m}_k)^T \mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k) + \sum_{i=1}^{d} \psi\left(\frac{\nu_k + 1 - i}{2}\right)$$
$$+ d\ln 2 + \ln|\mathbf{W}_k| + \psi(\alpha_k) + \psi(\sum_{k=1}^{K} \alpha_k), \tag{1}$$

and ψ is the gamma function.

(M)aximization Step : Hyperparameters are updated as follows:

$$\alpha_k = \alpha_0 + N_k, \qquad \beta_k = \beta 0 + N_k, \qquad \mathbf{m}_k = \frac{1}{\beta 0 + N_k}(\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k),$$

$$\nu_k = \nu_0 + N_k, \qquad \mathbf{W}_k^{-1} = \mathbf{W}_0^{-1} + N_k \mathbf{S}_k + \frac{\beta_0 N_k}{\beta_k}(\bar{\mathbf{x}}_k - \mathbf{m}_0)(\bar{\mathbf{x}}_k - \mathbf{m}_0)^T, \quad (2)$$

where,

$$N_k = \sum_{n=1}^{N} r_{nk}, \qquad \bar{\mathbf{x}}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n, \qquad \mathbf{S}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T. \quad (3)$$

Here $N_k, \bar{\mathbf{x}}_k, S_k$ are the intermediate values used to update the hyperparameters. Initial values for priors $\{\alpha_0, \beta_0, \nu_0, \mathbf{m}_0, \mathbf{W}_0\}$ are chosen as suggested in [3].

2.2 Universal Background Model (UBM)

In the VB-GMM-UBM system, the first step is to generate the UBM [2,4]. The UBM is a single and class-independent large GMM. Given N training examples, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we train the UBM using the VB-EM method described in the previous subsection. This gives the hyperparameters for each mixture component in the UBM. However, Bayesian adaptation needs Gaussian parameters of the mixture components, that are used to build the class specific models by updating the UBM parameters. Hence, the conversion of hyperparameters to the Gaussian parameters is performed in the UBM before the adaptation process starts. In the process of obtaining the Gaussian parameters, we first determine the probabilistic alignment of training examples to the UBM mixture components. The responsibility of a training example $\mathbf{x}_n$ for the k^{th} mixture in the UBM is computed as follows:

$$\zeta_{nk} = \frac{\alpha_k p_k(\mathbf{x}_n)}{\sum_{j=1}^{K} \alpha_j p_j(\mathbf{x}_n)}. \quad (4)$$

Here, $p_k(\mathbf{x}_n)$ is calculated using the student's t-distribution for k^{th} mixture of the UBM as $p_k(\mathbf{x}_n) = \mathrm{St}(\mathbf{x}_n | \mathbf{m}_k, \mathbf{L}_k, \nu_k + 1 - d)$, where $\mathbf{m}_k$ is the mean vector, $\mathbf{L}_k$ is the precision matrix given by $\mathbf{L}_k = (\beta_k(\nu_k + 1 - d)/(1 + \beta_k))\mathbf{W}_k$, and K is the number of mixtures in the UBM. Then we use the responsibility ζ_{nk} to compute the Gaussian parameters such as weight w_k, mean μ_k and variance σ_k^2 of k^{th} component as follows:

$$w_k = \sum_{n=1}^{N} \zeta_{nk}, \qquad \mu_k = \frac{1}{w_k}\sum_{n=1}^{N} \zeta_{nk}\mathbf{x}_n, \qquad \sigma_k^2 = \frac{1}{w_k}\sum_{n=1}^{N} \zeta_{nk}(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^t. \quad (5)$$

2.3 Adaptation of Class Model from UBM

Once the Gaussian parameters of the UBM are computed, the next step is to get the GMM for each class. The class model is obtained by adapting the Gaussian parameters of the UBM using the training examples of the corresponding class [2,5]. Given a UBM and R training vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_R\}$ from a class, we first determine the probabilistic alignment of the training vectors onto the Gaussian mixture components in UBM. For k^{th} mixture in the UBM, the responsibility ξ_{rk} is computed as follows:

$$\xi_{rk} = \frac{w_k p_k(\mathbf{x}_r)}{\sum_{j=1}^{K} w_j p_j(\mathbf{x}_r)}. \tag{6}$$

Then ξ_{rk} and $\mathbf{x}_r$ are used to compute the new parameters as follows:

$$\tilde{w_k} = \sum_{r=1}^{R} \xi_{rk}, \qquad \tilde{\mu_k} = \frac{1}{\tilde{w_k}} \sum_{r=1}^{R} \xi_{rk}\mathbf{x}_r, \qquad \tilde{\sigma_k}^2 = \frac{1}{\tilde{w_k}} \sum_{r=1}^{R} \xi_{rk}(\mathbf{x}_r - \tilde{\mu}_k)(\mathbf{x}_r - \tilde{\mu}_k)^t. \tag{7}$$

The new parameters $\{\tilde{w_k}, \tilde{\mu_k}, \tilde{\sigma_k}^2\}$ obtained from the class specific training data are used to update the old UBM parameters $\{w_k, \mu_k, {\sigma_k}^2\}$ for k^{th} mixture to get the adapted parameters as follows:

$$\hat{w_k} = [\alpha_k^w \tilde{w}_k / R + (1 - \alpha_k^w) w_k]\gamma, \tag{8}$$

$$\hat{\mu_k} = \alpha_k^m \tilde{\mu_k} + (1 - \alpha_k^m)\mu_k, \tag{9}$$

$$\hat{\sigma_k}^2 = \alpha_k^v \tilde{\sigma_k}^2 + (1 - \alpha_k^v)({\sigma_k}^2 + {\mu_k}^2) - \hat{\mu_k}^2, \tag{10}$$

where ${\mu_k}^2$ and $\hat{\mu_k}^2$ are diagonal elements of matrix $\mu_k {\mu_k}^t$ and $\hat{\mu_k}\hat{\mu_k}^t$ respectively.

The adaptation coefficients $\{\alpha_k^w, \alpha_k^m, \alpha_k^v\}$ control the balance between the old and new parameters for the weights, means, and variances respectively. The scale factor, γ is computed over all adapted mixture weights to ensure that their sum is unity. For each mixture and each parameter, a data-dependent adaptation coefficient $\alpha_k^\rho, \rho = \{w, m, v\}$, is defined as $\alpha_k^\rho = \frac{\tilde{w_k}}{\tilde{w_k} + r^\rho}$, where r^ρ is a fixed relevance factor for parameter ρ. The performance is insensitive to relevance factors in the range $\lfloor 8 - 20 \rfloor$ [2]. We have used a relevance factor of 16. The update of parameters described in Eqs. (8)-(10) can be derived from the general maximum a posteriori (MAP) estimation for a GMM using constraints on the prior distribution described in [6].

In the next section, we study the performance of conventional GMMs, GMM-UBM, VB-GMM and VB-GMM-UBM for audio clip classification task.

3 Studies and Results

In this section we present our studies on audio clip classification. We compare the performance of the proposed approach with that of conventional GMMs, GMM-UBM and VB-GMM.

The audio data set used in our study includes 180 audio clips from 5 categories : advertisement, cartoon, cricket, football, and news from different TV channels, where 120 clips used for training and remaining clips used for testing. The duration of an audio clip is approximately 20 seconds. The 14-dimension features used are clip-based, and are extracted from frame-based features [7]. The features for

Table 1. Audio clip classification accuracy (in %) for the GMMs, GMM-UBM, VB-GMMs, and VB-GMM-UBM

Audio Class	GMMs	GMM-UBM	VB-GMMs	VB-GMM-UBM
Ads	67.21	77.04	68.85	78.68
Cartoon	72.15	91.13	86.07	96.20
Cricket	66.67	70.00	78.33	93.33
Football	57.38	80.32	96.72	96.72
News	73.09	83.33	89.74	91.02
Average	67.30	80.36	83.94	91.18

Table 2. Performance of GMM-UBM and VB-GMM-UBM for adapting different sets of parameters (W=weight, M=mean, V=variance)

Adapted Set of Parameters	Performance(%)	
	GMM-UBM	VB-GMM-UBM
W	15.89	74.36
M	74.85	82.01
V	28.90	73.74
W, M	78.03	90.26
W, V	33.53	86.72
M, V	79.34	89.38
W, M, V	80.36	91.18

clip-based are volume standard deviation, volume dynamic range, volume undulation, 4Hz modulation energy, ZCR standard deviation, nonsilence ratio, pitch standard deviation, similar pitch ratio, nonpitch ratio, frequency centroid, bandwidth, and energy ratio in subbands.

Table 1 shows the classification accuracy of conventional GMMs, GMM-UBM, VB-GMMs, and VB-GMM-UBM on audio clip data. It is seen that the GMM-UBM performs better than the conventional GMMs. In GMM-UBM, adapted class models localize the most unique features for each target audio class and perform better than conventional GMMs. It is also seen that VB-GMM gives a better average performance than the GMM-UBM. This is because the UBM trained by ML method suffers from singularity and overfitting problems. These problems also propagate to the final class model through the adaptation process. Such problems are resolved when the UBM is trained using the variational learning method. The advantage of the variational approach is that the final number of Gaussian components is generally smaller than the initial ones. This is due to the fact that the VB method makes the model selection and parameter learning at the same time. The proposed VB-GMM-UBM model gives 91.18% classification accuracy. The performance is significantly higher compared to 67.30% in GMMs, 80.36% in GMM-UBM and 83.94% in VB-GMMs.

We study the effect of adapting different sets of parameters during generation of the audio class models. Table 2 shows the performance of GMM-UBM and VB-GMM-UBM on adaptation of different sets of Gaussian parameters. It is seen that the mean adaptation plays a more significant role in the GMM-UBM

Table 3. Confusion matrix for VB-GMM-UBM

Audio Class	**Classified Audio Class**				
	Ads	**Cartoon**	**Cricket**	**Football**	**News**
Ads	78.68	14.75	1.63	0.0	4.91
Cartoon	2.53	96.20	0.0	0.0	1.26
Cricket	1.66	3.33	93.33	1.66	0.0
Football	1.63	1.63	0.0	96.72	0.0
News	3.84	5.12	0.0	0.0	91.02

than the weight or variance adaptation. In VB-GMM-UBM, each individual parameter adaptation (weight, mean, and variances) plays an important role in adaptation process.

The confusion matrix for the VB-GMM-UBM that gives the best performance is shown in Table 3. It is seen that the Ads category is mainly confused with the Cartoon category and the News category.

4 Summary

This work studies the behavior of different types of Gaussian mixture model based classification models. Conventional GMMs, GMM-UBM, VB-GMMs and the proposed method VB-GMM-UBM are studied. The GMMs and GMM-UBM suffer from the overfitting problem, when the model complexity is high. In such cases, the VB-GMMs perform better. The VB-GMMs prune out the redundant components. The VB-GMMs are also free from singularity problems that arise frequently in GMMs and GMM-UBM. We could improve its performance by applying the adaptation process in VB-GMMs. Results of our studies on audio clip classification indicate that the proposed VB-GMM-UBM model performs better than the other models.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
2. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19–41 (2000)
3. Nasios, N., Bors, A.: Variational learning for Gaussian mixture model. IEEE Trans. System, Man, and Cybernetics 36, 849–862 (2006)
4. Zheng, R., Ulang, S., Xu, B.: Text-independent speaker identification using GMM-UBM and frame level likelihood normalization. In: Proc. ISCSLP, pp. 289–292 (2004)
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2000)
6. Gauvain, J.L., Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech and Audio Processing 2, 291–298 (1994)
7. Aggarwal, G., Bajpai, A., Khan, A.N., Yegnanarayana, B.: Exploring features for audio indexing. Inter-Research Institute Student Seminar, IISc Bangalore (March 2002)

Cross-Lingual Vocal Emotion Recognition in Five Native Languages of Assam Using Eigenvalue Decomposition

Aditya Bihar Kandali[1], Aurobinda Routray[1], and Tapan Kumar Basu[2]

[1] Department of Electrical Engineering, Indian Institute of Technology Kharagpur, PIN Code-721302, India
abkandali@rediffmail.com, aroutray@ee.iitkgp.ac.in
[2] Aliah University, Salt Lake City, Kolkota, India
basutk02@yahoo.co.in

Abstract. This work investigates whether vocal emotion expressions of full-blown discrete emotions can be recognized cross-lingually. This study will enable us to get more information regarding nature and function of emotion. Furthermore, this work will help in developing a generalized vocal emotion recognition system, which will increase the efficiency required for human-machine interaction systems. An emotional speech database was created with 140 simulated utterances (20 per emotion) per speaker, consisting of short sentences of six full-blown discrete basic emotions and one 'no-emotion' (i.e. neutral) in five native languages (not dialects) of Assam. A new feature set is proposed based on Eigenvalues of Autocorrelation Matrix (EVAM) of each frame of utterance. The Gaussian Mixture Model is used as classifier. The performance of EVAM feature set is compared at two sampling frequencies (44.1 kHz and 8.1 kHz) and with additive white noise with signal-to-noise ratios of 0 db, 5 db, 10 db and 20 db.

Keywords: Full-blown Basic Emotion, Cross-lingual Vocal Emotion Recognition, Gaussian Mixture Model, Eigenvalues of Autocorrelation Matrix.

1 Introduction

Human beings express emotions explicitly in speech, face, gait and other body languages. The vocal expressions are harder to regulate than other explicit emotional signals. So, it is possible to know the actual affective state of the speaker from her/his voice without any physical contact. But exact identification of emotion from voice is very difficult due to several factors. The speech consists broadly of two components coded simultaneously: (i) "What is said" and (ii) "How it is said". The first component consists of the linguistic information pronounced as per the sounds of the language. The second component consists of non-linguistic or paralinguistic or supra-segmental component which includes the prosody of the language i.e. pitch, intensity and speaking-rate rules to give lexical and grammatical emphasis for the spoken messages; and the prosody of emotion to express

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 519–524, 2009.

the affective state of the speaker. In addition, speakers also possess their own style, i.e. a characteristic articulation rate, intonation habit and loudness characteristic. The voice contains also information about the speaker's identity, age, gender, and body size.

The present work investigates a specific research question concerning vocal emotion recognition: "Are vocal emotion expressions of discrete emotions recognized cross-lingually?". This study will enable one to get more information about the nature and function of emotion. It will also help in developing a generalized voice emotion recognition system, which will increase the efficiency of human-machine interaction systems. Some applications are as follows: (i) to obtain more efficient and more accurate performance of automatic speech recognition and automatic speaker recognition systems, due to reduction of search space to models corresponding to pre-recognized emotions [1, 2]; (ii) to design an automatic speech translator across languages retaining the emotional content [1], (iii) to make more efficient automatic tutoring, alerting, and entertainment systems [3]. Picard [4] has explained about affective computing algorithms which can improve problem solving capability of a computer and make it more intelligent by giving it the ability to recognize and express emotions.

When a machine is trained with emotion utterances of one set of languages and tested with emotion utterances of a set of different languages, the process is called as cross-lingual (or cross-cultural) vocal emotion recognition. Very few studies of cross-lingual (i.e. cross-cultural) voice emotion recognition have been reported by researchers [5]. Among these noteworthy is the study by Scherer et al. [6], conducted in nine countries in Europe, the United States, and Asia on vocal emotion portrayals of anger, sadness, fear, joy, and neutral voice, which are produced by professional German actors. In this study, overall perception accuracy by human subjects is found to be 66%. Also, the patterns of confusion are found very similar across all countries, which suggest the existence of similar inference rules from vocal expression across cultures. Generally, accuracy decreases with increasing language dissimilarity from German in spite of the use of language-free utterances. So their conclusion is that culture- and language-specific paralinguistic patterns may influence the decoding process. Juslin and Laukka [5] also reported that cross-cultural decoding accuracy of voice expression of emotions is significantly higher than that expected by chance. Laukka [7] has reported that: (i) vocal expressions of discrete emotions are universally recognized, (ii) distinct patterns of voice cues correspond to discrete emotions, and (iii) vocal expressions are perceived as discrete emotion categories but not as broad emotion dimensions. All the above experiments are done mostly with a very few number of European and Asian languages. So, these findings need to be verified using more number of languages.

The present study is based on a modified Brunswikian lens model of process of vocal communication of emotion [8]. This model motivates research to determine the proximal cues i.e. the representation of voice acoustic cues in the basilar membrane of the cochlea, amygdala, and auditory cortex, which will lead to the perception of the vocal emotion. Based on studies by researchers [3, 8, 9], one can identify three broad types of proximal voice cues: (i) fundamental frequency or pitch frequency (F0) contour, (ii) continuous acoustic variables: magnitude of fundamental frequency, intensity, speaking rate, and spectral energy distribution; and (iii) voice quality (tense, harsh or breathy): described by high frequency energy, formant frequencies, precision articulation and glottal waveform. A description of relationships among archetypal emotions and the voice cues is given in [3, 8, 9].

In this paper, a new feature set is proposed based on 5 most significant Eigenvalues of Autocorrelation Matrix (EVAM) of each frame of utterance for automatic vocal emotion recognition. The 5 most significant EVAM of a signal represent the powers of 5 most prominent frequency components (though with some additive noise) in the signal [10]. The source-filter model of speech production describes speech as an acoustic excitation signal filtered due to resonances of the vocal tract. The vocal tract resonances are called formant frequencies, or formants; which are the prominent frequencies having relatively higher amplitudes than other frequency components in the speech signal. In general, a speech signal contains 5 to 6 formants. Hence, 5 or 6 most significant EVAM of a short-time frame of speech will represent the powers corresponding to the formants (though with some additive noise), if they are present in that speech frame. The EVAM feature set is also expected to be robust in presence of noise, since these eigenvalues corresponds to the most prominent signal subspace eigenvectors.

The Gaussian mixture model (GMM) classifier is used for classification [11]. The study of cross-lingual vocal emotion recognition is carried out using simulated utterances of 6 full-blown discrete basic emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*) and 1 '$no-emotion$' (i.e. *neutral*) in 5 Indian languages: Assamese, Bodo (or Boro), Dimasa, Karbi, and Mishing (or Mising), which are the native languages (not dialects) of the state of Assam. The performance of the EVAM feature set is compared at two sampling frequencies (44.1 kHz and 8.1 kHz) and with additive white noise with signal-to-noise ratios of 0 db, 5 db, 10 db and 20 db.

2 Data Collection

As a part of this research work, emotional utterances in native languages of Assam are collected as described below. The subjects are chosen mostly from students and faculty members of some educational institutions of Assam. Some subjects are lay actors and others are trained for the first time through a few rehearsals, so as to avoid exaggerated expressions. Thirty randomly selected volunteer subjects (3 males and 3 females per language) are requested for recording emotional utterances of the 5 native languages of Assam. The utterances are recorded in an almost noise-free small closed room with headphone-mic and notebook computer in a single channel with 44.1 kHz sampling frequency and 16 bit depth. Each subject is asked to utter a fixed set of 140 short sentences (20 per emotion) of variable length of her/his first language only. The subjects are asked to rehearse their acting a few times before final recording.

3 Listening Test

A listening test of the emotional speeches is carried out with the help of 6 randomly selected volunteer listeners (3 Males and 3 Females) for each language of the Multilingual ESDNEI database. The listeners have never heard the speeches of the languages of the Multilingual ESDNEI database. Some of the listeners are selected as different persons for each language while others remained the same, because of the unavailability of a complete set of different volunteer listeners for each language, who do not understand or never heard speeches in the above languages. The average scores of the listening test are given in Table 1.

Table 1. Percentage Cross-lingual Average Recognition Success Scores of Listening Test of the Utterances of the individual languages of the Multilingual ESDNEI database by 6 Human subjects (3 Males and 3 Females) who never heard any of these languages

Language→ Emotion↓	Assamese	Bodo	Dimasa	Karbi	Mishing	Average
Anger	97.22	70.97	95.42	88.19	86.67	87.69
Disgust	95.00	45.83	85.97	75.69	75.42	75.58
Fear	97.78	85.28	95.56	87.08	93.33	91.81
Happiness	85.69	74.86	84.17	78.75	79.44	80.58
Sadness	97.50	88.19	97.08	96.25	94.58	94.72
Surprise	87.78	67.36	78.06	69.58	71.81	74.92
Neutral	99.03	81.81	94.17	93.61	90.97	91.92
Average	94.29	73.47	90.06	84.17	84.60	85.32

4 Experiment: Cross-Lingual Vocal Emotion Recognition

A total of seven GMMs, one for each emotion, are trained using the Expectation-Maximization (EM) algorithm [11] and Leave-One-Out (LOO) cross-validation method [12], and EVAM feature vectors of 10 utterances of the subjects of 4 languages. After training, the classifier is tested with EVAM feature vectors of test utterances consisting of other 10 utterances of the subjects of the left-out language (test-language) as follows. The mean-log-likelihood of EVAM feature vectors of one test-utterance with respect to the trained GMM corresponding to each emotion-class is computed. The test-utterance is considered to belong to that emotion-class with respect to which the mean log-likelihood becomes the largest. The Percentage Average Recognition Success Score (PARSS) of each emotion and the Mean-PARSS (MPARSS) of all emotions are computed from the Recognition Success Scores (RSS) for all 5 combinations of train-test data. The initial means and the elements of diagonal covariance matrices of the GMM are computed by split-Vector Quantization algorithm [13]. The above procedure is repeated for GMMs with different number of components of Gaussian probability distribution functions i.e. M=8, 16 and 32, and the best result is considered. Henceafter, the Mean-PARSS will be referred as the 'Average'.

5 Feature Extraction

In this paper, the speech is preprocessed by detecting the end points and removing the silence periods and the dc component. The frame duration is chosen as 23.22 ms in case of sampling frequency of 44.1 kHz. The utterances are decimated to sampling frequency 8.1 kHz and in this case the frame duration is chosen as 31.6 ms. All the frames are rectangular windowed. The frames are taken with 50% overlaps with neighboring frames. For each frame, the autocorrelation matrix with lag p=32 in case of sampling frequency of 44.1 kHz and lag p=8 in case of sampling frequency of 8.1 kHz are computed. Then after eigen decomposition of the autocorrelation matrix, a 5-element feature vector is formed using the 5 most significant eigenvalues, from each frame. The EVAM features are normalized by subtracting mean and dividing by the standard deviation.

6 Results and Discussion

The percentage average scores of cross-lingual voice emotion recognition in each case are shown in Table 2. It can be observed that the performance of the proposed feature set for the case of original utterances at 8.1 kHz sampling frequency is a little better than at 44.1 kHz sampling frequency. The performance gradually decreases as the Signal-to-Noise Ratio (SNR) is reduced from 20 db to 0 db. It is observed that the performance is satisfactorily above that of the average human recognition score (i.e. 85.32% in Table 1) in the listening test. The results show high potential of EVAM features for emotion recognition from telephone channel voices which have 8 kHz sampling frequency.

Table 2. Percentage Average Score of Cross-lingual Vocal Emotion Recognition from original speech with Noise-not-Added (NNA) and in presence of additive white noise of 4 Signal-to-Noise Ratios (SNRs) [fs: Sampling Frequency, Number of Components in GMM: 32]

SNR (db)→	NNA	NNA	20	10	5	0
fs (kHz)→ Emotion↓	44.1	8.1	8.1	8.1	8.1	8.1
Anger	100.00	100.00	100.00	97.33	95.33	84.33
Disgust	99.67	99.00	97.33	91.67	83.67	74.33
Fear	99.67	100.00	100.00	99.33	98.33	94.33
Happiness	100.00	100.00	100.00	99.67	98.00	96.00
Sadness	98.67	99.67	99.33	98.00	97.00	87.00
Surprise	100.00	100.00	99.33	95.33	91.00	85.67
Neutral	99.00	99.00	98.33	97.67	95.67	93.67
Average	99.57	99.67	99.19	97.00	94.14	87.90

7 Conclusion

This study verified that the full-blown discrete basic vocal emotions are recognized cross-lingually with accuracies much above the chance level. It is also verified that there exist distinct patterns of voice cues corresponding to full-blown discrete basic emotions. The EVAM features have high potential for vocal emotion recognition in telephone channel.

Acknowledgment

The authors are grateful to the students and the faculty members of Jorhat Engineering College, Jorhat, Assam, and all the other people of Assam, who have actively cooperated during the collection of data for the present work. The authors also express gratefulness to all the students of Indian Institute of Technology Kharagpur, India, who have voluntarily taken part as listeners in the listening test.

References

[1] Holmes, J., Holmes, W.: Speech Synthesis and Recognition, 2nd edn. Taylor & Francis, New York (2001)

[2] Rose, P.: Forensic Speaker Identification, p. 302. Taylor & Francis, New York (2002)

[3] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. 18(1), 32–80 (2001)

[4] Picard, R.W.: Affective Computing. The MIT Press, Cambridge (1997)

[5] Juslin, P.N., Laukka, P.: Communication of Emotions in Vocal Expression and Music Performance. Psychological Bulletin 129(5), 770–814 (2003)

[6] Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. J. Cross-Cultural Psychology 32(1), 76–92 (2001)

[7] Laukka, P.: Vocal Expression of Emotion – Discrete-emotion and Dimensional Accounts. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 141, ACTA Universitatis Upsaliensis, Uppsala (2004)

[8] Scherer, K.R., Johnstone, T., Klasmeyer, G.: Vocal Expression of Emotion. In: Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (eds.) Handbook of Affective Science, Part IV, ch. 23, 1st edn. Oxford University Press, Oxford (2003)

[9] Ekman, P.: Basic Emotions. In: Dalgleish, T., Power, M. (eds.) Handbook of Cognition and Emotion, ch. 3. John Wiley & Sons, Ltd., Sussex (1999)

[10] Marple Jr., S.L.: Digital Spectral Analysis With Applications. Prentice Hall Inc., Englewood Cliffs (1987)

[11] Reynolds, D.A., Rose, R.C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech Audio Process. 3(1), 72–83 (1995)

[12] Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Morgan Kaufmann, Academic Press, New York (1990)

[13] Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications 28(1), 84–95 (1980)

Variable Length Teager Energy Based Mel Cepstral Features for Identification of Twins

Hemant A. Patil[1] and Keshab K. Parhi[2]

[1] Dhirubhai Ambani Institute of Information and Communication Technology, DA-IICT
Gandhinagar, India-382 007
[2] Department of Electrical and Computer Engineering, University of Minnesota,
Minneapolis, MN, USA- 55455
hemant_patil@daiict.ac.in, parhi@umn.edu

Abstract. An important issue which must be addressed for the speaker recognition system is how well the system resists the effects of determined mimics such as those based on physiological characteristics especially twins. In this paper, a new feature set based on recently proposed Variable Length Teager Energy Operator (VTEO) and state-of-the-art Mel frequency cepstral coefficients (MFCC) is developed. The effectiveness of the newly derived feature set in identifying twins in Marathi language has been demonstrated. Polynomial classifiers of *2nd* and *3rd* order have been used. The results have been compared with other spectral feature sets such as Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC) and baseline MFCC.

1 Introduction

Speaker recognition refers to the task of identifying a person from his or her voice with the help of machines. Fig.1 and Fig.2 show two speech signals (and their corresponding spectrograms, formant contours, pitch contours) produced by two male twin speakers. It is evident from the plots that the pattern of speech signals, pitch contours, formant contours and spectrograms for each twin speakers are very similar, if not identical. This makes speaker recognition of twins a challenging research problem. Similar studies were reported in [4]. Identification of professional mimics is also challenging task and is attempted in [7]. In this paper, we show the effectiveness of the newly derived feature set by amalgamating Variable Length Teager Energy Operator (VTEO) and Mel frequency cepstral coefficients (MFCC) for identification of twins.

2 Variable Length Teager Energy Operator (VTEO)

According to Teager [8], the airflow do not propagate in the vocal tract as a linear planar wave, but rather as separate and concomitant vortices are distributed throughout the vocal tract during phonation. He suggested the true source of sound production is actually the vortex-flow interactions, which are non-linear and a non- linear

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 525–530, 2009.

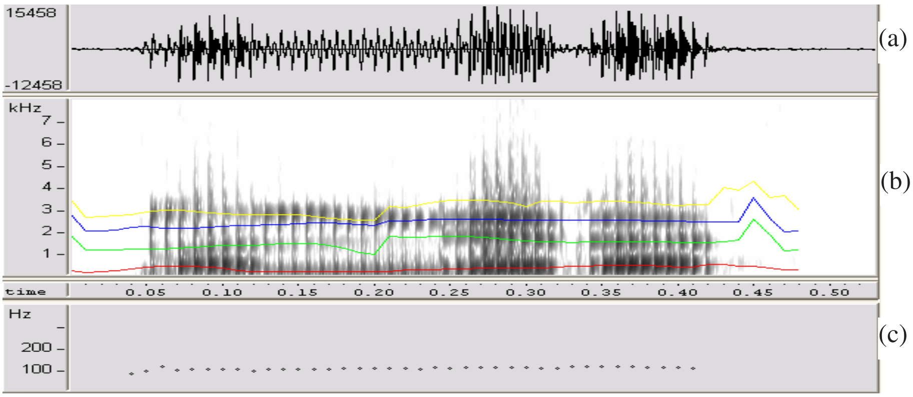


Fig. 1. (a) Speech signal, *viz.*, '*Mandirat*' by a twin male speaker A, (b) Spectrogram and formant contour of signal shown in (a), (c) Pitch contour of signal shown in (a)

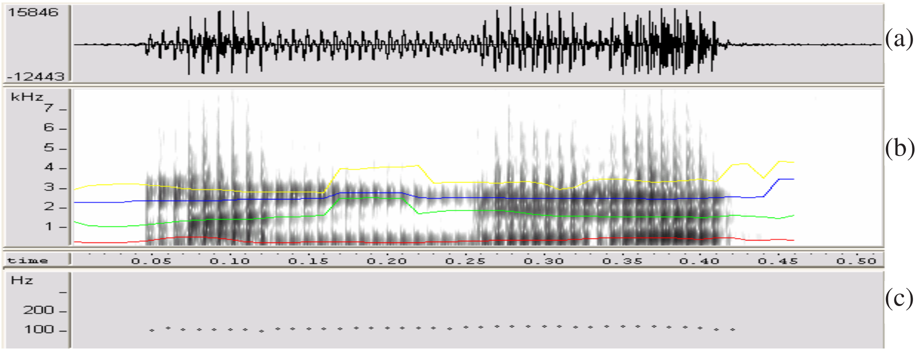


Fig. 2. (a) Speech signal, *viz.*, '*Mandirat*' by a twin male speaker B, (b) Spectrogram and formant contour of signal shown in (a), (c) Pitch contour of signal shown in (a)

model has been suggested based on the energy of airflow. Modeling the time-varying vortex flow is a formidable task and Teager devised a simple algorithm which uses a non-linear energy-tracking operator called as Teager Energy Operator (TEO) (in discrete-time) for signal analysis with the supporting observation that hearing is the process of detecting energy. The concept was further extended to continuous-domain by Kaiser [3]. According to Kaiser, energy in a speech frame is a function of amplitude and frequency as well and by using the dynamics of S.H.M. he developed TEO for discrete-time signal, as

$$TEO\left\{x(n)\right\} = x^2(n) - x(n+1)x(n-1) \qquad (1)$$

From (1) it is evident TEO of a signal involves non-linear operations (e.g. square) on the signal. TEO algorithm gives good running estimate of the signal energy when signal has sharp transitions in time-domain. However, in situations when the amplitude difference between two consecutive samples of signal is minute, then the TEO will give zero energy output which indicates that energy required to generate such sequence of samples is zero but that may not be the case in actual physical signal (e.g. speech). To alleviate this problem VTEO was proposed recently [9] and hence very briefly it is discussed here. The dynamics and solution (which is a S.H.M.) of mass-spring system are described and the energy is given by

$$\frac{d^2x}{dt^2}+\frac{k}{m}x=0 \Rightarrow x(t)=A\cos(\Omega t+\phi) \tag{2}$$

$$E=\frac{1}{2}m\Omega^2A^2 \Rightarrow E \propto \left(A\Omega\right)^2$$

From (2), it is clear that the energy of the S.H.M. of displacement signal $x(t)$ is directly proportional not only to the square of the amplitude of the signal but also to the square of the frequency of the signal. Kaiser and Teager proposed the algorithm to calculate the running estimate of the energy content in the signal. Now $x(t)$ can be expressed in discrete-time domain as

$$x(n)=A\cos(\omega n+\phi)$$

where A, ω and ϕ are the amplitude, digital frequency (rad/sec) and phase of sinusoidal signal of S.H.M. We have

$$\begin{aligned}
&x\left(n\pm i\right)=A\cos\left(\omega\left(n\pm i\right)+\phi\right); \qquad i>n\\
&\Rightarrow VTEO_i\left\{x\left(n\right)\right\}=x^2(n)-x(n+i)x(n-i)=A^2\sin^2\left(i\omega\right)\approx A^2\omega^2\approx E_n;\\
&\Rightarrow \quad VTEO_i \rightarrow TEO \quad as \quad i\rightarrow 1
\end{aligned} \tag{3}$$

where E_n gives the running estimate of signal's energy and we refer $VTEO_i\left\{x\left(n\right)\right\}$ variable length TEO for the dependency index i which is expected to give running estimate of signal's energy after considering i^{th} past and i^{th} future sample to track the *dependency* in the sequence of samples of speech signal. It should be noted that (3) is exact and unique if $i\omega$ is restricted to values less than $\pi/2$, the equivalent of one-fourth of the sampling frequency. In addition, if we limit $i\omega<\pi/4$, then the relative error is always below 11% [3], [9].

3 Variable Length TEO Based MFCC (VT-MFCC)

Traditional MFCC based feature extraction involves pre-processing; Mel-spectrum of pre-processed speech, followed by log-compression of subband energies, and finally DCT to get MFCC per frame [2]. In our approach, we employ VTEO for calculating the energy of speech signal. Now, one may apply VTEO in frequency domain, i.e., VTEO of each subband at the output of Mel-filterbank, but there is difficulty from implementation point of view as discussed in [6]. Let us now see the computational details of VT-MFCC.

Speech signal $x(n)$ is first passed through pre-processing stage (which includes frame blocking, Hamming windowing and pre-emphasis) to give pre-processed speech signal $x_p(n)$. Next we calculate the VTEO of $x_p(n)$:

$$VTEO_i[x_p(n)]={x_p}^2(n)-x_p(n+i)x_p(n-i)=\psi(n) \qquad (say)$$

The magnitude spectrum of the VTEO output is computed and warped to Mel frequency scale followed by usual log and DCT computation (of MFCC) to obtain VT-MFCC as:

$$VT-MFCC=\sum_{l=1}^{L}\log\left[\Psi(l)\right]\cos\left(\frac{k(l-0.5)}{L}\pi\right),k=1,2,....,Nc.$$

where $\Psi(l)$ is the filterbank output of $DFT\{\psi(n)\}$ and $\log[\Psi(l)]$ is the log of filterbank output and $VT-MFCC(k)$ is the k^{th} VT-MFCC. Proposed feature set, *viz.*, VT-MFCC differ from the traditional MFCC in the definition of *energy measure*, i.e., MFCC employs L^2 energy in frequency domain (due to Parseval's equivalence) at each subband whereas VT-MFCC employs variable length Teager energy in time domain. Fig. 3 shows the functional block diagram of MFCC and VT-MFCC. For DI=1(i.e., TEO), Maragos *et al.* [5] has proved that we can use this operator for formant estimation and hence VTMFCC may capture formant information in some sense. In addition, VTEO profile may capture speaker-specific airflow properties (along with adjacent sample dependencies) in the vocal tract of twin speakers.

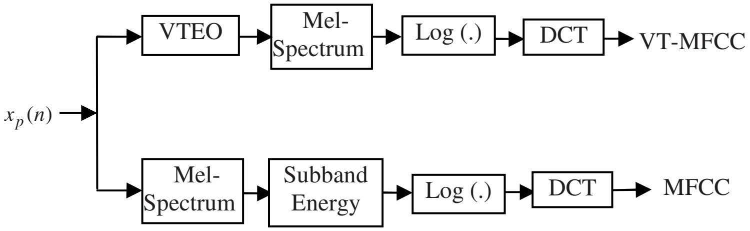


Fig. 3. Block diagram for VT-MFCC and MFCC implementation

4 Experimental Results

Database of 17 twins is prepared from different dialectal zones of Maharashtra State, India. Text material for recording consisted of isolated words, digits, combination-lock phrases and a contextual speech of considerable duration. Polynomial classifiers of *2*nd and *3*rd order approximation (based on discriminative training with MSE criterion) are used as the basis for all experiments. The testing feature vectors are processed by the polynomial discriminant function. Every speaker has speaker specific vector to be determined during machine learning and the output of a discriminant function is averaged over time resulting in a score for each speaker [1]. Feature analysis was performed using *12*th order LPC on a 23.2 ms mean subtracted frame with an overlap of 50%. Each frame was pre-emphasized with the filter 1-0.97z^{-1}, followed by Hamming window (similar pre-processing steps were performed for MFCC and VT-MFCC except mean removal). We have taken 2DI (i.e., dependency index) samples more to compute VT-MFCC than that for LPC, LPCC and MFCC because of VTEO processing. The results are shown as success rates (ratio of the number of correctly identified speakers to the total number of speakers used for machine learning) in Table 1 for different testing durations and the last row gives average success rate for 2^{nd} and 3^{rd} order polynomial approximation, respectively. Some of the observations from the results are as follows:

- For 2^{nd} order polynomial approximation, average success rates for VT-MFCC is found to be better than T-MFCC, MFCC, LPCC and LPC in majority of the cases.
- Average success rates for VT-MFCC, T-MFCC, MFCC, LPCC and LPC go up for 3^{rd} order polynomial approximation as compared to their 2^{nd} order counterparts. This may be due to the fact that *feature occupancy* decreases (and thus class separability increases) in high-dimensional feature space, i.e., for 3^{rd} order classifier.
- For 3^{rd} order polynomial approximation, VT-MFCC and T-MFCC performed equally well and these feature sets together outperformed LPC and LPCC. In this case, MFCC performed better than VT-MFCC, T-MFCC, LPCC, and LPC.
- It is interesting to note that in majority of the cases of misidentification, the misidentified person is the actual speaker's twin brother or sister (except in those cases where the twin pair is of different gender). This is evident by the sub-optimal success rates (which considers a correct identification if twin speaker is misidentified with his or her twin brother or sister) shown in brackets in Table 1.
- It is evident from the Table 1 that sub-optimal success rates are higher for LPC, LPCC and MFCC than proposed VT-MFCC in majority of the cases. This means that the confusion in capturing speaker-specific information is more for LPC, LPCC and MFCC (because they directly reflect in some sense physiological characteristics of twins).
- Fig. 4 shows VTEO for DI=2 and DI=1 (i.e., T-MFCC [6]) of pre-processed speech signals of Fig.1-2. It is evident that dependency index does play a role in capturing sample dependency in the airflow and hence perhaps speaker-specific information, whereas MFCC may extract similarity in *spectral* features of twins (Fig.1-2).

Table 1. Success Rates (%) for 2^{nd} Order Polynomial Approximation with 60 s Training Speech

TEST (SEC)	VT-MFCC (DI=2)	T-MFCC (DI=1)	MFCC	LPCC	LPC
1	76.47 (88.23)	52.94 (70.58)	73.52 (88.23)	64.70 (79.41)	67.64 (79.41)
3	76.47 (88.23	70.58 (85.29)	76.47 (88.23)	67.64 (82.35)	67.64 (82.35)
5	79.41 (91.17)	70.58 (85.29)	79.41 (94.11)	64.70 (85.29)	67.64 (88.23)
7	85.29 (94.11)	73.52 (85.29)	79.41 (94.11)	70.58 (85.29)	70.58 (88.23)
10	85.29 (94.11)	76.47 (85.29)	76.47 (94.11)	82.35 (94.11)	70.58 (88.23)
12	82.35 (91.17)	79.41 (85.29)	79.41 (94.11)	79.41 (94.11)	73.52 (91.17)
15	82.35 (91.17)	73.52 (85.29)	79.41 (94.11)	79.41 (94.11)	73.52 (88.23)
Order *2*	**81.09** (91.17)	71.00 (83.19)	77.73 (91.59)	72.68 (87.81)	70.16 (86.13)
Order *3*	84.87 (**95.37**)	86.13 (92.85)	**88.23** (**97.47**)	82.35 (93.27)	78.99 (93.69)

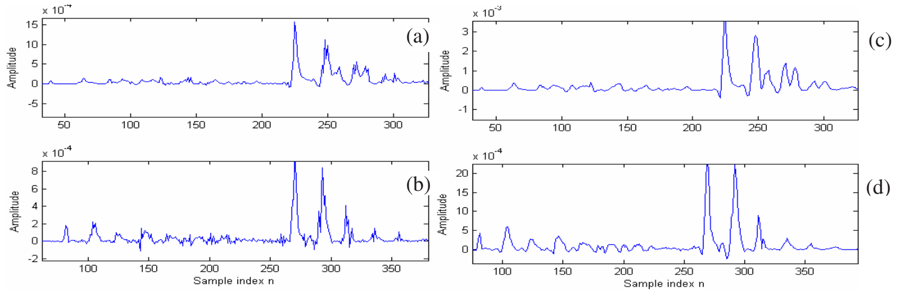


Fig. 4. (a)-(b) TEO (i.e., DI=1) of signal shown in Fig. 2 and Fig.1, respectively for *35th* frame. (c)-(d) VTEO of signal shown in Fig. 2 and Fig.1, respectively for *35th* frame of 516 samples (50% overlap) with DI=2.

5 Conclusion

In this paper, a new feature set, *viz.*, Variable Length Teager Energy based MFCC (VT-MFCC) is proposed for speaker identification of twins. Their performance was compared with conventional features and found to be effective. In this work, the variable length TEO was referred for variation in DI across experiments. Future work includes the investigation on optimal DI in VT-MFCC for speaker recognition. In addition to this, the present work could be extended to investigate choice of DI from segment to segment (e.g. vowel, transitions, etc).

References

1. Campbell, W.M., Assaleh, K.T., Broun, C.C.: Speaker recognition with polynomial classifiers. IEEE Trans. on Speech and Audio Processing 10, 205–212 (2002)
2. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech and Signal Processing 28, 357–366 (1980)
3. Kaiser, J.F.: On a simple algorithm to calculate the 'energy' of a signal. In: Proc. of Int. Conf. on Acoustic, Speech and Signal Processing, vol. 1, pp. 381–384 (1990)
4. Kersta, L.G., Colangelo, J.A.: Spectrographic speech patterns of identical twins. J. Acoust Soc. Amer. 47(1), 58–59 (1970)
5. Maragos, P., Quatieri, T., Kaiser, J.F.: Speech non-linearities, modulation and energy Operators. In: Proc. Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP 1991, Toronto, ON, Canada, pp. 421–424 (1991)
6. Patil, H.A., Basu, T.K.: The Teager energy based features for identification of identical twins in multi-lingual environment. In: Pal, N.R., Kasabov, N., Mudi, R.K., Pal, S., Parui, S.K. (eds.) ICONIP 2004. LNCS, vol. 3316, pp. 333–337. Springer, Heidelberg (2004)
7. Patil, H.A., Basu, T.K.: LP spectra vs. Mel spectra for identification of professional mimics in Indian languages. Int. J. Speech Tech. 11(1), 1–16 (2008)
8. Teager, H.M.: Some observations on oral air flow during phonation. IEEE Trans. Acoust., Speech, Signal Process. 28, 599–601 (1980)
9. Tomar, V., Patil, H.A.: On the development of variable length Teager energy operator (VTEO). In: Interspeech, Brisbane, Australia, pp. 1056–1059 (2008)

Unit Selection Using Linguistic, Prosodic and Spectral Distance for Developing Text-to-Speech System in Hindi

K. Sreenivasa Rao, Sudhamay Maity, Amol Taru, and Shashidhar G. Koolagudi

School of Information Technology
Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India
ksrao@iitkgp.ac.in, friendsudha@gmail.com, amol.taru@gmail.com,
koolagudi@yahoo.com

Abstract. In this paper we propose a new method for unit selection in developing text-to-speech (TTS) system for Hindi. In the proposed method, syllables are used as basic units for concatenation. Linguistic, positional and contextual features derived from the input text are used at the first level in the unit selection process. The unit selection process is further refined by incorporating the prosodic and spectral characteristics at the utterance and syllable levels. The speech corpora considered for this task is the broadcast Hindi news read by a male speaker. Synthesized speech from the developed TTS system using multi-level unit selection criterion is evaluated using listening tests. From the evaluation results, it is observed that the synthesized speech quality has improved by refining the unit selection process using spectral and prosodic features.

Keywords: Text-to-speech, unit selection, linguistic features, prosodic features and spectral features.

1 Introduction

In the concatenative speech synthesis approach, the speech is generated by concatenating the segments of natural speech waveforms corresponding to the sequence of sound units that are derived from the input text [1]. Earlier, the concatenative speech synthesis is performed by concatenating the sequence of sound units (phones or diphones or syllables), where the unique versions of the sound units are stored in the database. After the concatenation of basic sound units, the prosodic information will be incorporated using appropriate signal processing techniques. This method introduces distortion due to the manipulation of sound units by signal processing techniques. To overcome this distortion, corpus based (data driven) concatenation approach is proposed. In this approach, the database consists of huge labeled speech corpus, having the multiple replicas of the basic sound units. Since the database has multiple candidates for each sound unit, there should be a mechanism to choose the sequence of sound units in an accurate way, such that it requires minimal signal manipulation.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 531–536, 2009.

In this paper we are proposing multilevel unit selection criterion for choosing the sequence of units from the speech corpus. At the first level the unit selection is performed using the linguistic, positional and contextual features derived from the text. In the second level the unit selection is performed on the units selected in the first level using spectral and prosodic features separately. At the final level (3rd level), the unit selection is performed on the units selected at the second level, by combining spectral and prosodic features together.

There are some earlier attempts in the research of developing TTS systems for Indian languages. At IIT Madras, TTS system for Hindi was developed in early 90's using parametric approach [2]. N. Sridhar Krishna *et al.*, have proposed duration and prosodic phrasing models for developing the TTS system in Telugu [3]. Samuel Thomas *et al.*, have developed natural sounding TTS system in Tamil using syllable like units [4]. S. P. Kishore *et al.*, have proposed data-driven speech synthesis approach using syllables as basic sound units for developing TTS in Hindi [5]. At TIFR Mumbai, TTS system for Indian accent English was developed for browsing the web [6]. Sreekanth *et al.*, have developed festival based TTS system for Tamil [7]. Speech synthesizers in Hindi and Bengali were developed at IIT Kharagpur for visually challenged people [8].

The paper is organized as follows: The details of the development of speech corpora are discussed in section 2. In section 3, description of the base line TTS system using the proposed unit selection approach is provided. The proposed approach of unit selection process and performance of the TTS system by incorporating the proposed unit selection criterion is analyzed using listening tests are discussed in section 4. In the final section the summary of the paper is given along with future work that can improve the performance of the system further.

2 Speech Corpus

Speech corpus used in this work consists of Hindi broadcast news data read by male speaker. Duration of speech corpus is about 1 hour. The speech signal was sampled at 16 kHz, and each sample is represented as a 16 bit number. database is organized at sentence, word and syllable levels. Each of the syllables is labeled by 24 features representing the linguistic context and production constraints [9]. These features represent positional, contextual and phonological information of the syllable. The list of the features representing the syllable is given in the Table 1. Along with the linguistic features derived from the text, the syllables are also labeled with the prosodic (i.e., pitch, duration and energy) and spectral information.

3 Text-to-Speech System

Block diagram of the TTS system using the proposed unit selection approach is shown in Fig. 1. Text normalization module converts abbreviations, numbers etc., into the spoken equivalent text. Syllabification module derives the sequence of syllables from the normalized text. Syllabification module first identifies vowels, and

Table 1. List of the factors and features representing the linguistic context and production constraints of the syllable

Factors	Features
Syllable position in the phrase	Position of syllable from beginning of the phrase Position of syllable from end of the phrase Number of syllables in the phrase
Syllable position in the word	Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables in the word
Word position in the phrase	Position of word from beginning of the phrase Position of word from end of the phrase Number of words in the phrase
Syllable identity	Segments of the syllable (consonants and vowels)
Context of the syllable	Identity of the previous syllable Identity of the following syllable
Syllable nucleus	Position of the nucleus Number of segments before the nucleus Number of segments after the nucleus

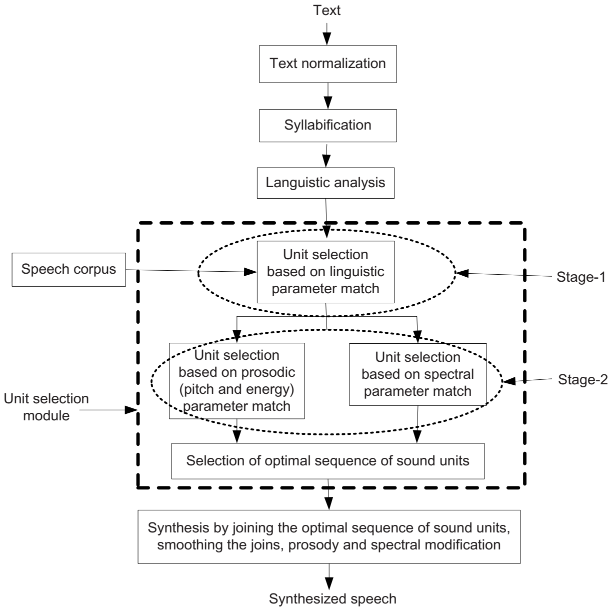

Text
Text normalization
Syllabification
Languistic analysis
Speech corpus
Unit selection based on linguistic parameter match
Stage-1
Unit selection based on prosodic (pitch and energy) parameter match
Unit selection based on spectral parameter match
Stage-2
Unit selection module
Selection of optimal sequence of sound units
Synthesis by joining the optimal sequence of sound units, smoothing the joins, prosody and spectral modification
Synthesized speech

Fig. 1. Block diagram of the TTS system using the proposed unit selection approach

then determines number of syllables present in text. Each syllable is derived by associating consonants to the appropriate vowel using linguistic rules. Linguistic analysis module derives 24 features (see Table 1) for each syllable. The first stage of the unit selection module derives the multiple realizations of the given sound units, which satisfy the matching criterion based on 24 features, derived from linguistic analysis module. These sound units are further fed to spectral and prosodic parameter analysis modules. The second stage of the unit selection module (i.e. spectral and prosodic analysis module) derives the unique sequence of sound units from its multiple realizations by minimizing the spectral and prosodic distances, between the adjacent units, to minimize the join cost. The final output sequence of sound units given by unit selection module is concatenated by taking care of smoothening the joints between successive units. After simple concatenation, prosodic and spectral manipulations are carried out according to the predicted prosody from the models. Finally, the speech synthesized after prosodic manipulation is the desired speech for the given input text.

4 Proposed Unit Selection Criterion

The basic goal of unit selection module is to derive the optimal sequence of sound units from the speech corpus by minimizing the cost function. Here the cost function may be derived from the three components (1) Linguistic match, (2) Spectral match and (3) Prosodic match. In this work we have proposed two stage unit selection criterion. At the first stage, for the derived sequence of sound units, multiple realizations are chosen using linguistic match criterion. The linguistic match is carried out by matching the 24 dimensional feature vector of each syllable derived from the text to the syllables present in the corpus [9]. At this stage we have considered five syllables for each target syllable. These five units correspond to the top five matched units with respect to the target unit. Unit selection based on the linguistic match is illustrated with an example sentence *"bhaarat ke pradhaanmanthri ne kahaa"*. The text analysis module derives the desired sequence of syllables for the above text as: *bhaa, rat, ke, pra, dhaan, man, thri, ne, ka, haa*. In this utterance there are 10 syllables. The identity of the first syllable is *"bhaa"*, the context of this syllable is represented by the syllable identities of the preceding and the following syllables. In this case the preceding syllable is absent, and the following syllable is *"rat"*. In view of positional information, the preset syllable is at the beginning of the word, beginning of the utterance and the word position is one. Here, for the target syllable *"bhaa"*, all the units of *"bhaa"* in the corpus are chosen. Then based on syllable context some realizations of *"bhaa"* are filtered out. Likewise, the selection process follows the sequence of filtering the units. The filtering process will be terminated when the number of realizations of the unit reaches to 5.

In the proposed unit selection approach spectral and prosodic matches are performed on the realizations of the target units derived from the first stage of unit selection (i.e., linguistic match). Spectral matching between adjacent units is very crucial in view of minimizing perceptual distortion. Therefore while searching units, appropriate weightage has to be given to the spectral distances between the adjacent units. From the first stage of unit selection module, we get roughly 5 realizations of each unit. For performing spectral match between adjacent units, we need to compute spectral distances between the five realizations of the present unit and the five realizations of the following unit (i.e., total of 15 spectral distances). Among the five realizations of present unit and the five realizations of following units, a pair of units is selected based on the minimal spectral distance measures.

In this work prosodic match is estimated using the differences in average pitch and energy between the adjacent units. For deriving the unique sequence of sound units based on prosodic match, prosodic distance is estimated between the five realizations of the present and the following units. The optimal pair is selected based on the minimum distance criterion. For implementing the combined spectral and prosodic matching to select the sequence of units, distances between all possible pairs of units derived from the first stage of unit selection, need to be computed. The optimal sequence is derived by minimizing combined distance derived from prosodic and spectral matches together.

Performance of proposed multilevel unit selection process is evaluated by conducting listening tests on synthesized speech samples. Speech samples are synthesized using concatenative speech synthesizer, by implementing the proposed unit selection methodology. Listening tests are conducted using 25 research scholars in the age group of 25-35 years. Four sets of sentences are synthesized using the proposed unit selection process at different levels: (1) The first set of sentences are synthesized by concatenating the sound units derived from unit selection module without implementing spectral and prosodic matching (i.e., unit selection based on the linguistic match only). (2) The second set of sentences are synthesized by sequence of sound units from unit selection module using linguistic match and spectral match (i.e., without prosodic match). (3) The third set of sentences are synthesized by deriving sound units from unit selection module using linguistic match and prosodic match (i.e., without spectral match). (4) The fourth set of sentences are synthesized by deriving sound units using linguistic, spectral and prosodic matches together.

The tests are conducted in the laboratory environment by playing the speech signals through headphones. In the test, subjects were asked to judge the perceptual distortion and quality of the speech on a 5-point scale for each of the sentences. Each listener has to give the opinion scores for each of the five utterances in all four cases (altogether 20 scores) mentioned above. Mean opinion scores (MOS) indicating the quality of synthesized speech are given in Table 2. Different approaches for selecting units are given in the first column of Table 2. The second column in Table 2 shows the MOS for speech quality. The obtained MOS's indicate that the synthesized speech quality has improved by using the proposed multilevel unit selection criterion for choosing optimal sequence of sound units from speech corpus.

Table 2. Mean opinion scores for the quality of synthesized speech for different unit selection approaches

Unit selection approach	Mean opinion score (MOS)
Linguistic match	2.1
Linguistic match + Spectral match	2.5
Linguistic match + Prosodic match	2.6
Linguistic match + Spectral match + Prosodic match	2.8

5 Summary and Conclusion

Multilevel unit selection methodology was proposed to develop the TTS system to enhance the quality of synthesized speech. Hindi broadcast news speech corpus was used for developing the baseline Hindi TTS system. Linguistic, spectral and prosodic features were explored in the unit selection module to choose the optimal sequence of sound units from their multiple realizations. The efficiency of proposed unit selection approach was evaluated by developing Hindi TTS system and carrying out perceptual analysis on synthesized speech. The perceptual analysis showed that quality of synthesized speech was improved by performing the unit selection process using linguistic, spectral and prosodic features in a combined way. The performance of proposed TTS system can be enhanced by manipulating spectral and prosodic features of the concatenated units using the predicted prosody and spectral models.

References

1. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Atlanta, Georgia, USA, May. 1996, vol. 1, pp. 373–376 (1996)
2. Yegnanarayana, B., Murthy, H.A., Sundar, R., Ramachandran, V.R., Kumar, A.S.M., Alwar, N., Rajendran, S.: Development of text-to-speech system for Indian languages. In: Proc. Int. Conf. Knowledge Based Computer Systems, Pune, India, December 1990, pp. 467–476 (1990)
3. Krishna, N.S., Murthy, H.A.: A new prosodic phrasing model for Indian language Telugu. In: INTERSPEECH 2004 - ICSLP, October 2004, vol. 1, pp. 793–796 (2004)
4. Thomas, S., Rao, M.N., Murthy, H.A., Ramalingam, C.S.: Natural sounding TTs based on syllable-like units. In: Proc. 14th European Signal Processing Conference, Florence, Italy (September 2006)
5. Kishore, S.P., Kumar, R., Sangal, R.: A data-driven synthesis approach for indian languages using syllable as basic unit. In: Int. Conf. Natural Language Processing, Mumbai, India (December 2002)
6. Sen, A., Vijaya, K.S.: Indian accent text to speech system for web browsing, Sadhana (2002)
7. Sreekanth, M., Ramakrishnan, A.G.: Festival based maiden TTS system for Tamil language. In: Proc. 3rd Language and Technology Conf., Poznan, Poland, October 2007, pp. 187–191 (2007)
8. Basu, A., Sen, D., Sen, S., Chakrabarthy, S.: An Indian language speech syn- thesizer: Techniques and its applications. In: National Systems Conference, IIT Kharagpur, Kharagpur, India (2003)
9. Rao, K.S., Yegnanarayana, B.: Modeling durations of syllables using neural networks. Computer Speech and Language 21, 282–295 (2007)

Exploring Speech Features for Classifying Emotions along Valence Dimension

Shashidhar G. Koolagudi and K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India
koolagudi@yahoo.com, ksrao@iitkgp.ac.in

Abstract. Naturalness of human speech is mainly because of the embedded emotions. Today's speech systems lack the component of emotion processing within them. In this work, classification of emotions from the speech data is attempted. Here we have made an effort to search, emotion specific information from spectral features. Mel frequency cepstral coefficients are used as speech features. Telugu simulated emotion speech corpus (IITKGP-SESC) is used as a data source. The database contains 8 emotions. The experiments are conducted for studying the influence of speaker, gender and language related information on emotion classification. Gaussian mixture models are use to capture the emotion specific information by modeling the distribution. An average emotion detection rate of around 65% and 80% are achieved for gender independent and dependent cases respectively.

Keywords: Emotion; Emotion recognition; Gausssian mixture models; Telugu emotional speech database; Prosody; Spectral features; Valence.

1 Introduction

Most of state-of-the-art speech systems can efficiently process neutral speech, leading to incomplete and imperfect communication. Human beings always encapsulate the message, with in an envelop of desired emotion. This inbuilt emotion successfully conveys the intended meaning of the message, from the speaker, to the listener. So speech systems capable of processing emotional content of the signal along with proper message, are claimed to be more complete and meaningful. To bring in, the missing naturalness into the processed speech, one has to explore the mechanism of capturing emotions from the natural speech. Emotions through a nonverbal communication, play an important role, in the analysis of, telephonic conversations, call center dialogues and interactive voice response systems(IVRS) [1]. Medical doctors may use emotional content of the patient's voice as a diagnosing tool. Extracting emotions from the tapped telephone conversation of crime suspects, may help forensic department to nab the culprits. Robotic pets and humanoid partners may be more natural and enjoyable, if they can express and recognise emotions . So today's applications prefer the speech systems that can understand and produce emotions.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 537–542, 2009.

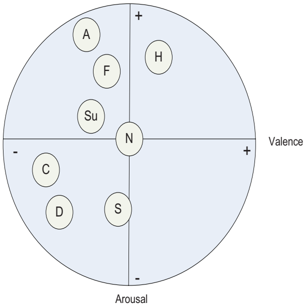


Fig. 1. The distribution of 8 emotions on a two dimensional emotional plane of Arousal and Valence. (A-Anger, C-Compassion/Sad, D-Disgust, F-Fear, H-Happy, N-Neutral, S-Sarcastic, Su-Surprise).

The emotional content of the speech can be visualized in a 3-dimensional space. The three dimensions are, arousal or activation, valence or pleasure and power or dominance [2]. Arousal is a loudness in expression. Valence is a perceptual pleasure, indicating positivism or negativity of the emotion. Power indicates dominance or weakness of an expression. Fig.1 shows the projected three dimensional emotional space, on the plane containing activation and valence as axes. One can observe here that the emotions anger and happy, are not easily distinguishable using arousal, but they can be using valence.

Normally human beings use long term speech features like: energy profile, intonation pattern, duration variations and so on, for detecting the emotions [3]. These are known as prosodic features. This is the reason that, most of the emotional speech related literature is shaped around the close vicinity of either prosodic features or their derivatives. But it is difficult to distinguish the emotions, that share common acoustic and prosodic features, using only these longterm features. It is shown in the Table 1, that quite frequently the emotions like anger and happy are inter miss-classified, with only prosodic features.

Table 1. Percentage of inter miss-classification of Anger and Happy emotions quoted by different researchers in the literature

Reference	Language	Percentage of anger utterances classified as happy	Percentage of happy utterances classified as anger
Serdar yeldirim, et.al. [4]	English	42	31
Dimitrios Virviridis, et.al. [5]	Scandinavian language (danish)	20	14
Felix Burkhardt, et.al. [6]	German (Berlin Emotion Database)	–	12
Oudeyer Pierre, et.al. [7]	Concatenated synthesis (English)	35	30
S G. Koolagudi, et. al.[8]	Telugu	–	34
S. G. Koolagudi	Berlin, German	27	20
Raquel tato [9]	German	24	25

2 The Collection of Speech Database and Selection of Features

The word 'emotion' is often open to quite large number of interpretations. In reality, emotions are basically pervasive in nature, but research models are mostly being built for their full blown versions. The emotional speech corpora may be collected in three major ways. The expert artists are asked to produce verbal emotions to get *simulated database.* The artists or normal people are made to produce different emotions by creating respective situations, without the knowledge of the speaker, giving *induced emotional database. Naturalistic data corpus* contains the recordings of natural emotional conversations. In this work, simulated speech corpus of Telugu language, is used for emotion analysis. The 8 emotions considered are anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise . 10 speakers (5 male and 5 female) of varied age and experience from All India Radio (AIR) station, are recorded for preparing the corpus. 15 linguistically neutral sentences are uttered by each speaker in 10 different sessions. So the database contains 1500 utterances ($15 sentences \times 10 sessions \times 10 speakers$) for each emotion and 12000 utterances ($1500 \times 8 emotions$) in total. Proper care has been taken to include all phoneme classes, speaker, gender, text and session variations. The speech is recorded with a sampling frequency of 16kHz and each sample is stored as a 16 bit number [8].

Here we tried to identify the reliable speech features, that are least influenced by gender, speaker, language, cultural and contextual variations, while detecting emotions. Valence or perceptual pleasure is mainly observed in the speech due to conscious vocal effort through articulator activities. So appropriate spectral features - representing the vocal tract characteristics may be used as the features of interest. Therefore MFCC features are used to represent valence information of emotional utterances. The motivation to use these features, is that 'MFCC's' represent vocal tract characteristics of a speaker and pleasure (valence) of an emotional utterance. In fact MFCC's are derived using audio critical bands of human perceptive system and understanding of an emotion, by human being, is largely an individualistic perception. In this work the experiments are conducted to justify the hypothesis 'Spectral features alone are sufficient to classify the emotions along valence axis'.

3 Results and Discussion

The identification of different emotions would be reasonably possible, only when all the emotions under study, are properly discriminated in an emotional space using suitable features. Here the spectral features are explored for characterizing the emotions, with respect to the valence dimension. To capture the distribution of emotions, Gaussian mixture models are used. Single GMM is used for one emotion. So the model, in total contains 8 GMM's. 12 MFCC features extracted from a frame of a signal, along with one energy value, formed a feature vector of size 13. All the utterances of specific emotion are taken from the first 8 sessions

Table 2. Confusion Matrix for classification of emotions for the model trained using single female utterances. 13 MFCC's were used to construct feature vector, GMM contains 128 components and converged with 200 iterations.

	Anger	Compassion	Disgust	Fear	Happy	Neutral	Sarcastic	Surprise
Anger	60	0	27	0	0	3	10	0
Compassion	0	87	3	7	3	0	0	0
Disgust	23	0	70	0	0	0	7	0
Fear	10	3	0	84	0	0	0	3
Happy	0	0	0	0	84	10	3	3
Neutral	3	0	0	0	0	97	0	0
Sarcastic	3	0	3	0	0	0	94	0
Surprise	0	3	0	3	23	0	0	71

of the corpus and used for training the GMM's. 30 randomly selected utterances of each emotion, from the remaining 2 sessions are used for validating the trained models.

In this work the effort has been made to reduce the discrimination discrepancies caused due to prosodic features, by classifying the emotions on the basis of valence dimension. Table 2 is a confusion matrix obtained from the emotion classification results of the single female actor's utterances. GMM's with 128 components and 200 epochs towards convergence, yielded an average emotion detection rate of 80.42%. The spectral features (MFCC's) here, are able to clearly classify the emotions like anger and happy, which share similar acoustic characteristics along arousal or activation axis. It may be observed from Table 2, that none of either happy or anger utterances is inter miss-classified.

The summary of the classification results obtained using different, GMM configurations and text dependent training sets is briefed in Table 3. Column 2A represents the emotion classification performance of the models, built using single male speaker utterances. Similar results of female speaker are tabulated in column 2B. Performance of female emotion recognition is better due to clear expression of emotions by female artists. It is also supported by the MOS of subjective listening tests. Columns 2C and 2D represent the classification results, where speaker related information is generalised by training the models with the utterances of multiple speakers of the same gender. Column 2E represents the performance, where emotion recognition models are built with both the genders.

Table 4. consolidates the emotion recognition performance for text independent cases. Here the texts of training and testing utterances are different. This experiment is to verify the effect of phoneme related information on the classification results. Almost near to similar results are observed for text dependent and independent cases for female voices. Observe the columns 2A as well as 2B of tables 3 and 4 respectively . The higher performance in case of male emotion recognition, for text dependent case is obvious because of phonemic information playing role during classification (see the columns 2A of Tables 3 and 4) and male voices are less expressive compared to female voices. But the similar trend is not observed for female emotive utterances, the slight improvement in the emotion

Table 3. Average emotion classification performance for text dependent case. Emotions considered for analysis are anger,compassion, disgust, fear, happy neutral, sarcastic, surprise.

1 GMM's C-No.of components I-No.of epochs	2 Training Sets				
	2A Single Male	2B Single Female	2C 3 Males	2D 3 Females	2E 3Males + 3Females
64C-100I	73.33	80.63	66.25	73.25	59.25
64C-200I	74.58	79.17	65.75	74.75	63.38
128C-100I	75.83	80.42	73.25	76.37	61.63
128C-200I	77.92	80.42	73.37	76.75	63.75

Table 4. Average emotion classification performance for male and female voices in text independent case.Average emotion classification performance for text dependent case. Emotions considered for analysis are anger,compassion, disgust, fear, happy neutral, sarcastic, surprise.

1 Gaussian Mixture Models C-No.of components I-No.of epochs	2 Training Sets	
	2A Single Male	2B Single Female
64C-100I	63.33	84.12
64C-200I	63.75	80.00
128C-100I	66.25	85.42
128C-200I	65.83	83.33

recognition performance of text independent case justifies the very little role of phoneme based information towards classification. It reveals the fact that along with the speaker and phoneme specific information, spectral features also contain robust emotion specific information. The clear classification of emotions like, anger and happy motivates one to attribute this robust emotion specific information to MFCC features and hence the classification can be justified to be along valence or pleasure dimension. The above experiments are designed to show the minimal influence of speaker, gender and phoneme related information during the classification.

4 Summary and Conclusions

In this work, features representing the characteristics of vocal tract system (Spectral features) were proposed to discriminate the emotions. The hypothesis considered is that, 'the characteristics of VT system will follow the valence dimension of the emotions'. It has been shown here that, the emotions can be efficiently discriminated using the features contributing to emotional valence. Valence being pleasure in perception, is mainly contributed by the spectral features. So MFCC features were used for classifying the emotions. These are found to be robust amongst other features, while classifying the emotions amidst speaker, gender and language variations. Maximum average emotion recognition performances for 8 given emotions, in case of single male and female speakers are around 78% and 80% respectively.

It is still a difficult challenge to classify the closely spaced emotions along pleasure and power axes. For example anger and disgust are very close in the negative region along pleasure axis, and they are on the either side of the origin along power axis. It is important to note that one cannot claim the spectral features, as the sole carriers of emotion specific information in their pleasure space, but results have shown that they have major contribution towards the valence. The combination of prosodic and spectral features may be more robust to represent the all 3 known dimensions of emotions.

References

1. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. IEEE Trans. Speech and Audio Processing 13, 293–303 (2005)
2. Jin, X., Wang, Z.: An Emotion Space Model for Recognition of Emotions in Spoken Chinese. In: Tao, J., Tan, T., Picard, R.W. (eds.) ACII 2005. LNCS, vol. 3784, pp. 397–402. Springer, Heidelberg (2005)
3. Rao, K.S., Yegnanarayana, B.: Intonation modeling for indian languages. CSL (2008)
4. Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Busso, C., Deng., Z., Lee, S., Narayanan, S.: An acoustic study of emotions expressed in speech. In: Int'l Conf. on Spoken Language Processing (ICSLP 2004), Jeju island, Korean (October 2004)
5. Ververidis, D., Kotropoulos, C., Pitas, I.: Automatic emotional speech classifcation. In: ICASSP 2004, pp. I593–I596. IEEE, Los Alamitos (2004)
6. Burkhardt, F., Sendlmeier, W.F.: Verification of acousical correlates of emotional speech using formant-synthesis. In: ITRW on Speech and Emotion, Newcastle, Northern Ireland, UK, September 5-7, pp. 151–156 (2000)
7. Oudeyer, P.-Y.: The production and recognition of emotions in speech: features and algorithms. International Journal of Human Computer Studies 59, 157–183 (2003)
8. Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Rao, K.S.: IITKGP-SESC: Speech Database for Emotion Analysis. In: Communications in Computer and Information Science, JIIT University, Noida, India, August 17-19. Springer, Heidelberg (2009)
9. Tato, R., Santos, R., Pardo, R.K.J.: Emotional space improves emotion recognition. In: 7th International Conference on Spoken Language Processing, Denver, Colorado, USA, September 16-20 (2002)

A Novel Time Decaying Approach to Obstacle Avoidance

Sankalp Arora and S. Indu

Faculty of Delhi College of Engineering, Delhi College of Engineering, Delhi University,
New Delhi-110042, India
ar.sankalp@gmail.com, s.indu@rediffmail.com

Abstract. One of the basic issues in navigation of mobile robots is the obstacle avoidance task which is commonly achieved using reactive control paradigm where a local mapping from perceived states to actions is acquired. The algorithms of this class suffer from a major drawback of exhibiting cyclic behavior when encountered with certain obstacle configurations. This paper presents a cognitive time decaying approach to overcome this cyclic behavior .The Dynamic Window algorithm is taken as an example for implementing this approach. To build a dynamic window based obstacle avoider, we use time decaying heuristic function for history mapping - which innately eliminates local minima even for a cluttered environment and gives the robot an exploratory nature best suited for map building purposes. The algorithm is successfully tested on a simulation, where it is shown to avoid the U bend problem of local minima.

Keywords: obstacle avoidance, dynamic window, local minima, cognitive.

1 Introduction

Many mobile robot systems combine a global path-planning module with a local obstacle avoidance module to perform navigation. While the global path planner determines a suitable path based on a map of the environment, the obstacle avoidance algorithm determines a suitable direction of motion based on recent sensor data. Obstacle avoidance is performed locally in order to ensure that real-time constraints are satisfied. A fast update rate allows the robot to safely travel at high speeds. Obstacle avoidance paradigm was introduced by Khatib in 1986 [1] through the concept of artificial potential fields. There were several reactive algorithms suggested [2], [3], [4] after that but none of them took the complete dynamics of the platform involved in their world models. In 1997 though Dieter Fox, Wolfram Burgard and Sebastian Thrun introduced the dynamic window approach [5] to obstacle avoidance. It differs from previous approaches by searching the velocity space for the appropriate controlling command. This is done by reducing the search space to the dynamic window which consists of velocities reachable in a short interval of time.

Although the dynamic window method has been very successful and is extensively in use, it suffers from the local minima problem that is getting stuck in cyclic behavior for certain configurations of obstacles. To avoid this cyclic behavior Global Dynamic Window Approach [6] was suggested by O. Brock and O. Khatib in 1999. It

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 543–548, 2009.

used a continuous navigation function with only one absolute minimum at the goal. This algorithm also had tendency to somehow move away from the goal for certain obstacle configurations as pointed out by Petter Ögren and Naomi Ehrich Leonard in the paper "A Convergent Dynamic Window approach to Obstacle Avoidance"[7] in 2005. The work of Petter Ögren and Naomi Ehrich Leonard suggests model predictive control method and lyapunov function based approach to resolve the issues of local minima and map trotting. Both the above works assume that the simultaneous localization and mapping problem is solved. Though this is possible using the range of SLAM techniques available but these algorithms are processor intensive and compromise the reactive nature of the algorithm.

In this paper, inspiration is drawn from the cognitive behavior of history mapping shown by humans in order to plan paths in an unknown environment. To implement the same, time decay based approach is suggested, to avoid cyclic behavior in obstacle avoiding algorithms. This approach considers and maps the recent past activities of the robot in a small local map and leaves time decaying stamps on its trail, when considering future paths these time decaying stamps are taken into account and path which is least traversed as well as more suited for reaching the goal is chosen. The tendency of the robot to take the least traversed path makes sure that it does not exhibit cyclic behavior. The flexibility of the approach lies in the fact that the increasing localization error can be dealt with by increasing the rate of decay of the time stamps and hence the approach can be implemented with or without using SLAM techniques .Therefore retaining the reactive nature of the original algorithm.

2 Local Obstacle Avoidance – Dynamic Window Approach

The dynamic window approach originally proposed is especially designed to deal with the constraints imposed by limited velocities and accelerations, because it is derived directly from the motion dynamics of synchro-drive mobile robots. In a nutshell, the approach considers periodically only a short time interval when computing the next steering command to avoid the enormous complexity of the general motion planning problem. The approximation of trajectories during such a time interval by circular curvatures results in a two-dimensional search space of translational and rotational velocities. This search space is reduced to the admissible velocities allowing the robot to stop safely.

Due to the limited accelerations of the motors a further restriction is imposed on the velocities: the robot only considers velocities that can be reached within the next time interval. These velocities form the dynamic window which is centered around the current velocities of the robot in the velocity space. Among the admissible velocities within the dynamic window the combination of translational and rotational velocity is chosen by maximizing an objective function. The objective function includes a measure of progress towards a goal location, the forward velocity of the robot, and the distance to the next obstacle on the trajectory. By combining these, the robot treads to its desire to move fast towards the goal and its desire to ship around obstacles.

3 The Cognitive Time Decaying Approach

It is observed that that when humans get stuck in complex obstacle configuration, without the knowledge of global maps, history mapping is innately used by them along with approximate map building techniques. As a general tendency, humans during maze solving are least likely to take back the path on which they have just traversed. The time decay approach proposes to introduce the same cognitive behavior in obstacle avoidance algorithms.

Above is achieved by leaving a trail of time decaying time stamps on the map, wherever the robot treads. This trail is then used as history model to inform the algorithm about how much time back the robot had visited a cell. Using this past information a **time heuristic** is introduced to motivate the robot to take the least visited or the least recently visited path while avoiding the obstacles and moving towards the goal as suggested by the algorithm it is applied on. **As the robot has the tendency to take the least recently visited path, its probability of getting stuck at the local minima or exhibiting cyclic behavior is also highly reduced**. It should be noted here that the rate of decay of time stamps can be controlled as per the localization errors of a given platform; hence this approach can be used with or without SLAM techniques.

It is understandable that updating the whole map in every iteration would be difficult, hence only the active area (area around the robot) is updated for decayed trail values. Thus retaining the reactive nature of the algorithm as well as making it less susceptible to cyclic behavior. It is important here to note that only the area of map which was being refreshed by new proximity sensor reading is being refreshed now, even though there is a complete map present in the memory. This makes the algorithm more memory intensive without having any considerable affects on the runtime.

4 The Time Heuristic

The time decaying approach introduces a new Time_function, added to the original heuristic, so that the algorithm knows about its recent past and moves towards the goal while circumventing collisions as fast as it can and also avoiding the recently taken paths so as to avoid entering into the local minima situations. The resulting algorithm might not give the best path towards its goal but is highly reliable in terms of its reaching the goal.

The time function for a given candidate path is calculated by summing the time objectives of each cell that a candidate path encounters.

4.1 The Time Objective –

The time objective of a given cell is calculated by inverting the subtraction of current time from the time stamp left on the cell earlier, whole raise to power N. The previously unvisited cells are given a zero time objective value.

$$\text{T\_obj}(t) = [1/(\text{current\_time} - \text{cell\_time\_stamp})]^{N} \quad (1)$$

4.2 The Time Function –

The time_function for any given path is calculated by summing the time objectives of all the cells to be encountered in a given path and multiplying it with a normalizing value. Mathematically it is given as-

$$T\_function(t) = (-1)*\psi* \sum T\_obj(t) \tag{3}$$

The time function is thus obtained for every probable path in the given dynamic window. This function is then added to the existing heuristic to allow the algorithm to take an informed decision, considering its past states as well. As can be observed that the time objective for a given cell is always positive and attains a maxima when the time stamp of the cell is very close to current time. The time function inverts this relationship between time stamps and time objective values by introducing an extra negative sign. This ensures that the time function is maximum(0), for a path with no visited cells in it and minimum for the path having most recently visited cells in it. As we try to maximize the total heuristic the paths with least recently visited cells are automatically given more priority than the others, thus resulting in behavior that avoids cyclicity innately.

4.3 Variation with N –

N signifies the rate of decay of time_obj values with time for a given cell. As N increases the rate of decay increases exponentially. For situations where the positioning error of a robot is less and does not increase rapidly with distance traversed then small values of N are acceptable for efficient implementation of algorithm. But if the positioning error of the robot increases rapidly with distance traversed then large values of N are required for the algorithm to work efficiently.

In the second case the robot acquires more exploratory nature and hence might take more time to get out of the obstacle configuration.

4.4 Variation with Ψ –

Psi signifies the normalizing weightage allotted to the property of the robot to distinguish between visited and non – visited paths, and choosing the most non-visited path more than other paths. Greater psi is kept , lesser is the probability of the robot to choose the visited path. Although increasing the value of psi to a great extent would lead to the robot only choosing the least visited path even if the other functions like heading and distance heavily favor other paths and hence may lead to the failure of the algorithm. It should be noted that the time function is introduced in order to help the algorithm to take an informed decision and should not be the only criterion based on which the next state is decided. The new heuristic after incorporating time-decaying approach is given by-

$$G = \alpha*heading(v,w)+ \beta*dist(v,w)+ \gamma*velocity(v,w)+ Time\_func(t) \tag{4}$$

4.5 Parameter Settings

Although the performance of the obstacle avoidance depends on the weighting parameters N and psi, it is stable against slight changes of their values. Without any exhaustive tuning of these parameters we found values of 1.5 and 25% of total weightage, for N and psi to give good results. The tendency of the robot to avoid the past taken paths is defined by the relative value of psi and the memory of the algorithm is a function of N. By choosing different values for parameters, global knowledge about the environment can be transferred to the local obstacle avoidance. While higher values produce good results in wide environments, smaller values are more appropriate in narrow and populated hallways. An improvement upon the parameter tuning approach would be – self learning of parameters during run time. But the stated technique may need extensive overhauling of the simple history model currently used.

5 Results

As can be seen from figure 1, the simple Dynamic Window Approach algorithm is exhibiting cyclic behavior inside the U shape obstacle configuration but successfully circumvents the obstacle (Fig. 2), using the time decaying approach. The robot is continuously trying to move towards its goal while avoiding the past paths that it has taken. The value of N in this case was taken to be 1 and psi was used for normalization and hence was taken to be 25% of the total weightage. It is important here to mention that this approach is not only limited to dynamic window based obstacle avoiders but also to other reactive algorithms with minor changes according to their heuristics.

Figure 3 shows the path treaded by the robot under a more exploratory heuristic, the value of N in this case was taken to be 3 and value of psi was also increased to an optimal value as can be seen the robot circumvents the obstacle but is more inquisitive in nature , and explores more area before reaching its goal. It is this nature of the algorithm that makes it suitable for use in conjunction with mapping techniques for map building purposes. Otherwise if the robot is used only for goal reaching purposes, it works best in conjunction with a SLAM technique.

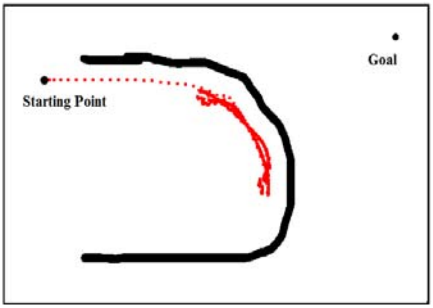


Fig. 1.

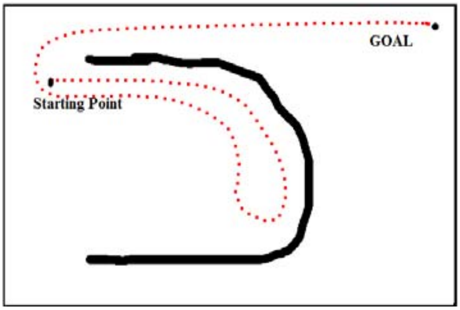


Fig. 2.

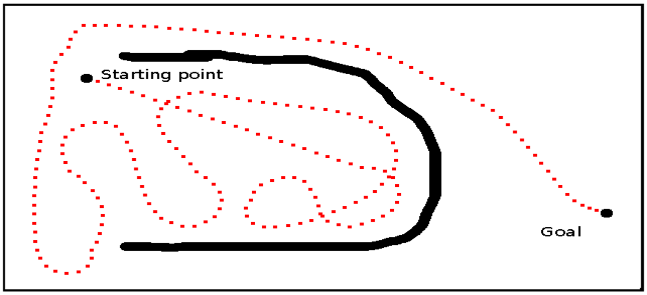


Fig. 3.

6 Conclusion

In this paper we first compare various obstacle avoidance techniques and then present the well established dynamic window approach in detail. We then presented a cognitive history mapping technique, achieved by leaving a trail of decaying time stamps along the robots past positions. As an example, the dynamic window approach was combined with time decaying methods, which has shown promising results in simulation. The approach is shown to innately avoid cyclic behavior without compromising on local and reactive nature of the algorithm. The algorithm is able to do so without considerably adding to the computation costs, whereas in other provably convergent approaches the addition on computation costs is high, and the algorithm shifts from local to global paradigm. The approach so presented is also shown to be flexible towards localization errors and gives an exploratory nature to the robot, best suitable for map building purposes. The time decaying approach can be combined with other obstacle avoidance algorithms as well with little changes in implementation details.

References

[1] Khatib, O.: Real-time obstacle avoidance for robot manipulator and mobile robots. The International Jour. of Rob. Res. 5(1), 90–98 (1986)

[2] Moravec, H.P.: Sensor fusion in certainty grids for mobile robots. AI Magazine, 61–74 (1988)

[3] Koren, Y., Borenstein, J.: Potential field methods and their inherent limitations for mobile robot navigation. In: Proc. IEEE Int. Conf. Robotics and Automation (April 1991)

[4] Borenstein, J., Koren, Y.: The vector field histogram - fast obstacle avoidance for mobile robots. IEEE Trans. Robot. and Auto. 7(3), 278–288 (1991)

[5] Fox, D., Burgard, W., Thrun, S.: The dynamic window approach to collision avoidance. IEEE Robot. Autom. Mag. 4, 23–33 (1997)

[6] Brock, O., Khatib, O.: High-speed navigation using the global dynamic window approach. In: Proc. IEEE Int. Conf. Robot. Autom., De troit, MI, May 1999, pp. 341–346 (1999)

[7] Ögren, P., Leonard, N.E.: A Convergent Dynamic Window approach to Obstacle Avoidance. IEEE Trans. Robot 21(2), 188–195 (2005)

Chaotic Synchronization and Secure Communication Using Contraction Theory

Bharat Bhushan Sharma* and Indra Narayan Kar

Department of Electrical Engineering,
Indian Institute of Technology, Delhi,
Hauz Khas, New Delhi- 110016, India
bbs.iit@gmail.com,
ink@ee.iitd.ac.in

Abstract. Here, observer based synchronization and secure communication scheme is presented for chaotic systems. In proposed scheme, extended Kalman filter based receiver is selected for given transmitter system. The stability results are derived using virtual system concept. Observer gains for synchronization are obtained as a solution of matrix Riccati equation. For secure communication, n-shift ciphering algorithm is used with one of the chaotic state chosen as key. Numerical simulations are presented in the end to verify the efficacy of proposed scheme.

Keywords: Synchronization, Observer, Secure Communication.

Synchronization of chaotic systems and its application to secure communication gained momentum after the work of Pecora et al. [1]. Synchronization uses output of drive system to control the slave system s.t. its output matches with that of drive system. Various synchronization schemes have been successfully applied to synchronization of chaotic systems [2,3,4]. Some observer based schemes for synchronization and secure communication can be found in [5,6,7,8,9]. Extended Kalman filter (EKF) based approach for secure communication using multi-shift ciphering approach is given in [6]. Most of these techniques are based on Lyapunov based stability analysis. Here, we analyze synchronization of chaotic systems using EKF based observer approach. Exponential convergence of synchronizing scheme is presented using contraction framework [10,11]. This approach is different from Lyapunov analysis as it does not require explicit knowledge of specific attractor. In the scheme presented here, a virtual system is defined for actual & observer system and is shown to be contracting by selecting observer gains suitably. As per contraction theory, if virtual system is contracting then its particular solutions converge to each other, exponentially. For secure communication, an encrypted information signal is transmitted to the receiver where original information is recovered using decryption algorithm. For masking the information, chaotic system states are used as carrier as well as key for improving

[*] Corresponding author: Bharat Bhushan Sharma, Research Scholar, Deptt. of Electrical Engineering, I.I.T. Delhi,India-110016; Tel. No. +91-11-26596133.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 549–554, 2009.

the security. The encryptor uses n-shift cipher involving a nonlinear function. At receiver, both key and carrier are reconstructed to recover original information. Speech signal transmission and retrieval is shown to justify the effectiveness of proposed strategy.

1 Contraction Theory Results

A nonlinear system is contracting if trajectories of perturbed system return to the nominal behaviour with an exponential convergence rate. Consider a nonlinear system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) \tag{1}$$

where $\mathbf{x} \in \Re^m$ and $\mathbf{f}(\mathbf{x}, t)$ is continuously differentiable vector function defined as $\mathbf{f} : (\Re^m \times \Re) \rightarrow \Re^m$. Considering $\delta\mathbf{x}$ to be virtual displacement in $\mathbf{x}$, first variation of system (1) will be $\delta\dot{\mathbf{x}} = \frac{\partial \mathbf{f}(\mathbf{x},t)}{\partial \mathbf{x}}\delta\mathbf{x}$. Defining squared distance between neighbouring trajectories as $\left(\delta\mathbf{x}^T\delta\mathbf{x}\right)$, one can write

$$\frac{d}{dt}\left(\delta\mathbf{x}^T\delta\mathbf{x}\right) = 2\delta\mathbf{x}^T\frac{\partial \mathbf{f}}{\partial \mathbf{x}}\delta\mathbf{x} = 2\delta\mathbf{x}^T\mathbf{J}\delta\mathbf{x} \leq 2\lambda_m(\mathbf{x}, t)\delta\mathbf{x}^T\delta\mathbf{x} \tag{2}$$

Here, $\lambda_m(\mathbf{x}, t)$ represents the largest eigenvalue of the symmetric part of Jacobian matrix $\mathbf{J}$. If $\lambda_m(\mathbf{x}, t)$ is strictly uniformly negative definite (UND), then any infinitesimal length $\|\delta\mathbf{x}\|$ converges exponentially to zero. It is assured from (2) that all the solution trajectories of the system (1) converge exponentially to single trajectory, independently of the initial conditions.

Definition 1. *For system in (1), a region of state space is called contracting region if Jacobian* $\mathbf{J}$ *is UND in that region i.e. there exists a scalar* $\alpha > 0$, $\forall\mathbf{x}$, $\forall t \geq 0$ *s.t.* $\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \leq -\alpha\mathbf{I} < \mathbf{0}$. *It further implies that,* $\frac{1}{2}\left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} + \frac{\partial \mathbf{f}^T}{\partial \mathbf{x}}\right) \leq -\alpha\mathbf{I} < \mathbf{0}$.

Lemma 1. *For system (1), any trajectory which starts in a ball of constant radius centered about a given trajectory and contained at all times in a contraction region, remains in that ball and converges exponentially to given trajectory. Further, global exponential convergence to given trajectory is guaranteed if whole state space region is contracting.*

Various other results related to contraction theory can be found in [10,11,12].

1.1 Stability Using Virtual System Theory

Let the dynamics of a given system is rearranged as $\dot{\mathbf{x}} = \rho(\mathbf{x}, \mathbf{p}, t)\mathbf{x}$, where $\mathbf{x} \in \Re^m$, $\mathbf{p}$ is parameter vector and function $\rho(\mathbf{x}, \mathbf{p}, t) \leq \alpha\mathbf{I} < \mathbf{0}$. To show this system to be uniformly globally exponentially stable, virtual system is defined as $\dot{\mathbf{y}} = \rho(\mathbf{x}, \mathbf{p}, t)\mathbf{y}$. This system is having origin and actual system as its particular solutions. In differential framework, $\delta\dot{\mathbf{y}} = \rho(\mathbf{x}, \mathbf{p}, t)\delta\mathbf{y}$. Using UND nature of $\rho(\mathbf{x}, \mathbf{p}, t)$, states of virtual systems are ensured as contracting. It implies exponential convergence of states of actual system to each other, being a particular solution of virtual system.

In case of observer design, virtual system is defined exactly like observer. For actual system $\dot{\mathbf{x}}_s = \mathbf{x}_s + u$, let observer be selected as $\dot{\hat{\mathbf{x}}} = \hat{\mathbf{x}} + u + L(\mathbf{x}_s - \hat{\mathbf{x}})$. Here, u is control input and gain L is chosen s.t. $L > 1$. Let the virtual system is defined as $\dot{\mathbf{x}} = \mathbf{x} + u + L(\mathbf{x}_s - \mathbf{x})$. For contraction, Jacobian of virtual system should be UND. It will lead to exponential convergence of observer states $\hat{\mathbf{x}}$ to actual system states $\mathbf{x}_s$ as both are particular solutions of the virtual system. These results can be extended to nonlinear observer design accordingly [11].

2 Problem Formulation

Consider the dynamics of transmitter system be

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \theta, t) \tag{3}$$

with output as $\mathbf{y} = \mathbf{s}(\mathbf{x}, t)$, where $\mathbf{x} \in \Re^n$, $\mathbf{y} \in \Re^p$, θ is parameter vector and $\mathbf{s}$ is linear function of measurable states. Let EKF based receiver is given as

$$\dot{\varsigma} = \mathbf{f}(\varsigma, \theta, t) + \mathbf{K}(\varsigma, t)\,(\mathbf{y} - \mathbf{s}(\varsigma, t)) \tag{4}$$

Here, $\mathbf{K}(\varsigma, t)$ is time varying observer gain matrix and ς represents the estimate of actual system states $\mathbf{x}$. The output of this system is $\widehat{\mathbf{y}} = \mathbf{s}(\varsigma, t)$.

2.1 Synchronization and Secure Communication Scheme

In transmission-retrieval scheme shown in fig. 1, $\widehat{\mathbf{x}}$ is estimate of system states $\mathbf{x}$. Transmitter state x_c is used as carrier for modulating the information $m(t)$. Masked signal given by $S_m(t) = x_c + m(t)$ is encrypted using n-shift cipher algorithm [4], where key signal $k(t)$ is used n-times to encrypt the modulated signal. To increase the security, one of chaotic state x_k is used as key signal. Encrypted signal is given as $e(S_m(t)) = \varphi\left(\ldots, \varphi\left(\varphi\left(S_m(t), k(t)\right), k(t)\right), \ldots, k(t)\right)$. Here, signal $S_m(t)$ is shifted n times using nonlinear function $\varphi\left(S_m(t), k(t)\right)$ given as

$$\begin{aligned}
\varphi\left(x, k\right) &= (x+k) + 2l, \qquad -2l \le (x+k) \le -l \\
&= (x+k), \qquad\quad\;\; -l < (x+k) < l \\
&= (x+k) - 2l, \qquad l \le (x+k) \le 2l
\end{aligned} \tag{5}$$

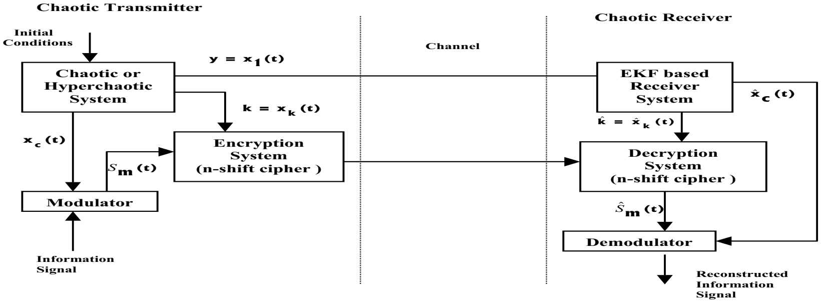

Chaotic Transmitter
Initial Conditions
Chaotic or Hyperchaotic System
y = x₁(t)
k = x_k(t)
x_c(t)
Encryption System (n-shift cipher)
S_m(t)
Modulator
Information Signal
Channel
Chaotic Receiver
EKF based Receiver System
x̂_c(t)
k̂ = x̂_k(t)
Decryption System (n-shift cipher)
Ŝ_m(t)
Demodulator
Reconstructed Information Signal

Fig. 1. Synchronization and secure communication scheme

Parameter l is chosen s.t. both x & $k(t)$ lie in the range $(-l,\ l)$. Decryption rule $\widehat{S}_m(t) = \varphi\left(\ldots, \varphi\left(\varphi\left(S_m(t), -\widehat{k}(t)\right), -\widehat{k}(t)\right), \ldots, -\widehat{k}(t)\right)$, is used to recover the information. i.e. information is recovered from $\widehat{S}_m(t)$ as $\widehat{m}(t) = \widehat{S}_m(t) - \widehat{x}_c$.

2.2 Stability Analysis for Observer Based Communication Scheme

For the systems in (3) & (4), let observer gain $\mathbf{K}(\varsigma, t)$ is defined as

$$\mathbf{K}(\varsigma, t) = \mathbf{P}(t)\mathbf{C}^T(\varsigma, t)\mathbf{S}^{-1} \tag{6}$$

For transmitter (3), define matrices $\mathbf{A}(\varsigma, t) = \frac{\partial \mathbf{f}(\mathbf{x},\theta,t)}{\partial \mathbf{x}}|_{\mathbf{x}=\varsigma}$ and $\mathbf{C}(\varsigma, t) = \frac{\partial \mathbf{s}(\mathbf{x},t)}{\partial \mathbf{x}}|_{\mathbf{x}=\varsigma}$. To obtain $\mathbf{K}(\varsigma, t)$, time varying uniformly positive definite matrix $P(t)$ is determined as a solution of matrix differential Riccati equation:

$$\dot{\mathbf{P}}(t) = \mathbf{A}\mathbf{P}(t) + \mathbf{P}(t)\mathbf{A}^T - \mathbf{P}(t)\mathbf{C}^T\mathbf{S}^{-1}\mathbf{C}\mathbf{P}(t) + \mathbf{Q} \tag{7}$$

Here, $\mathbf{A} = \mathbf{A}(\varsigma, t),\ \mathbf{C} = \mathbf{C}(\varsigma, t)$ and covariance matrices $\mathbf{Q}$ & $\mathbf{S}^{-1}$ are assumed to be constant symmetric positive definite (s.p.d.) matrices.

Assumption 1: In the Riccati equation (7), the solution matrix $\mathbf{P}(t)$ is assumed to be uniformly positive definite and bounds on $\mathbf{P}(t)$ are defined as $P_{min} \leq \mathbf{P}(t) \leq P_{max}$ where P_{min} and P_{max} are positive real numbers.
Synchronization results are presented in the form of a lemma here [12].

Lemma 2. *The EKF based receiver system (3) with gains in (6) gets synchronized with transmitter system (3), exponentially under the assumption 1.*

Proof: Let the virtual system for the systems in (3) & (4) is defined as

$$\dot{\mathbf{x}}_v = \mathbf{f}(\mathbf{x}_v, \theta, t) + \mathbf{K}(\varsigma, t)\left(\mathbf{y} - \mathbf{s}(\mathbf{x}_v, t)\right) \tag{8}$$

with output as $\mathbf{y}_v = \mathbf{s}(\mathbf{x}_v, t)$. In differential framework, one can write,

$$\delta\dot{\mathbf{x}}_v = \left[\frac{\partial}{\partial \mathbf{x}_v}\mathbf{f}(\mathbf{x}_v, \theta, t) - \mathbf{K}(\varsigma, t)\frac{\partial}{\partial \mathbf{x}_v}\mathbf{s}\mathbf{x}_v, t)\right]\delta\mathbf{x}_v = \left[\mathbf{A} - \mathbf{K}(\varsigma, t)\mathbf{C}\right]\delta\mathbf{x}_v \tag{9}$$

Time derivative of weighted square distance $(\delta\mathbf{x}_v^T\mathbf{M}\delta\mathbf{x}_v)$ will be

$$\frac{d}{dt}\left(\delta\mathbf{x}_v^T\mathbf{M}\delta\mathbf{x}_v\right) = \delta\mathbf{x}_v^T\mathbf{M}\delta\dot{\mathbf{x}}_v + \delta\mathbf{x}_v^T\frac{d}{dt}\mathbf{M}\delta\mathbf{x}_v + \delta\dot{\mathbf{x}}_v^T\mathbf{M}\delta\mathbf{x}_v$$

By defining $\mathbf{M} = P^{-1}$ and simplifying using (9), one can get

$$\frac{d}{dt}\left(\delta\mathbf{x}_v^T\mathbf{P}^{-1}\delta\mathbf{x}_v\right) = \delta\mathbf{x}_v^T\mathbf{P}^{-1}\left[\mathbf{A}_c\mathbf{P} - \dot{\mathbf{P}} + \mathbf{P}\mathbf{A}_c^T\right]\mathbf{P}^{-1}\delta\mathbf{x}_v \tag{10}$$

where $\mathbf{A}_c = (\mathbf{A} - \mathbf{K}\mathbf{C})$. Using equation (7) and observer gain $\mathbf{K}(\varsigma, t)$, we get

$$\frac{d}{dt}\left(\delta\mathbf{x}_v^T\mathbf{P}^{-1}\delta\mathbf{x}_v\right) = -\delta\mathbf{x}_v^T\mathbf{C}^T\mathbf{S}^{-1}\mathbf{C}\delta\mathbf{x}_v - \delta\mathbf{x}_v^T\mathbf{P}^{-1}\mathbf{Q}\mathbf{P}^{-1}\delta\mathbf{x}_v$$

$$\Rightarrow \frac{d}{dt}\left(\delta\mathbf{x}_v^T\mathbf{P}^{-1}\delta\mathbf{x}_v\right) = -\delta y_v^T\mathbf{S}^{-1}\delta y_v - \delta\mathbf{x}_v^T\mathbf{P}^{-1}\mathbf{Q}\mathbf{P}^{-1}\delta\mathbf{x}_v \tag{11}$$

This is because differential output $\delta y_v = \mathbf{C}\delta\mathbf{x}_v$. As $\mathbf{S}^{-1}$ is constant symmetric positive definite matrix, so equation (11) can be reduced to

$$\frac{d}{dt}\left(\delta\mathbf{x}_v^T\mathbf{P}^{-1}\delta\mathbf{x}_v\right) \leq -\delta\mathbf{x}_v^T\mathbf{P}^{-1}\mathbf{Q}\mathbf{P}^{-1}\delta\mathbf{x}_v \leq -\kappa\delta\mathbf{x}_v^T\mathbf{P}^{-1}\delta\mathbf{x}_v \tag{12}$$

where $\kappa = q_{min}/P_{max}$, q_{min} is smallest eigenvalue of matrix $\mathbf{Q}$ and P_{max} is largest eigenvalue of $\mathbf{P}$. It clearly shows contracting nature of virtual system. Hence, receiver states converge to transmitter states, exponentially. $\diamond$

3 Numerical Simulations

To analyze the synchronization in transmitter-receiver configuration, consider the chaotic Chen system to be transmitter having dynamics as

$$\begin{aligned}\dot{x}_1 &= -a\,x_1 + a\,x_2\\ \dot{x}_2 &= -x_1x_3 + c\,x_2 + (c-a)\,x_1\\ \dot{x}_3 &= x_1x_2 - b\,x_3\end{aligned} \tag{13}$$

Lt the output is given as $y = x_1$. Here, a, b & c are positive real constants of the system. System in (13) exhibits chaotic behaviour for $a = 40, c = 31$ and $b = 3$. For EKF based receiver system, output is taken as $\widehat{y} = \varsigma_1$ and matrix $\mathbf{A}$ and $\mathbf{C}$ are evaluated as discussed in section 2.2. Initial conditions for transmitter and receiver are taken as $\mathbf{x}(0) = (2\ 1\ 2)^T$ and $\varsigma(0) = (1\ 2\ 1)^T$, respectively.

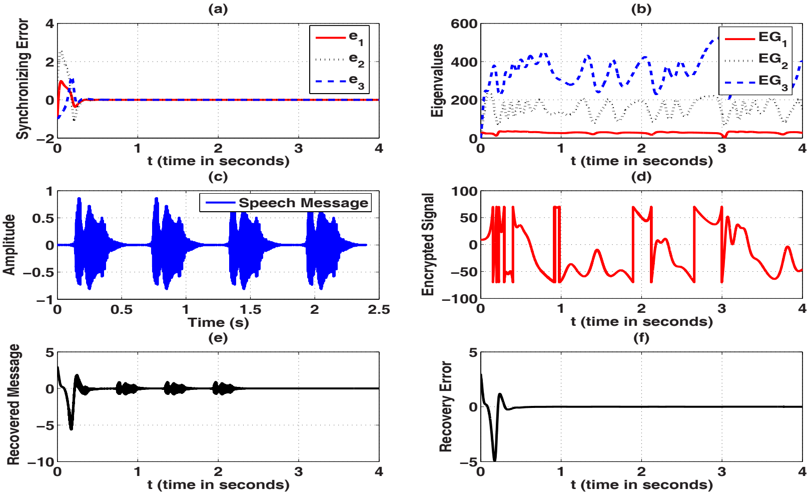


Fig. 2. Synchronization and secure communication: (a) Synchronization error; (b) Variation of eigenvalues of $P(t)$; (c) Speech message; (d) Encrypted signal; (e) Recovered message & (f) Message recovery error

P and **Q** are initialized as diagonal matrices with $P_{i,i} = 1 \times 10^{-4}$ and $Q_{i,i} = 0.35 \times 10^4$, respectively. Fig. 2 gives various plots related to synchronization and communication scheme. Multiple envelopes of a speech signal corresponding to word "HELLO" is transmitted after modulating it with carrier x_2. The signal is sampled at 16 KHz frequency and is having total duration of 0.6 seconds. State x_3 is used as the key signal for encryption. These figures clearly show the effectiveness of proposed approach in secure transmission and recovery of the message signal.

4 Conclusion

Synchronization and secure communication scheme for chaotic systems is presented where exponential stability is shown using virtual system concept in quite a simple manner. Systems once synchronized, can be used for information transmission and recovery. For secure communication, multi-shift key algorithm for encryption is utilized which uses one of chaotic state as key for encryption.

References

1. Pecora, L.M., Carroll, T.L.: Synchronization in Chaotic Systems. Phys. Rev. Lett. 64, 821–824 (1990)
2. Carroll, T.L., Pecora, L.M.: Synchronizing Chaotic Circuits. IEEE Trans. Circ. Syst.-I 38(4), 453–456 (1991)
3. Yang, L.X., Chu, Y.D., Zhang, J.G., Li, F.L., Chang, Y.X.: Chaos Synchronization in Autonomous Chaotic System via Hybrid Feedback Control. Chaos, Solitons and Fractals 41(1), 214–223 (2009)
4. Yang, T., Wu, C.W., Chua, L.O.: Cryptography based on Chaotic Systems. IEEE Trans. on Circ. Syst.-I 44(5), 469–472 (1997)
5. Boutayeb, M., Darouach, M., Rafaralahy, H.: Generalized State-space Observers for Chaotic Synchronization with Applications to Secure Communication. IEEE Trans. Circ. Syst.-I 49(3), 345–349 (2002)
6. Fallahi, K., Raoufi, R., Khoshbin, H.: An Application of Chen System for Secure Chaotic Communication based on Extended Kalman Filter and Multi-Shift Cipher Algorithm. Comm. in Nonlinear Science and Num. Sim. 13(4), 763–781 (2008)
7. Grassi, G., Mascolo, S.: Nonlinear Observer Design to Synchronize Hyperchaotic Systems via a Scalar Signal. IEEE Trans. Circ. Syst.-I 44(10), 1011–1014 (1997)
8. Liao, T.L., Huang, N.S.: An Observer based Approach for Chaotic Synchronization and Secure Communication. IEEE Trans. Circ. Syst.-I 46(9), 1144–1149 (1999)
9. Morgul, O., Solak, E.: Observer based Synchronization of Chaotic Systems. Phys. Rev. E 54(5), 4803–4811 (1996)
10. Lohmiller, W.: Contraction Analysis of Nonlinear Systems, Ph.D. Thesis, Department of Mechanical Engineering, MIT (1999)
11. Lohmiller, W., Slotine, J.J.E.: On Contraction Analysis for Nonlinear Systems. Automatica 34(6), 683–696 (1998)
12. Jouffroy, J., Slotine, J.J.E.: Methodological Remarks on Contraction Theory. In: IEEE Conf. on Decision Control, Atlantis, Bahamas, pp. 2537–2543 (2004)

Fault Diagnosis of an Air-Conditioning System Using LS-SVM

Mahendra Kumar[1] and I.N. Kar[2]

[1] Working as Deputy Chief Electrical Engineer in Northern Railway, Delhi-110001
[2] Department of Electrical Engineering, Indian Institute of Technology, Delhi-110016
mahendray71@gmail.com, ink@ee.iitd.ac.in.

Abstract. This paper describes fault diagnosis of an air-conditioning system for improving reliability and guaranteeing the thermal comfort and energy saving. To achieve this goal, we proposed a technique which is model based fault diagnosis technique. Here, a lumped parameter model of an air-conditioning system is considered and then characteristics of twelve faults are investigated in an air-conditioning system provided in passenger coach of an Indian Railway. Based on the variations of the system states under normal and faulty conditions of different degrees, the faults can be detected efficiently by using residual analysis method. The residual code is obtained through simple threshold testing of residuals, which are the output of a general scheme of residual generators. The pattern of residual is classified by using multi-layer LS-SVM classification. The diagnosis results show that LS-SVM classifier is effective with very high accuracy.

Keywords: LS-SVM, Residuals generator, Air-Conditioning System, FDD.

1 Introduction

The reliability of an air-conditioning (AC) system has a prime importance as failure of AC system can lead to occupant's discomfort, higher health and safety risks. The faulty and non optimal operation of an AC system wastes an estimated 15% to 30% of energy used in commercial building. Hence, fault detection and diagnosis (FDD) methods for an air-conditioning system is an important area. The main purpose of FDD is, to detect, locate and if possible, predict the presence of the defects causing faulty operation. A fault is detected when the observed behaviour of a system differs from the expected behaviour by some threshold. The expected behaviour of the system is expressed in a model whether physical, statistical or fuzzy. In this paper we have used model based on physical laws. The fault diagnosis has been done by residual analysis. A residual is often a time varying signal that is used as a fault detector. Normally, the residual is designed to be zero (or small in a realistic case where process is subjected to noise and the model is uncertain) in fault free case and deviate significantly from zero when fault occurs. So fault is identified by comparing residual with threshold value. The sequence of residual is identified using LS-SVM. The least square support vector machine (LS-SVM) is used for classification of different residuals pattern. In this paper twelve faults have been considered for diagnosis and

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 555–560, 2009.

eight residuals have generated for fault classification. Fault diagnosis of an air conditioning system has been the subject of many studies during the past decades. Generally, the fault detection and diagnosis (FDD) methods can be divided into three types [7]: the feature-based method , the model-based method , and a combination of both. Although these methods have been applied in a number of industrial processes with good performance, their application in the HVAC system is still at the research stage in laboratories [6]. The main contributions of this paper are as follows:

1. A model based fault diagnosis schemes have been developed by using residuals for diagnosing major air-conditioning faults.

2. The fault diagnosis is done by residuals pattern classification using LSSVM as classifier.

3. The residuals of the fault are generated by using an air-conditioning system provided in passenger coach of an Indian Railway and simulation of the proposed fault diagnosis scheme has been done.

2 Design of Model Based Fault Diagnosis Scheme

The first requirement of model based fault diagnosis scheme is model of the plant which represents expected behaviour of the system. In this paper we have considered lumped parameter model of an air conditioning system. In model based fault diagnosis scheme the output and control input of the plant and the output and control input of fault free model is compared and residuals are generated. A pattern of residuals is generated for a fault and in this paper we have proposed fault diagnosis scheme using LS-SVM as shown in Fig.1.

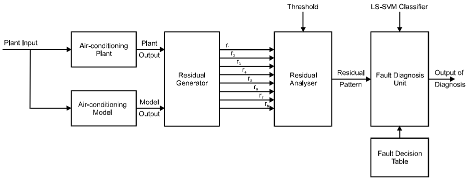


Fig. 1. The Block diagram of model based fault diagnosis scheme

The performance of the proposed fault diagnosis scheme is investigated on a lumped parameter model [1] in which complete dynamical model of an air-conditioning system has been considered. This dynamic model has been developed based on principles of mass and energy conservation. The dynamic mathematical model can be represented in state space form.

$$\dot{x}=g(x,u,d) \tag{1}$$

$$y=h(x,u) \tag{2}$$

where $x=[x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \quad x_8 \quad x_9 \quad x_{10} \quad x_{11} \quad x_{12}]'$ are the states. Control input vector is given by u , disturbance vector is d and output vector is y

$$u=[u_1 \quad u_2 \quad u_3 \quad u_4]' \quad d=\left[d_1 \quad d_2 \quad d_3 \quad d_4 \quad d_5\right]' \quad y=[y_1 \quad y_2 \quad y_3 \quad y_4]'$$

The details of the model is presented in [1]. The general scheme of residual generator is given below:

Consider a state-space realization of the continuous-time plant model given by

$$\dot{x}(t) = Ax(t) + Bu(t) + Ed(t) \tag{3}$$

$$y(t) = Cx(t) + Du(t) + Fd(t) \tag{4}$$

Where $d(t)$ is the fault vector to be detected. In general, $d(t)$ consists of faults which may occur in the actuators, the plant dynamics (components), or the sensors. The initial state $x(0)$ is assumed to be zero. Taking the s-transform of Eq.(3)-(4) gives

$$y(s) = G_u(s)u(s) + G_d(s)d(s) \tag{5}$$

where $G_u(s) = C(sI - A)^{-1}B + D$ and $G_d(s) = C(sI - A)^{-1}E + F$

A residual generator can be expressed in a general form as

$$r(s) = F(s)u(s) + H(s)y(s) \tag{6}$$

with the properties that for all $u(s)$,

i) $r_i(s) = 0$ if $d_i(s) = 0$ and (7)

ii) $r_i(s) \neq 0$ if $d_i(s) \neq 0, i = 1,2,..........,q$ (8)

where $F(s)$ and $H(s)$ are stable and proper transfer matrices that are realizable in a real system. When no failure occurs, all the residuals are equal or close to zero. In presence of *i-th* failure signal, residual signal will become distinguishably nonzero.

3 Fault Detection Based on Residual Analysis

3.1 Residual Analysis

Residual analysis is an easy way to study deviation of actual value from normal value. The residual can be computed by taking the difference between state variable under normal and faulty condition. Figs.2-3 show the residuals for various faults. In the above simulation the threshold are set at ± 2°C, ±0.5 bar, ±0.5 bar, -1.25°C, -5% for evaporator wall temperature, evaporator pressure, condenser pressure, coach air temperature and relative humidity respectively. If the residual between the system output and actual output exceed the threshold, a fault alarmed.

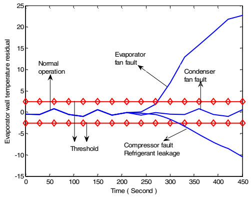


Fig. 2. Evaporator wall temperature residual

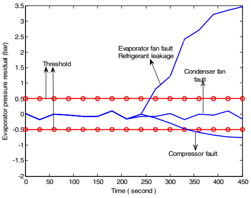


Fig. 3. Evaporator pressure residual

3.2 Fault Decision Table

In this section we will discuss different residual for different faults. The residual patterns for different faults are shown in Table 1. These residuals have been generated by using an air-conditioning unit provided in Indian Railway passenger coach faults.

r_1 = Evaporator wall temperature, r_2 = Thermal space temperature, r_3 = Absolute humidity of thermal space, r_4 = Evaporator Pressure, r_5 = Condenser pressure, r_6 = Condenser wall temperature, r_7 = Speed of compressor, r_8 = Air flow rate.

Fault 0: No fault, Fault 1: Evaporator fan speed less/ Air filters are choked wall temperature sensor defective. Fault 2: Condenser fan speed less. Fault 3: Compressor defective. Fault 4: Refrigerant leakage. Fault 5: Evaporator wall temperature sensor defective. Fault 6: Thermal space temperature sensor defective. Fault 7: Relative humidity sensor defective. Fault 8: Evaporator pressure sensor defective. Fault 9: Condenser pressure sensor defective. Fault 10: Condenser temperature sensor defective. Fault 11: Compressor speed sensor defective. Fault 12: Air flow rate sensor defective.

Table 1. Fault symptoms

Fault modes	Residuals							
	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
0	→	→	→	→	→	→	→	→
1	↓	↑	↑	↓	↓	→	→	↓
2	→	↑	↑	→	↑	↑	→	→
3	↑	↑	↑	↑	↓	↓	↓	→
4	↑	→	→	↑	↑	↑	→	→
5	↑↓	→	→	→	→	→	→	→
6	→	↑↓	→	→	→	→	→	→
7	→	→	↑↓	→	→	→	→	→
8	→	→	→	↑↓	→	→	→	→
9	→	→	→	→	↑↓	→	→	→
10	→	→	→	→	→	↑↓	→	→
11	→	→	→	→	→	→	↑↓	→
12	→	→	→	→	→	→	→	↑↓

4 Fault Diagnosis Using LS-SVM Classifier

The patterns of residuals given in Table 1 can be classified by using some classifier. In this paper a LSSVM classifier is designed to diagnose air-conditioning faults. In case of LS-SVM, original Vapnik's SVM classifier formulation has been modified as follows:

$$\min_{w,b,e} \Im(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2$$

subject to the equality constraints

$$y_k [w^T \varphi(x_k) + b] = 1 - e_k, \quad \text{k=1,.....,N} \,. \tag{9}$$

The important differences with standard SVMs are the equality constraints and the sum squared error term. Then the Lagrangian of the system can be constructed like:

$$\ell(w,b,e;\alpha)=\Im(w,e)-\sum_{k=1}^{N}\alpha_k\{y_k[w^T\varphi(x_k)+b]-1+e_k\} \tag{10}$$

where α_k are the Lagrange multipliers. The condition of optimality can be given by

$$\begin{bmatrix}0 & Y^T\\ Y & \vartheta+\gamma^{-1}I\end{bmatrix}\begin{bmatrix}b\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\ I\end{bmatrix} \tag{11}$$

$$\begin{aligned}&Z=[\varphi(x_1)^T y_1,........,\varphi(x_N)^T y_N]\\ &Y=[y_1,.........y_N],I=[1,.......1]\\ &\alpha=[\alpha_1,.......\alpha_N]\ and\ \vartheta\text{=ZZ}^{\text{T}}\end{aligned} \tag{12}$$

or, $$\vartheta_{kl}=y_k y_l\varphi(x_k)^T\varphi(x_l)=y_k y_l K(x_k,x_l) \tag{13}$$

$K(.,.)$ is the kernel function. LS-SVM classifier is constructed as follows:

$$y(x)=sign[\sum_{k=1}^{N}\alpha_k y_k K(x,x_k)+b]\ . \tag{14}$$

5 Results and Discussions

In this paper, in order to achieve better on-line FDD performance, a heuristic algorithm using LSSVM is used to classify 12 different faults. Since the LS-SVM classifier presented in Fig. 4 can only be used to deal with three class case, a multi-layer LS-SVM frame work has to be designed for the FDD problem with various

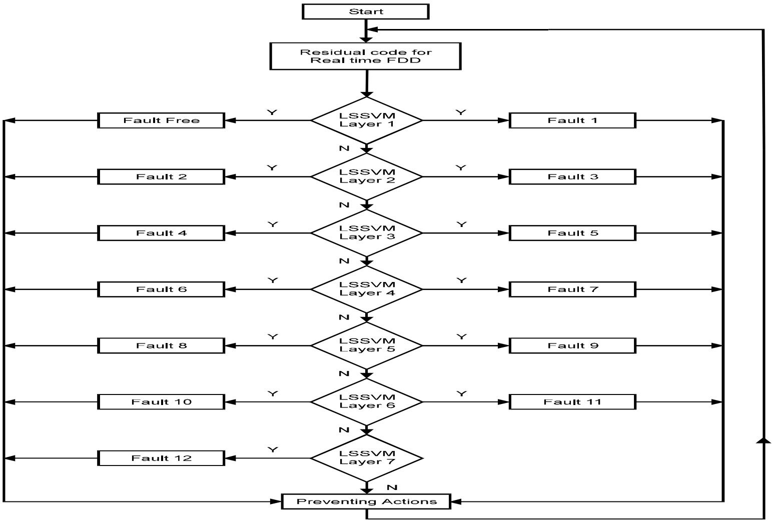


Fig. 4. Flow chart of the seven –layer LS-SVM classifier

faulty positions. It is interesting to know that for conventional neural network classifiers, the architecture and neuron number must be properly designed to achieve high classification accuracy. But for the SVM classifier, the key is to choose a proper kernel function. In general, linear function, polynomial function, radial basis function (RBF), sigmoid function etc. can be adopted as the kernel function. In this paper, polynomial function is used in the designed LSSVM classifier. as it has excellent performance in many applications.

The performance of this FDD method is measured in terms of diagnosis accuracy. The steady state variable data between 4 and 10 h are used to build the seven layer SVM classifier, the data with in the threshold under the normal condition indicate fault free, and the data beyond the threshold indicate faults 1-12. For each normal/ faulty condition, 256×12 groups of sample data are used, in which 256 groups are used for each fault degree. Therefore a total of 8(condition)$\times256\times$ 12(samples) are collected. The LSSVM classifier has been tested by using data other than training data, in absence as well as presence of noise. It has been found that LSSVM classifier is able to diagnose all twelve faults.

6 Conclusions

This paper proposes a fault diagnosis schemes having model based approach for diagnosing major faults in an air-conditioning system. In model based fault diagnosis using LS-SVM classifier very good diagnosis accuracy has been achieved.

(i) The most essential condition for achieving good accuracy is that model should be accurate. The inaccuracy in model will directly reduce fault diagnosis accuracy.

(ii) Fault diagnosis accuracy also depends upon threshold selection.

(iii)The practical systems have always noises and robustness of these fault diagnosis schemes against sensor noises needs to be examined.

References

[1] Kumar, M., Kar, I.N., Ray, A.: State Space based Modelling and Performance Evaluation of air-conditioning system. International Journal of HVAC & R Research 14(5) (September 2008)

[2] Liang, J., Du, R.: Model-based fault detection and diagnosis of HVAC systems using Support Vector Machine method. International Journal of refrigeration 30, 1104–1114

[3] Pau-Lo-Hsu, Lin, K.-L., Shen, L.-C.: Diagnosis of Multiple Sensor and Actuator Failures in Automotive Engines. IEEE Transactions on Vehicular Technology 44(4) (November 1995)

[4] Stoecker, W.F., Jones, J.W.: Refrigeration & Air conditioning. McGraw-Hill Book Company, New York (1982)

[5] The Math Works, Inc., MATLAB. The Math Works, Inc., Natick, Massachusetts (2007)

[6] Katipamula, S., Brambley, M.R.: Methods for fault detection, diagnostics, and prognostics for building systems-a review, part I. International Journal of HVAC&R Research 11, 3–25 (2005)

[7] Du, R.: Monitoring and Diagnosis of Sheet Metal Stamping Processes. In: Gao, R., Wang, L.H. (eds.) Condition-based Monitoring and Control for Intelligent Manufacturing. Springer, New York (2005)

Classification of Power Quality Disturbances Using GA Based Optimal Feature Selection

K.R. Krishnanand[1], Santanu Kumar Nayak[1], B.K. Panigrahi[2], V. Ravikumar Pandi[2], and Priyadarshini Dash[3]

[1] Department of Electrical Engineering, Silicon Institute of Technology, Bhubaneswar, India
[2] Department of Electrical Engineering, Indian Institute of Technology, Delhi, India
[3] Department of Electrical Engineering, NIT Durgapur, India
krishkr09@gmail.com, santanu.nayak87@gmail.com, bkpanigrahi@ee.iitd.ac.in, ravikumarpandi@gmail.com, priyadarshini.nitdgp@gmail.com

Abstract. This paper presents a novel technique for power quality disturbance classification. Wavelet Transform (WT) has been used to extract some useful features of the power system disturbance signal and Gray-coded Genetic Algorithm (GGA) have been used for feature dimension reduction in order to achieve high classification accuracy. Next, a Probabilistic Neural Network (PNN) has been trained using the optimal feature set selected by GGA for automatic Power Quality (PQ) disturbance classification. Considering ten types of PQ disturbances, simulations have been carried out which show that the combination of feature extraction by WT followed by feature reduction using GGA increases the testing accuracy of PNN while classifying PQ signals.

Keywords: Gray-coded Genetic Algorithm, Power quality disturbances, Wavelet transform, Probabilistic Neural Network.

1 Introduction

In recent years, power quality has become a significant issue for both utilities and customers. Power quality issues [1] and the resulting problems are the consequences of the increasing use of solid state switching devices, non-linear and power electronically switched loads, unbalanced power systems, lighting controls, computer and data processing equipments as well as industrial plant rectifiers and inverters. A power quality (PQ) problem usually involves a variation in the electric service voltage or current, such as voltage dips and fluctuations, momentary interruptions, harmonics and oscillatory transients causing failure or mal-operation of the power service equipment. Hence to improve power quality, fast and reliable detection of the disturbances and the sources and causes of such disturbances must be known before any appropriate mitigating action can be taken. However, in order to determine the causes and sources of disturbances, one must have the ability to detect and localize these disturbances. In the current research trends in power quality studies, Wavelet transform (WT) [2-4] is widely used in analyzing non-stationary signals for power quality assessment [5,6] . In order to identify the type of disturbance present in the power

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 561–566, 2009.

signal more effectively, several authors have presented different methodologies based on combination of wavelet transform (WT) and artificial neural network (ANN) [7]. Using the multi-resolution properties of WT [6], the features of the disturbance signal are extracted at different resolution levels and are used for the classification purpose. Gaing [8] demonstrated the classification of 7 types of PQ events by using wavelets and probabilistic neural network (PNN) [9].

In this paper, we have tried to decompose the power system disturbance signal up to 13 level of decomposition as mentioned in [8]. Instead of taking only one statistical feature of the decomposition levels, seven different statistical measures like energy, entropy, standard deviation, mean, kurtosis. ITD and skewness are applied to each of the decomposition level. This leads to increase the feature set length up to 91. It is generally well known that as the length of the feature set increases, the classification accuracy increases but the computational complexity becomes very high. Hence, an attempt has been made to reduce the dimensionality of the feature space by finding the optimal feature set using gray coded Genetic Algorithm [12].

2 Wavelet Transform

The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently. The DWT is calculated based on two fundamental equations: the scaling function $\phi(t)$, and the wavelet function $\psi(t)$, where

$$\phi(t) = \sqrt{2}\sum_k h_k \phi(2t-k) \ . \tag{1}$$

$$\psi(t) = \sqrt{2}\sum_k g_k \phi(2t-k) \ . \tag{2}$$

The scaling and wavelet functions are the prototype of a class of orthonormal basis functions of the form

$$\phi_{j,k}(t) = 2^{\frac{j}{2}} \phi(2^j t - k); \quad j,k \in Z \ . \tag{3}$$

$$\psi_{j,k}(t) = 2^{\frac{j}{2}} \psi(2^j t - k); \quad j,k \in Z \ . \tag{4}$$

The parameter j controls the dilation or compression of the function in time scale and amplitude. The parameter *k* controls the translation of the function in time. *Z* is the set of integers. Once a wavelet system is created, it can be used to expand a function *f(t)* in terms of the basis functions

$$f(t) = \sum_{l \in Z} c(l)\phi_l(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{\infty} d(j,k)\ \psi_{j,k}(t) \ . \tag{5}$$

Where, the coefficients $c(l)$ and $d(j,k)$ are calculated by inner product as

$$c(l) = \langle \phi_l \mid f \rangle = \int f(t)\phi_l(t)dt \ . \tag{6}$$

$$d(j,k) = \langle \psi_{j,k} \mid f \rangle = \int f(t)\psi_{j,k}dt \ . \tag{7}$$

The expansion coefficients *c(l)* represent the approximation of the original signal *f(t)* with a resolution of one point per every 2^J points of the original signal. The expansion coefficients *d(j,k)* represent details of the original signal at different levels of resolution. *c(l)* and *d(j,k)* terms can be calculated by direct convolution of *f(t)* samples with the coefficients h_k and g_k , which are unique to the specific mother wavelet chosen. Daubechies wavelet family is one of the most suitable wavelet families in analyzing power system transients In the present work, the db4 wavelet has been used as the wavelet basis function for power quality disturbance detection and classification.

3 Feature Extraction and Reduction

3.1 Feature Extraction

Data preparation is same as the one detailed in [10]. The detailed co-efficient D_{ij} at each decomposition level is used to extract the features. Statistical features like energy, standard deviation, mean, kurtosis, skewness, Shannon entropy and, Instantaneous Transient Disturbance (ITD) of the decomposition coefficients Dij are calculated.

Thus in the present case, with a '13' level of decomposition of the original signal, the feature vector adopted is of length '7*13' i.e., 91. The feature vector is denoted as

$$\text{Feature}= [ED_1ED_2 \ldots ED_{13}\sigma_1\sigma_2 \ldots\ldots\sigma_{13}\mu_1\mu_2 \ldots\mu_{13}KRT_1\ KRT_2 \ldots\ KRT_{13}SK_1\ SK_1 \ldots\ldots \ldots\ldots SK_{13}\ ENT_1ENT_2 \ldots ENT_{13}ITD_1ITD_2 \ldots\ldots ITD_{13}] . \qquad (8)$$

3.2 Gray Coded Genetic Algorithm and Optimal Feature Selection:

Genetic Algorithm (GA) [11] is a well known meta-heuristic algorithm which mimics the evolution process employed by nature for finding the best chromosome. Here, each decision variable provides a value for finding the best vector all together. A variant of GA termed as Gray Coded Genetic Algorithm [12], developed by replacing the binary representation scheme of classical GA with the gray coded binary representation scheme was proposed to remove the problem of premature convergence encountered in binary GA. In this paper a discrete version of GGA is used, wherein variables are allowed to pass through discrete values only; for better performance in problems where the search range consists of discrete values. Chaos Gray coded genetic algorithm (CGGA) is also developed in [13]. Pseudo-code for GGA is given below:

1. Initialization**:** Initialize the chromosome population with *k* random vectors. $\vec{x}_1,\ldots\ldots\ldots\ldots\vec{x}_k$ which are binary coded vectors
2. Evaluation: Each vector is evaluated to obtain the respective fitness value
3. Crossover: The exchange of binary values between chromosomes is done by roulette wheel selection and subsequent crossover.
4. Mutation: The mutation is done by random change of 1 to 0 and 0 to 1.
5. Reinsertion: The new chromosomes thus evolved are compared against the current population and the chromosomes with lesser fitness value in the population are replaced by them.
6. Repeat from step 2 until a certain number of iterations has been performed.

The parameters of the search are

- k, the size of the memory. A typical value is in the order of 10 to 50.
- $P_{crossover}$ the probability of crossover. A typical value is 0.6
- $P_{mutation}$ the probability of mutation. A typical value is 0.01

Here in our problem we have to select some of the features among the whole set in order to get better classification accuracy. In the present work, dimension of each vector of the GGA was taken to be 13 and each component of the vectors are initialized with 13 values out of random permutation of integers between 1 and 91. Thus, the proposed technique allows the algorithm to select a number of features which is equal to 13. Also, during each mutation step the target vectors are allowed to pass through integral values in this region only. As such, the proposed technique searches a small range and hence is expected to show better convergence rate. During the search process the fitness of each vector is evaluated using PNN classification accuracy so that our objective becomes maximization of classification accuracy. Finally, when a stopping criterion is reached (either by end of iterations or through stagnation) components of the vector with the best fitness in the population, gives the optimally selected features.

4 Classification of Disturbances Using Probabilistic Neural Network (PNN)

The PNN model is one among the supervised learning networks. The learning speed of the PNN model is very fast, making it suitable for fault diagnosis and signal classification problems in real time. Fig. 1 shows the architecture of a PNN model that is composed of the radial basis layer and the competitive layer. The learning and recalling processes of the PNN for classification problems can be found in [14].

5 Results and Discussion

Ten classes (C1-C10) of different PQ disturbances are taken for classification and they are as follows:

C1→	Voltage Sag	C6→	Sag with Harmonic
C2→	Voltage Swell	C7→	Swell with Harmonic
C3→	Harmonic Distortion	C8→	Interruption
C4→	Flicker	C9→	Spike
C5→	Oscillatory transients	C10→	Notch

The power quality signals corresponding to these ten classes are generated in Matlab [15] using parameterized models with different parameter values. Wavelet transform of these data samples are then performed to decompose the signals up to 13th level. The statistical features of the decomposed levels constitute the feature vector. Based on the extracted feature, the feature data sets for training and testing are constructed separately. The classification accuracy of the data set is computed using PNN

Table 1. Optimally selected features by GGA

Serial no.	Selected Index	Selected Feature Type	Serial no.	Selected Index	Selected Feature Type
1	24	Std. deviation of 11th level	8	69	Entropy of 4th level
2	40	Kurtosis of 1st level	9	75	Entropy of 10th level
3	41	Kurtosis of 2nd level	10	81	ITD of 3rd level
4	43	Kurtosis of 4th level	11	82	ITD of 4th level
5	56	Skewness of 4th level	12	88	ITD of 10th level
6	57	Skewness of 5th level	13	90	ITD of 12th level
7	60	Skewness of 8th level			

Table 2. Comparison of classification accuracy results

Features	Classification accuracy	Features	Classification accuracy
Energy (13)	98.86	Entropy (13)	96.84
Std. deviation (13)	98.77	ITD (13)	98.59
Mean (13)	85.76	Energy and Entropy (26)	98.60
Kurtosis (13)	98.86	All (91)	93.89
Skewness (13)	97.21	GGA selected features(13)	**99.30**

classifier for automatic classification of PQ events using all 91 features, only energy features, only entropy features and, entropy-energy features combined together, successively. Next, GGA is used to select the optimal feature set amongst available 91 features and overall classification accuracy is next compared with that of the previous cases. 13 features are selected by GGA as shown in table 1 and overall classification accuracy is compared in table 2.

6 Conclusion

In this paper an attempt has been made to improve the overall performance of a PNN classifier using optimal feature selection. GGA algorithm has been used for that purpose and results corroborate the feature extraction mechanism showing high classification accuracy of PQ disturbance at reduced complexity.

Acknowledgement

The authors acknowledge the financial grant by Department of Science and Technology (DST), Govt of India, for the research project on Power Quality Assessment in Distribution Network, to be carried out in EE Department, IIT Delhi (SR/S3/EECE/46/2007).

References

1. Bollen, M.H.J.: Understanding Power Quality: Voltage sags and Interruptions. IEEE Press, NewYork (2000)
2. Daubechies, I.: The wavelet transform, time/frequency location and signal analysis. IEEE Transactions on Information Theory 36, 961–1005 (1990)
3. Mallat, S.G.: A theory of multi resolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 674–693 (1989)
4. Meyer, Y.: Wavelets and Operators. Cambridge University Press, London (1992)
5. Santoso, S., Powers, E.J., Grady, W.M., Hofmann, P.: Power quality assessment via wavelet transform analysis. IEEE Transactions on Power Delivery 11, 924–930 (1996)
6. Gaouda, A.M., Salama, M.M.A., Sultan, M.K., Chikhani, A.Y.: Power Quality Detection and Classification Using Wavelet-Multi resolution Signal Decomposition. IEEE Transactions on Power Delivery 14, 1469–1476 (1999)
7. Santoso, S., Powers, E.J., Grady, W.M., Parsons, A.: Power quality disturbance waveform recognition using wavelet-based neural classifier, Part 1: theoretical foundation. In: The 1997 IEEE/PES Winter Meeting, New York,U.S.A (1997)
8. Gaing, Z.L.: Wavelet-Based Neural Network for Power Disturbance Recognition and Classification. IEEE Trans. on Power Delivery 19, 1560–1568 (2004)
9. Specht, D.F.: Probabilistic neural networks. Neural Networks 3, 109–118 (1990)
10. Panigrahi, B.K., Ravikumar Pandi, V.: Optimal feature selection for classification of power quality disturbances using wavelet packet-based fuzzy k-nearest neighbour algorithm. IET Generation Trans. Distr. 3, 296–306 (2009)
11. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. Springer, Berlin (1999)
12. Chakraborti, N., Mishra, P., Erkoc, S.: A Study of the Cu Clusters Using Gray-Coded Genetic Algorithms and Differential Evolution. Journal of Phase Equilibria and diffusion 25, 16–21 (2004)
13. Yang, X., Yang, Z., Yin, X., Li, J.: Chaos gray-coded genetic algorithm and its application for pollution source identifications in convection–diffusion equation. Comm. In non linear science and numerical simulation 13, 1676–1688 (2008)
14. MATLAB, Math Works, Inc., Natick, MA, USA (2000)

Constrained Control of Weakly Coupled Nonlinear Systems Using Neural Network

Dipak M. Adhyaru[1], I.N. Kar[2], and M. Gopal[2]

[1] Instrumentation and Control Engineering Department, Institute of Technology, Nirma University, Ahmedabad, India-382481
[2] Department of Electrical Engineering, Indian Institute of Technology, Delhi-110016
dipak.adhyaru@nirmauni.ac.in, ink@ee.iitd.ac.in, mgopal@ee.iitd.ac.in

Abstract. In this paper, a new algorithm is proposed for the constrained control of weakly coupled nonlinear systems. The controller design problem is solved by solving Hamilton-Jacobi-Bellman(HJB) equation with modified cost to tackle constraints on the control input and unknown coupling. In the proposed controller design framework, coupling terms have been formulated as model uncertainties. The bounded controller requires the knowledge of the upper bound of the uncertainty. In the proposed algorithm, Neural Network (NN) is used to approximate the solution of HJB equation using least squares method. Necessary theoretical and simulation results are presented to validate proposed algorithm.

Keywords: Weak coupling, HJB equation, Bounded control, Nonlinear system, Lyapunov stability.

1 Introduction

Many physical systems are naturally weakly coupled such as power systems, flexible space structures etc. The weakly coupled linear systems were introduced to the control audience by Kokotovic [1]. Optimal control of weakly coupled system studied with parallel processing in [2-5]. The results of [5] are based on the idea of the recursive reduced-order scheme for solving the algebraic riccati equation of weakly coupled systems [2] bilinear system [6]. Recently Lim and Kim [12] proposed a similar approach for nonlinear systems using Successive Galerkin Approximation (SGA). It is to be noted that in all the above mentioned approaches, it was assumed that (coupling coefficient) ^2=0, to do parallel processing. It is well known that SGA is computationally complex [7]. It is even more difficult to handle constraints on the control input. Due to limitations of the actuators one should consider constraints on the control input. Constrained optimal controller design proposed for nonlinear system using nonquadratic performance function in [8-9]. Khalaf et. al [8] used NN based HJB solution to solve it. Compared to SGA it is less complex and can handle constraints. Also one can avoid assumptions made by earlier work to form parallel processing. However, the weak coupling theory has been studied so far only without constraints on the input. The main contribution in this paper is the constrained controller design

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 567–572, 2009.

of weakly coupled nonlinear system. It can be achieved in the following steps: (1) Terms related to the weak coupling has been treated as uncertainties. So, original constrained controller design for weakly coupled nonlinear system now becomes constrained robust controller design of a nonlinear uncertain system. (2) A constrained optimal control problem is formulated for the nominal dynamics of the nonlinear system. It is solved using HJB equation with modified performance functional to tackle constraints and unknown coupling. The paper is organized as follows: In section 2, controller design problem of weakly coupled nonlinear system is formulated as a robust control problem. Robust – Optimal control framework has been described in section 3. Stability issues are discussed. In section 4, NN based HJB solution is used to find constrained robust-optimal control law. Solution of NN based HJB equation found by least squares method. Numerical example is given in section 5 for the validity of the approach. Proposed work is concluded in section 6.

2 Controller Framework for Weakly Coupled Nonlinear System

Consider a weakly coupled nonlinear system
$\dot{x}_1 = f_{11}(x_1) + g_{11}(x_1)u_1 + \varepsilon\left(f_{12}(x)\right) + \varepsilon(g_{12}(x))u_2$ (1) , $\dot{x}_2 = f_{22}(x_2) + g_{22}(x_2)u_2 + \varepsilon\left(f_{21}(x)\right) + \varepsilon(g_{21}(x))u_1$ (2)
where $x_1 \in \mathbb{R}^{n1}, x_2 \in \mathbb{R}^{n2}, u_1 \in \mathbb{R}^{m1}, u_2 \in \mathbb{R}^{m2}$ and ε is a small positive coupling coefficient. Also $x=[x_1^T \; x_2^T]^T$ is state vector and $u=[u_1^T \; u_2^T]^T$ is the control input vector. Each component of u_i is bounded by a positive constant λ. i.e.,$|u_i| \leq \lambda \in \mathbb{R}$ (3)

We assume that f_{1i}, f_{2i} and g_{ij} are Lipschitz continuous on the set Ω. We also assume that $f_{1i}(0)=0$ and $f_{2i}(0)=0$. Our aim is to design a control law u such that it will stabilize coupled systems defined by (1) and (2). It can be achieved by considering controller design problem of weakly coupled nonlinear system as a controller design problem of nonlinear uncertain system having uncertainties in the form of unknown coupling terms. Rewriting equations (1) and (2) as, $\dot{x}=f(x)+g(x)u+pd(x)+ph(x)u$ (4)
where $f(x)=\left[f_{11}(x_1) \;\; f_{22}(x_2)\right]^T$, $g(x)=[g_{11}(x_1) \;\; g_{22}(x_2)]^T$, $p = \varepsilon$ and $h(x)=[g_{12}(x) \;\; g_{21}(x)]^T$.

Origin is assumed as an equilibrium point of the system (4). It is also assumed that $pd(x)$ and $ph(x)$ are bounded by a known functions, $D_{\max}(x)$ and $H_{\max}(x)$ respectively. i.e. $\|pd(x)\| \leq D_{\max}(x); \|ph(x)\| \leq H_{\max}(x)$ (5)

In this paper we seek a constrained optimal control that will compensate for the uncertainty related to p.

(A) Robust control problem:

For the open loop system (4), find a feedback control law $u = K(x)$ such that the closed-loop system is asymptotically stable for all admissible uncertainties p.

This problem can be formulated into an optimal control of the nominal system with appropriate cost functional.

(B) Optimal control problem:

For the nominal system

$$\dot{x} = f(x) + g(x)u \tag{6}$$

find a feedback control $u = K(x)$ that minimizes the cost functional

$$\int_0^\infty (\mu_1 D_{\max}^2(x) + \mu_2 H_{\max}^2(x) + x^T Q x + M(u))dt \cdot \tag{7}$$

where $M(u) = 2\int_0^u \lambda \tanh^{-1}(v/\lambda) R dv = 2\lambda u R \tanh^{-1}(u/\lambda) + \lambda^2 R \ln(1 - u^2/\lambda^2) > 0$ (8)

is non-quadratic term expressing cost related to constrained input. The matrices Q and R are positive definite matrices showing the weightage of system states and control inputs, respectively. $\mu_i > 0$, act as a design parameters.

In this paper, we addressed the following problems:
1. Solutions of the problem (A) and (B) are equivalent.
2. Solve the optimal control problem using NN.

To solve the optimal control problem, let $V(x_0) = \min_u \int_0^\infty (\mu_1 D_{\max}^2(x) + \mu_2 H_{\max}^2(x) + x^T Q x + M(u))dt$ be the minimum cost of bringing (6) from x_0 to equilibrium point 0. The HJB equation gives $\min_u (\mu_1 D_{\max}^2(x) + \mu_2 H_{\max}^2(x) + x^T Q x + M(u) + V_x^T(f(x) + g(x)u)) = 0$ (9)
where $V_x = \partial V(x)/\partial x$. It is assumed that $V(x)$ is only a function of x. If $u = K(x)$ is the solution to optimal control problem then according to Bellman's optimality principle [9], it can be found by solving following HJB equation:

$$\mathrm{HJB}(V(x)) = \mu_1 D_{\max}^2(x) + \mu_2 H_{\max}^2(x) + x^T Q x + M(u) + V_x^T(f(x) + g(x)u) = 0 \tag{10}$$

The optimal control law is computed by solving $\partial \mathrm{HJB}(V(x))/\partial u = 0$ (11)

It gives, $u = K(x) = -\lambda \tanh\left(0.5(\lambda R)^{-1} V_x^T g(x)\right)$ (12)

With this basic introduction, following result is stated to show the equivalence of the solution of robust and the solution of optimal control problem.

Theorem 1: Consider the nominal system (6) with the performance function (7). Assume that there exit a solution of HJB equation (10). Using this solution, (12) ensures asymptotic closed loop stability of (4) if following condition is satisfied:

$$\mu_1 D_{\max}^2(x) + \mu_2 H_{\max}^2(x) \geq \left\|V_x^T(x) D_{\max}(x)\right\|^2 + \left\|V_x^T(x) H_{\max}(x)\right\|^2 + \lambda^2 \tag{13}$$

Proof: Here $u = K(x)$ is an optimal control law defined by (12) and $V(x)$ is the optimum solution of HJB equation (10). We now show that the equilibrium point system (4) is asymptotically stable for all possible uncertainties $p(x)$. To do this it is shown that $V(x)$ is a Lyapunov function. Clearly, $V(x)$ is a positive definite function i.e., $V(x) > 0,\ x \neq 0$ and $V(0) = 0$

Using equations (5), (10) and (13), $\dot{V}(x) = (\partial V/\partial x)^T (dx/dt) \leq -x^T Q x \leq 0$.

Thus conditions for Lyapunov local stability theory are satisfied. □

Hence by knowing the exact solution of HJB equation, one can find robust control law in the presence of uncertainties which eventually an optimal control of a weakly coupled nonlinear systems (1) and (2). Theorem 1 is valid if we know the exact solution of HJB equation, which is difficult problem. In the next section NN is used to approximate value-function V.

3 NN Based Robust-Optimal Control

It is well known that an NN can be used to approximate smooth time-invariant functions on prescribed compact sets [13]. Let $\mathbb{R}$ denote the real numbers. Given $x_k \in \mathbb{R}$, define $x=[x_0,x_1,...x_n]^T$, $y=[y_0,y_1,...y_m]^T$ and weight matrices $W_L=[w_1\ w_2 \cdots w_L]^T$. Then the ideal NN output can be expressed as $y=W_L^T\sigma_L(x)$ with $\sigma_L(x)=[\sigma_1(x),\sigma_2(x),...,\sigma_L(x)]^T$ as the vector of basis function. Let NN structure be defined as $\hat{V}(x,t)=\sum_{j=1}^{L} w_j\sigma_j(x)=W_L^T\sigma_L(x)$ (14)

It gives $\hat{V}_x(x)=\partial\hat{V}/\partial x=W_L^T\,\partial\sigma_L(x)/\partial x=W_L^T\nabla\sigma_L(x)$ (15)

$\sigma(x)$ is selected such that $\hat{V}(0)=0$ and $\hat{V}(x)>0$ for $\forall\ x\neq 0$. With this background, we propose NN based robust-optimal control framework in the next section.

3.1 NN Based Controller Formulation

Using (14) and (15) NN based HJB equation can be written as

$$\text{HJB}(\hat{V}(x))=\mu_1 D_{\max}^2(x)+\mu_2 H_{\max}^2(x)+x^TQx+M(\hat{u})+\hat{V}_x^T(f(x)+g(x)\hat{u})=e \quad (16)$$

Khalaf et. al. [8] proved existence of NN based HJB solution. The existence of it for modified performance functional can be proved in the same line of it. So, (16) can be written as $\text{HJB}(\hat{V}(x))=\mu_1 D_{\max}^2(x)+\mu_2 H_{\max}^2(x)+x^TQx+M(\hat{u})+\hat{V}_x^T(f(x)+g(x)\hat{u})\approx 0$ (17)

The optimal control law can be found by taking derivative of (17) w.r.to $\hat{u}$. It can be found as $\hat{u}(x)=-\lambda\tanh\left(0.5(\lambda R)^{-1}g(x)^T\hat{V}_X\right)=-\lambda\tanh\left(0.5(\lambda R)^{-1}g(x)^T W\nabla\sigma^T(x)\right)$ (18)

It is an optimal control law defined by (18) and $\hat{V}(x)$ is the solution of the HJB equation (17). We can show that with this control, the system remains asymptotically stable for all possible p. Using (14) and $\sigma(x)$, $\hat{V}(0)=0$ and $\hat{V}(x)>0$ for $\forall\ x\neq 0$. Also $\dot{\hat{V}}(x)=d\hat{V}/dt<0$ for $x\neq 0$, can be proved similarly as theorem 1 by replacing $V(x)$ by $\hat{V}(x)$ if condition (19) is satisfied. $\mu_1 D_{\max}^2(x)+\mu_2 H_{\max}^2(x)\geq\left\|\hat{V}_x^T(x)D_{\max}(x)\right\|^2+\left\|\hat{V}_x^T(x)H_{\max}(x)\right\|^2+\lambda^2$ (19)

From the above results, it can be proved that NN based robust control stabilizes system having uncertainty. Hence controller designed by (18) stabilizes the weakly coupled systems (1) and (2). Next section is about the utilization of the least squares method for finding a HJB solution.

4 HJB Solution by Least-Square Method

The unknown weights are determined by projecting the residual error e onto de/dW and setting it to zero using the inner product, i.e. $\langle de/dW, e\rangle=0$ for $\forall x\in\Omega\subseteq\mathbb{R}^n$ (20)

This method can be applied to solve robust-optimal control problem for the system having matched uncertainties. According to this method, by using definitions (14),(15) and (16), we can write (20) as

$$\left\langle\nabla\sigma(x)(f(x)+g(x)\hat{u}),\nabla\sigma(x)(f(x)+g(x)\hat{u})\right\rangle W+\left\langle\mu_1 D_{\max}^2(x)+\mu_2 H_{\max}^2(x)+x^TQx+M(\hat{u}),\nabla\sigma(x)(f(x)+g(x)\hat{u})\right\rangle=0 \quad (21)$$

Hence weight updating law is

$$W=-\left\langle\nabla\sigma(x)(f(x)+g(x)\hat{u}),\nabla\sigma(x)(f(x)+g(x)\hat{u})\right\rangle^{-1}\left\langle\mu_1 D^2_{\max}(x)+\mu_2 H^2_{\max}(x)+x^T Qx+M(\hat{u}),\nabla\sigma(x)(f(x)+g(x)\hat{u})\right\rangle \quad (22)$$

By solving this equation, one can find control law using (18) which is the solution of controller design problem of weakly coupled systems. In the next section, simulation carried out on three uncertain systems to validate proposed algorithm.

5 Simulation Experiments

Consider the a weakly coupled nonlinear systems (1)and (2) with terms defined as, $x_1=[x_{11} \quad x_{12}]^T$, $x_2=[x_{21} \quad x_{22}]^T$, $u=[u_1 \quad u_2]^T$, $f_{11}(x_1)=\left[-1.93x_{11}^2 \quad 1.394x_{11}x_{12}\right]^T$, $f_{12}(x)=[0 \quad -4.2x_{21}x_{22}]^T$, $f_{21}(x)=\left[-1.3x_{12}^2 \quad 0.95x_{11}x_{21}-1.03x_{12}x_{22}\right]^T$, $f_{22}(x_2)=\left[-0.63x_{21}^2 \quad 0.413x_{21}-0.426x_{22}\right]^T$, $g_{11}(x_1)=\left[-1.274x_{11}^2 \quad 0\right]^T$, $g_{12}(x)=[0 \quad -6.5x_{22}]^T$, $g_{21}(x)=[0.75x_{11} \quad 0]^T$, $g_{22}(x_2)=[-0.718x_{21} \quad 0]^T$. Control input is bounded by $|u_i|\le 5,$ For simplicity let assume that $\varepsilon=0.1$.Clearly, $\left|\varepsilon\left(f_{12}(x)\right)\right|\le\left[0 \quad x_{21}^2+x_{22}^2\right]^T$, $\left|\varepsilon\left(g_{12}(x)\right)\right|\le\left[0 \quad x_{22}^2\right]^T$ $\left|\varepsilon\left(f_{21}(x)\right)\right|\le\left[x_{12}^2 \quad x_{11}^2+x_{12}^2+x_{21}^2+x_{22}^2\right]^T$, and $\left|\varepsilon\left(g_{21}(x)\right)\right|\le\left[x_{11}^2 \quad 0\right]^T$.

It gives $D_{\max}(x)=\left[0 \quad x_{21}^2+x_{22}^2 \quad x_{12}^2 \quad x_{11}^2+x_{12}^2+x_{21}^2+x_{22}^2\right]^T$ and $H_{\max}(x)=\left[0 \quad x_{22}^2 \quad x_{11}^2 \quad 0\right]^T$.

Our aim is to find the optimal control law that will stabilize the weakly coupled nonlinear system for all possible ε. For the nominal system, $\dot{x}_1=f_{11}(x_1)+g_{11}(x_1)u_1$ and $\dot{x}_2=f_{22}(x_2)+g_{22}(x_2)u_2$ we have to find a feedback control law $u=K(x)$ that minimizes $\int_0^\infty (D_{\max}^T(x)\mu_1 D_{\max}(x)+H_{\max}^T(x)\mu_2 H_{\max}(x)+x^T Qx+M(u))dt$ where, $\Phi=-\tanh\left(0.5W^T\nabla\sigma(x)B(x)\right)$ Q and R have been taken as identity matrices of appropriate dimensions. $\mu_1=500I_{4\times 4}$ and $\mu_2=100I_{4\times 4}$ have been selected for the simulation purpose. This problem can be solved by using (18) and (22). Here we have selected $\hat{V}(x)=w_1x_1^2+w_2x_2^2+w_3x_3^2+w_4x_4^2+w_5x_1x_2+w_6x_1x_3+w_7x_1x_4+w_8x_2x_3+w_9x_2x_4+w_{10}x_3x_4$.

This is a power series NN of 10 activation functions containing power of the state variable of the system upto 2nd order.

It gives W =[0.54 -38.69 0.18 2857 -6.52 6.95 1.49 46.06 -31.12 60.73].

The optimal control law can be found using (18). It can be observed from figures 1(a) and 1(b) that all the system states converge to the equilibrium point. Control signal remains bounded i.e. $|u_i|\le 5,$ as shown in figure-1(c). Condition (19) is verified and shown in figure 1(d).

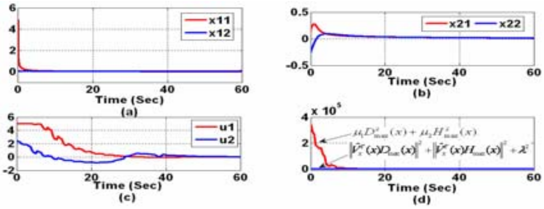


Fig. 1. (a) System (1) states Vs. time, (b) System (2) states Vs. time, (c) Variation of Control input, (d) Verification of condition (19)

6 Conclusions

The contribution of this paper is a methodology for designing constrained controllers for a weakly coupled nonlinear system. It is achieved by formulating the controller design problem of weak coupled nonlinear system into a controller design problem of a nonlinear system having uncertainty in the form of unknown coupling terms. The exact information about uncertainty is not required except some restrictive norm bound. We have adopted NN based HJB solution to design optimal control law that satisfies a prescribed bound. Modifications are done on the earlier approaches to handle constraints on the input and uncertainty related to coupling terms. Simulation results show the good agreement with that of theoretical observations.

References

[1] Kokotovic, P., Perkins, W., Cruz, J., D'Ans, G.: e-coupling for near optimum design of large scale linear systems. Inst. Elect. Eng. Proc. Part D 116, 889–892 (1969)

[2] Gajic, Z., Shen, X.: Decoupling transformation for weakly coupled linear systems. Int. J. Control 50, 1515–1521 (1989)

[3] Gajic, Z., Shen, X.: Parallel Algorithms for Optimal Control of Large Scale Linear Systems. Springer, London (1992)

[4] Aganovic, Z., Gajic, Z.: Optimal control of weakly coupled bilinear systems. Automatica 29, 1591–1593 (1993)

[5] Aganovic, Z., Gajic, Z.: Linear optimal control of bilinear systems: With applications to singular perturbations and weak coupling. Springer, London (1995)

[6] Cebuhar, W., Costanza, V.: Approximation procedures for the optimal control for bilinear and nonlinear systems. J. Optim. Theory Appl. 43(4), 615–627 (1984)

[7] Abu-Khalaf, M., Lewis, F.L.: Nearly optimal state feedback control of constrained nonlinear systems using a neural networks HJB approach. Annual Reviews in Control 28(2), 239–251 (2004)

[8] Abu-Khalaf, M., Huang, J., Lewis, F.L.: Nonlinear H_2/H_∞ constrained feedback control: A practical design approach using neural networks. Springer, Heidelberg (2006)

[9] Gopal, M.: Modern Control System Theory, 2nd edn. New Age International Publishers, New Delhi (1993)

[10] Kim, Y.J., Kim, B.S., Lim, M.T.: Composite control for singularly perturbed nonlinear systems via successive Galerkin approximation. DCDIS, Series B: Appl. Algorithms 10(2), 247–258 (2003)

[11] Kim, Y.J., Kim, B.S., Lim, M.T.: Finite-time composite control for a class of singularly perturbed nonlinear systems via successive Galerkin approximation. Inst. Elect. Eng. Proc.- Control Theory Appl. 152(5), 507–512 (2005)

[12] Kim, Y.J., Lim, M.T.: Parallel Optimal Control for Weakly Coupled Nonlinear Systems Using Successive Galerkin Approximation. IEEE Transactions on Automatic Control 53(6) (2008)

[13] Hornik, K., Stinchcombe, M., White, H.: Universal approximation of an unknown mapping and its derivatives using multilayer feed-forward networks. Neural Network 3, 551–560 (1990)

A Network Coding and Genetic Algorithm Based Power Efficient Routing Algorithm for Wireless Sensor Networks*

Wen-wei Lu, Jian Pan, and Yi-hua Zhu

College of Computer Science and Technology, Zhejiang University of Technology,
Hangzhou, Zhejiang 310023, China
lu.wen.wei@yahoo.com.cn, pj@zjut.edu.cn, yhzhu@ieee.org

Abstract. In a wireless sensor network (WSN), retransmission and acknowledgement (ACK) are required to make reliable packet delivery. In this paper, a Network Coding based Power Efficient Routing (NCPER) algorithm integrated with multi-path routing algorithms is proposed to eliminate retransmission and ACK, which guarantees that the receiving node can decode the original data sent by a source node so that the source node need not care whether the transmitted packets are lost or not. In addition, numeric experiments are conducted to show impacts of packet size and finite field size on energy cost of the NCPER. The NCPER expends energy efficiently and alleviates radio interferences among nodes.

Keywords: Wireless sensor network; routing; network coding.

1 Introduction

In a Wireless Sensor Network (WSN), data packets are delivered to a sink via multi-hop manner. To overcome the problem of packet loss, retransmission and acknowledgement (ACK) are applied in some protocols to make packets delivered reliably. Clearly, the probability that a packet is successfully delivered from a node to the sink declines when the number of communication hops increase, which may cause timeouts or ACK packet loss at some nodes so that these nodes retransmit periodically, i.e., retransmission becomes more seriously.

Most nodes in a WSN are powered by battery with limited energy. Hence, retransmitting and ACK lead nodes to quickly expend their available energy so that lifetimes of the nodes and the WSN are shortened. As a result, reducing number of retransmissions and ACKs is significant for a WSN to run effectively and efficiently.

Multi-path routing[1] is one of effective approaches to enhancing reliability of packet delivery. In addition, network coding [2][3] has the ability of making a

* This project is supported by National Natural Science Foundation of China under grant No. 60873228 and by Zhejiang Provincial Natural Science Foundation of China under grant No. Y1080483.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 573–578, 2009.

destination node recover data packets sent by several source nodes if the destination node receives a sufficient number of coded packets. Moreover, network coding can significantly reduce the number of retransmissions in lossy networks compared to an ARQ (automatic repeat request) scheme [4]. In this paper, we integrate multi-path routing with network coding and propose a routing algorithm called Network Coding based Power Efficient Routing (NCPER) to eliminate retransmission and ACK in the WSNs so that energy is expended efficiently. Under the NCPER, a source node splits data into multiple pieces and chooses a group of linear independent vectors to code these data pieces to generate some new coded packets. Then, the coded packets are transmitted via distinct paths to the sink. The main contribution of the paper are: 1) under the NCPER, retransmission and ACK are not needed; and 2) the NCPER can reduce energy cost and alleviate radio interference.

2 NCPER

We assume there is only one sink in the WSN and it knows the topology of the WSN. We use S to represent a source node in the WSN and assume there are K paths from node S to the sink, denoted by $P(1), P(2), \cdots, P(K)$ respectively, which are constructed by a multipath routing algorithm (e.g., AOMDV [1]). As depicted in Fig.1, each path consists of some wireless links. Besides, we assign triple $(h_i, P_i, E_i(k))$ to $P(i)$, where h_i, P_i and $E_i(k)$ stand for the number of hops, probability of successfully delivering a packet from S to the sink, and energy expended for delivering a k-bit packet to the sink through $P(i)$, respectively. We denote $P(i)$ by $(e_{i,1}, e_{i,2}, \cdots, e_{i,h_i})$, where $e_{i,j}$ stands for the j-th wireless link in $P(i)$, $j = 1, 2, \cdots, h_i$.

In a WSN, Energy consumption is mainly caused by radio communications. Thus, we only consider energy consumption of the radio and ignore other energy consumption in the sensor node. In addition, we apply the same model as proposed in [5], i.e., the energy costs of transmitting and receiving a k-bit packet between two nodes being d meters apart are, respectively, as follows [5]:

$$E_{Tx}(k,d) = k(E_{elec} + \epsilon_{amp} d^{\gamma}) \tag{1}$$

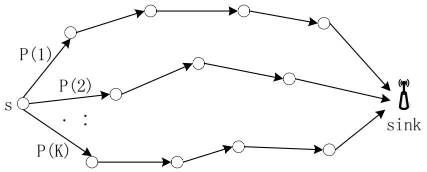


Fig. 1. K paths from S to the sink

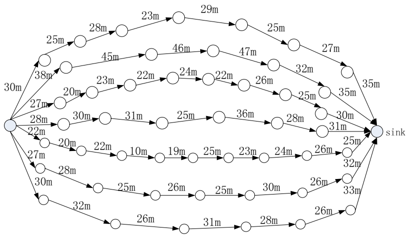


Fig. 2. Multiple paths from source node S to the sink

$$E_{Rx}(k) = kE_{elec} \tag{2}$$

where $\gamma \in [2, 4]$ is the path loss exponent; E_{elec} denotes the energy consumption due to digital coding, modulation, filtering, and spreading of the signal, etc., and ϵ_{amp} is the energy consumed by the transmitter power amplifier.

Assume failure probability of transmitting a packet on wireless link $e_{i,j}$ is p, i.e., successful probability is $1 - p$, where $i = 1, 2, \cdots, K, j = 1, 2, \cdots, h_i$. Let $H(i)$ be the maximum number of hops of a packet being transmitted on $P(i)$ before it is lost. Then, we have the following

$$p_i = (1 - p)^{h_i}, i = 1, 2, \cdots, K$$

$$Pr\{H(i) = n\} = \begin{cases} p(1-p)^n, 0 \leq n < h_i \\ (1-p)^{h_i}, n = h_i \end{cases}$$

$$\overline{H}(i) = \sum_{n=0}^{h_i} nPr\{H(i) = n\} = \frac{1 - (1-p)^{h_i+1}}{p} - 1 \tag{3}$$

where $\overline{H}(i)$ is the average number of hops. Thus, the energy cost for a k-bit packet to reach the sink from node S through $P(i)$ is as follows:

$$E_i(k) = \sum_{j=1}^{\overline{H}(i)} [E_{Tx}(k, d(e_{i,j})) + E_{Rx}(k)]. \tag{4}$$

Assume the message to be delivered to the sink from the source node S is composed of a string of characters from finite field $\mathbb{F}_{2^n}$[6]. Although data are generated continuously by sensor nodes, they are collected periodically in some applications. Thus, we consider the situation that there is a fixed number of packets at a sensor node to be delivered to the sink. Especially a node only sends N blocks of data, say $\alpha_1, \alpha_2, \cdots, \alpha_N$, once a time. Assume each block consists of m characters from $\mathbb{F}_{2^n}$. Additionally, node S produces the following coded data block (CDB) based on network coding:

$$\beta_i = r_{i1}\alpha_1 + r_{i2}\alpha_2 + \ldots + r_{iN}\alpha_N \tag{5}$$

where $r_i \equiv (r_{i1}, r_{i2}, \cdots, r_{iN})$ is termed as *i-th coding vector*(CV) and $r_{ij} \in \mathbb{F}_{2^n}$.

As soon as the sink receives N linearly independent CDBs $\beta_1, \beta_2, \cdots, \beta_N$, we, from (5), have $(\beta_1, \beta_2, \cdots, \beta_N)^T{=}R(\alpha_1, \alpha_2, \cdots, \alpha_N)^T$,where $R{=}(r_1, r_2, \cdots, r_N)^T$. Thus, the sink can recover the original data as follows.

$$(\alpha_1, \alpha_2, \cdots, \alpha_N)^T = R^{-1}(\beta_1, \beta_2, \cdots, \beta_N)^T \tag{6}$$

Obviously, solution of (6) exists if matrix R is full rank, i.e., the CVs $r_1, r_2, \cdots, r_N$ are linearly independent.

3 Combinational Optimization with the NCPER

The key of the NCPER is an optimization problem to find an optimal pair (L, N) such that all of the following are satisfied:

(i) $N \leq L \leq K$;

(ii) L CDBs are generated by node S such that arbitrary N of them are linearly independent;

(iii) The sink is guaranteed to receive at least N CDBs;

(iv) The energy cost per bit (ECPB) for delivering N CDBs through L distinct paths is minimized.

Condition (ii) can be easily met by taking the following four steps: 1) limit $L \leq 2^n$;2) take L different elements $\theta_1, \theta_2, \cdots, \theta_L$ in $\mathbb{F}_{2^n}$; 3) let $r_i \equiv (1, \theta_i, \theta_i^2, \cdots, \theta_i^{N-1})$, $i = 1, 2, \cdots, N$; and 4) generate CDBs as follows:

$$\beta_i = 1\alpha_1 + \theta_i\alpha_2 + \theta_i^2\alpha_3 + \ldots + \theta_i^{N-1}\alpha_N (i = 1, 2, \cdots, L). \tag{7}$$

The reason that arbitrary N of CDBs $\beta_1, \beta_2, \cdots, \beta_L$ are linearly independent is that the CVs of these N CDBs form a full rank Vandermonde determinant valued $\prod_{1 \leq j < i \leq N}(\theta_i - \theta_j) \neq 0$.

Now, we consider the other three conditions. Let $\Phi(x)$ be number of bits x contains. Thus,$\Phi(r_{ij}) = n$ since r_{ij} is in $\mathbb{F}_{2^n}$, and besides,$\Phi(\alpha_i) = nm$ as α_i contains m characters in $\mathbb{F}_{2^n} (i = 1, 2, \cdots, N)$. Moreover, from (6), we have $\Phi(\beta_i) = \Phi(r_{i1}\alpha_1 + r_{i2}\alpha_2 +, \ldots, + r_{iN}\alpha_N)$. Noting that α_i contains m characters in $\mathbb{F}_{2^n}$ and $r_{ij} \in \mathbb{F}_{2^n}$, it can be proved that $r_{i1}\alpha_1 + r_{i2}\alpha_2 + \ldots + r_{iN}\alpha_N$ is also a string of m characters in $\mathbb{F}_{2^n}$.This is because finite field $\mathbb{F}_{2^n}$ has the property that any linear combination of its elements is also in it. As a result, $\Phi(\beta_i) = \Phi(\alpha_i) = nm, i = 1, 2, \cdots, L$.

Clearly, a CDB has to be transmitted with a CV so that the sink is able to recover the original data from N received CDBs. Consequently, a packet being transmitted in a wireless link consists of three portions: the packet header, a CDB, and a CV. We ignore the number of bits the header contains and assume the size of a packet is B. Then, the number of bits a packet contains is $\Phi(\beta_i) + \Phi(r_i) = \Phi(\beta_i) + N\Phi(r_{ij}) = n(m + N)$, i.e.,

$$n(m + N) = B \tag{8}$$

Let x_i be an indicator: $P(i)$ is chosen for routing a packet if $x_i = 1$, otherwise if $x_i = 0$. Thus, the number of packets successfully reaching the sink through $P(i)$ is $x_i p_i$. Besides,we have $L = \sum_{i=1}^{K} x_i$. To guarantee that there are N packets arriving at the sink, the following should hold:

$$N \leq \sum_{i=1}^{K} x_i p_i \tag{9}$$

From (8), the number of bits of payload in N packets is $N\Phi(\alpha_i) = Nnm = N(B - nN)$. Hence, from (4),we have ECPB as follows.

$$E_{bit}(X) = \frac{\sum_{i=1}^{K} x_i E_i(B)}{N(B - nN)} = \frac{\sum_{i=1}^{K} x_i \sum_{j=1}^{\overline{H}(i)} [E_{Tx}(B, d(e_{i,j})) + E_{Rx}(B)]}{N(B - nN)} \tag{10}$$

Therefore, the optimization problem earlier mentioned is equivalent to the following combination optimization problem (COP):

$$Min\{\frac{\sum_{i=1}^{K} x_i \sum_{j=1}^{\overline{H}(i)} [E_{Tx}(B,d(e_{i,j}))+E_{Rx}(B)]}{N(B-nN)}\}$$
$$s.t. \begin{cases} N = \lfloor \sum_{i=1}^{K} x_i p_i \rfloor \\ x_i \in \{0,1\}, i = 1,2,\cdots,K \end{cases} \tag{11}$$

where $\lfloor . \rfloor$ is the floor function. We adopt a genetic algorithm (GA) to solve the COP. A chromosome is defined as $\xi \equiv x_1 x_2 \cdots x_K$, where $x_i \in \{0,1\}, \xi \in \Omega \equiv \{0,1\}^K$. Moreover, we define fitness as

$$f(\xi) \equiv \frac{1}{E_{bit}(X)} = \frac{N(B-nN)}{\sum_{i=1}^{K} x_i \sum_{j=1}^{\overline{H}(i)} [E_{Tx}(B,d(e_{i,j})) + E_{Rx}(B)]} \tag{12}$$

The size of population is set to *M*. The initial population is generated by producing *M* chromosomes, with each being a *K*-bit binary string. Selection operation is performed by calculating fitness in (12) for each chromosome and choosing a chromosome with probability $f(\xi)/\sum_{\eta \in \Omega} f(\eta)$ according to roulette wheel. Crossover operation is based on single random cross point with probability p_c. For instance, for two chromosomes $\xi = x_1 x_2 \cdots x_K$ and $\eta = y_1 y_2 \cdots y_K$, crossover operation make two new chromosomes as $\xi^{'} = x_1 x_2 \cdots x_i y_{i+1} y_{i+2} \cdots y_K$ and $\eta^{'} = y_1 y_2 \cdots y_i x_{i+1} x_{i+2} \cdots x_K$, if the cross point is $i (1 \leq i < K)$. Mutation is performed by choosing a gene in a chromosome with probability p_m and changing the gene by xoring it with 1.

The NCPER is based on multiple disjoint paths, which are constructed by AOMDV [1]. Considering that a destination node in AOMDV knows all the paths from a source node to itself, we solve the COP in (11) by running GA at the sink which has ample power. Then, the optimal *N* together with the optimal *L* paths are sent to source node *S*, in which *N* is used to create CDBs that are delivered through the *L* received distinct paths to the sink. Eventually, the sink decodes the received packets to obtain the original data.

4 Numeric Experiments

Our experiments are based on Fig.2, in which there are 7 paths from source node *S* to the sink. An arrow represents a wireless link between two nodes, and the distance of two nodes is labeled beside the wireless link connecting the two nodes.Following [5], we set $\gamma = 2, E_{elec} = 50nJ/bit$ and $E_{amp} = 10pJ/bit/m^2$. In addition,we take n=8.Moreover, for the GA, we choose $M = 100$, $p_c = 0.8$ and $p_m = 0.05$. The GA evolves 100 generations. Fixing p=0.05, 0.10, 0.15, respectively, and letting *B* vary, lead to Fig. 3, which indicates that 1) for a particular *p*, ECPB decreases as packet size *B* increases, i.e., our NCPER will bring more benefits if a larger *B* is used; and 2) ECPB increases as *p* goes high when B is larger than 128bits.

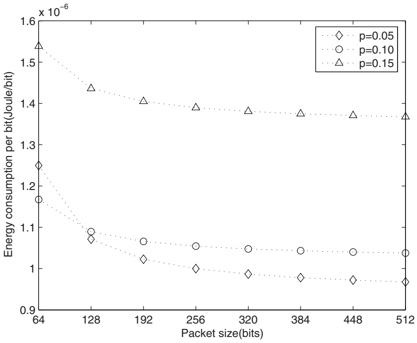


Fig. 3. ECPB vs. B

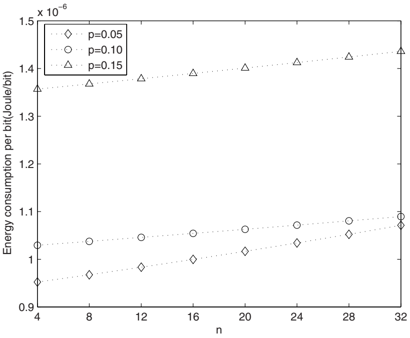


Fig. 4. ECPB vs. n

Setting B=512 and letting n vary, we have Fig. 4, which shows that ECPB increases with n growing. This is because the energy cost for transmitting CVs increases as n increases since more bits are transmitted. In other words, we should choose a small n so that energy cost is reduced. However, it should be stressed that network coding will be impractical if n is too small. For instance, n can not be less than 8, if node S and the sink communicate with symbols consisting of ASCII characters, with each composed of 8bits.

5 Conclusion

The proposed NCPER eliminates retransmission and ACK and is able to guarantee the sink can decode the original data. The numeric experimental results in the previous section show that larger size of packet and smaller size of finite field should be chosen for the NCPER. In future, we will be comparing the NCPER with some well-known multi-path routing algorithms.

References

1. Marina, M.K., Das, S.R.: On-demand Multipath Distance Vector Routing for Ad Hoc Networks. In: Ninth International Conference on Network Protocols, November 11-14, pp. 14–23 (2001)
2. Fragouli, C., Soljanin, E.: Network coding applications. Now Publisher Inc., USA (2007)
3. Ahlswede, R., Cai, N., Li, S.-Y.R., Yeung, R.W.: Network information flow. IEEE Transactions on Information Theory 46, 1204–1216 (2000)
4. Ghaderi, M., Towsley, D., Kurose, J.: Reliability Gain of Network Coding in Lossy Wireless Networks. In: The 27th IEEE Conference on Computer Communications, INFOCOM 2008, April 13-18, pp. 2171–2179 (2008)
5. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for Wireless microsensor networks. IEEE Transactions on Wireless Communications 1, 660–670 (2002)
6. Bartee, T.C., Schneider, D.J.: Computation with finite fields. Information and Computing 6, 79–98 (1963)

Towards Situation Awareness in Integrated Air Defence Using Clustering and Case Based Reasoning

Manish Gupta and Sumant Mukherjee

Institute for Systems Studies and Analyses
Defence Research & Development Organisation
Metcalfe House, Delhi. India 110 054
manish.iitdelhi@gmail.com

Abstract. For integrated air defence in network centric environment, it is required to process information (collected from multiple sensors) for enhancing Situation Awareness (SA) at the command and control nodes. SA is the process of building comprehensive pictures of the battlespace to the decision maker who can further utilized it for threat evaluation. A novel approach for enhancing situation awareness in integrated air defence perspective via Clustering and Case Based Reasoning (CBR) has been proposed in this paper. Clustering is applied on track data generated from Level I of multi sensor data fusion to aggregate entities in the target area. CBR further provide information about air package type, size and purpose of the aggregated entities using cluster attribute records. The effectiveness of the proposed approach has been illustrated on simulation data generated to depict typical integrated defence scenario.

Keywords: Network Centric Warfare, Integrated Air Defence, Situation Awareness, Clustering, Case Based Reasoning.

1 Introduction

Network Centric Warfare (NCW)[2] is the term used in military circles to define information-based war fighting. Network centric battle space can be visualized as network of sensors, platforms and operators (military forces) and Command & Control (C2) entities where free flow of information takes place unlike stovepipe flow of conventional C2 hierarchy.For NCW, data from multiple sources is collected and fused to create the 'Common Operational Picture' (COP)[7,8] using various techniques of Multi-Sensor Data Fusion (MSDF)[10].MSDF is an evolving technology, concerning the problem of how to fuse data from multiple sensors intelligently in order to make a more accurate estimation of the environment. Applications of data fusion cross a wide spectrum, including environment monitoring, automatic target detection and tracking, battlefield surveillance, remote sensing, situation awareness, etc.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 579–584, 2009.

Situation Awareness (SA)[3,4] was envisioned as the main part of Level 2 processing in the JDL model of MSDF[10]. Assuming that object identification has been done at Level I, Level II tries to understand situation by grouping identified objects and interpreting patterns emerging out of this grouping. Essentially four processes take place at this stage object aggregation, event aggregation, context interpretation and multi-perspective assessment. Pattern recognition techniques and automated reasoning techniques are used to implement these processes.

The paper provides a novel approach for enhancing situation awareness in integrated air defence perspective via Clustering[5] and Case Based Reasoning (CBR)[1,6,9]. Clustering is applied outputs generated from level I of MSDF to aggregate entities in the target area. CBR further provide information about unit type, unit size and unit purpose of the aggregated entities using cluster attribute records. The effectiveness of the proposed approach has been illustrated on simulation data generated to depict typical integrated defence scenario.

Section 2 provides an overview of Situation Awareness (SA) and its components. Proposed approach along with application scope of clustering and case based reasoning in SA is described in section 3. Section 4 shows the simulation experiments to illustrate the effectiveness of the proposed approach. Concluding remarks with future scope of the study are given in the last section.

2 Situation Awareness: An Overview

Situation awareness has been formally defined by Endsley[3] as ”the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future”. Situation Awareness (SA) involves being aware of what is happening around you to understand how information, events, and your own actions will impact your goals and objectives, both now and in the near future and deriving relations among entities, e.g., the aggregation of object states (i.e., classification and location).

2.1 Endsley’s Model

Endsley[3] specifies three levels of Situation Awareness namely, Perception, comprehension and projection. The first step in achieving SA is to perceive the status, attributes, and dynamics of relevant elements in the environment. Thus, Level 1 SA leads to an awareness of multiple situational elements (objects, events, people, systems, environmental factors) and their current states (locations, conditions, modes, actions). Level 2 SA integrate disjointed Level 1 information through the processes of pattern recognition to understand how it will impact upon the individual’s goals and objectives. This includes developing a comprehensive picture of the battlefield. Level 3 SA is achieved through knowledge of the status and dynamics of the elements and comprehension of the situation (Levels 1 and 2 SA), and then extrapolating this information forward in time to determine how it will affect future states of the operational environment.

3 Proposed Approach

The paper proposes a unique approach based on clustering and case based reasoning to achieve situation awareness for integrated air defence environment. The flow diagram of the proposed approach for target detection and situation awareness is shown in Figure 1. It is two stage processes in which targets are aggregated geographically using hierarchical clustering[5] and aggregated objects are further evaluated to build common operational picture using CBR. Feature vector is extracted on the output data of Level-I of MSDF. Clusters are formed based on the position of the targets which further generates cluster attribute records. The cluster attribute records are further matched using CBR technique with the existing CBR library to determine air package size, type and purpose. The details of the proposed approach are given in the following subsections.

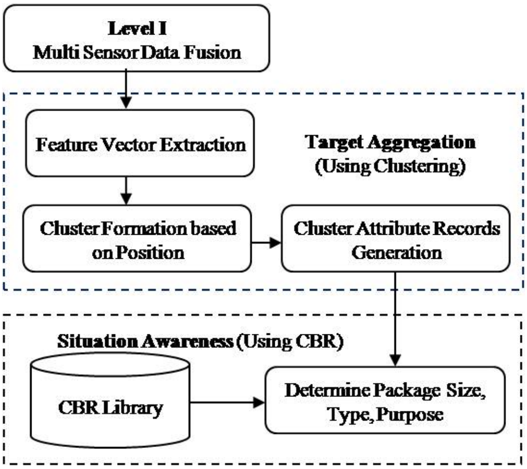

Level I
Multi Sensor Data Fusion
Feature Vector Extraction
Target Aggregation
(Using Clustering)
Cluster Formation based on Position
Cluster Attribute Records Generation
Situation Awareness (Using CBR)
CBR Library
Determine Package Size, Type, Purpose

Fig. 1. Proposed Approach for Target Aggregation and Situation Awareness

3.1 Clustering Algorithm for Targets Aggregation

The feature vector obtained from Level-I of MSDF for each air target contains the information such as (x,y,z)-position components and the class beliefs are between 0.0 and 1.0. The feature vectors for a battlespace are to be updated from tracks at certain times to yield a current set of target feature vectors for SA processing. The clustering of these vectors is done only with respect to position features (x,y,z) in a given area of 20x20x1 kilometers.

Once clusters are for formed as target aggregation using hierarchical clustering[5], cluster attribute records are generated by considering number of individual target type and their average class beliefs as weight for every cluster. Cluster attribute records serve as the problem part for further analysis of Case-based reasoning (CBR). The problem is to be compared to the problem parts of cases in the CBR library to extract a similar case whose solution is in the form of air package type, size, and purpose of the cluster.

3.2 Case Based Reasoning for Building Common Operational Picture

Case-based reasoning (CBR) by Roger Schank[9] in early 1980s is the process of solving new problems based on the solutions of similar past problems. Generally, CBR is regarded as a vital approach when it is difficult to formulate rules describing the situations e.g. car repairing by Mechanic, Lawyer etc. It is a four step process namely retrieve, reuse, revise and retain.

Dynamic memory model[9,6] has been used for case storage in CBR Library. Each case in the SA case-base is made up of problem and solution parts. The new cluster attribute record is compared with the problem parts of cases of CBR library and extract solution of a similar case using case retrieval methods such as nearest neighbor, induction and template retrieval. We have applied the nearest neighbor method, which compares the similarity between the cases of CBR library and the input problem. The case similarity calculation uses the information of quantity proximity, the degree of belief, and the weight of importance of each target class. One simulation study was carried out to illustrate the effectiveness of the proposed approach in integrated air defence scenario.

4 Simulation Results

The proposed approach has been implemented using simulation study. Table 1 shows the feature vector of ten targets obtained from Level I of MSDF. Feature vector shows the positional vector of the targets along with class and its class belief calculated at identity estimation algorithm of Level I. Time and date can also used for dynamic case of creating common operational picture (COP). The last column of the Table 1 shows the cluster index of the targets using hierarchical clustering. The clustering results are described in further subsection.

Table 1. Feature Vector of Tracked Data and Cluster Index

Tgt No.	Time	Date	X	Y	Z	Class	Class Belief	Cluster Index
1	14.01	14/4/2009	2092	3451	15009	2	0.7	1
2	14.01	14/4/2009	2083	3480	14500	1	0.8	1
3	14.01	14/4/2009	2507	3076	15890	2	0.7	1
4	14.01	14/4/2009	2465	2946	16879	1	0.6	2
5	14.01	14/4/2009	2345	3215	12036	2	0.7	3
6	14.01	14/4/2009	2106	3514	14569	3	0.8	1
7	14.01	14/4/2009	2365	3489	13568	1	0.7	4
8	14.01	14/4/2009	2678	3012	15442	2	0.6	1
9	14.01	14/4/2009	1817	3489	17899	3	0.7	5
10	14.01	14/4/2009	1716	3128	16454	2	0.7	2

4.1 Clustering Results

Hierarchical clustering has been applied to obtain clusters of the given targets. Threshold is also obtained the 70% of the maximum distance between the targets. It is observed that five clusters are obtained for the simulation data. Cluster attribute records are further generated by considering number of individual target type and their average class beliefs as weight for every cluster. Table 2 shows the cluster attribute records of the all five clusters obtained using clustering. CBR are further applied on the cluster attribute records for obtaining the best match to determine package type, size and purpose from the CBR library. The subsequent section describes the CBR procedure in detail.

Table 2. Cluster Attribute Records

Type 1 Tgts No.	Type 1 Tgts %	Type 1 Weight	Type 2 Tgts No.	Type 2 Tgts %	Type 2 Weight	Type 3 Tgts No.	Type 3 Tgts %	Type 3 Weight
1	20	0.8	3	60	0.67	1	20	0.8
1	50	0.6	1	50	0.7	0	0	0
0	0	0	1	100	0.7	0	0	0
1	100	0.7	0	0	0	0	0	0
0	0	0	0	0	0	1	100	0.7

4.2 CBR Results

Case Base Reasoning (CBR) Library is required for applying CBR on cluster attribute records. The library records contain two part i.e. problem part and solution part. The problem part contains the same fields given in the cluster attribute records. The solution part consists of package type (e.g. bomber, fighter etc), package size (e.g. single, Multiple & Squadron etc) and purpose (Defensive, Offensive, Surveillance etc.). Table 3 shows the solution part for all five clusters obtained by matching the most similar case of the CBR Library using standard similarity matrix. The CBR results can further be used for creating common operational picture (COP) and threat evaluation at MSDF Level III. It is to be noted that solution part can be different for different application. It can be *Lethality* and *Attack Profile* of the target but similar CBR Library is prerequisite for such study.

Table 3. CBR Results

Cluster No.	Package Type	Package Size	Package Purpose
1	Fighter A/c	Multiple(3)	Offensive
2	Multi Role A/c	Unknown	Surveillance
3	Fighter A/c	Multiple(3)	Offensive
4	Multi Role A/c	Multiple(5)	Defensive
5	Multi Role A/c	Unknown	Surveillance

5 Concluding Remarks

The paper proposes a novel approach of enhancing situation awareness using clustering and case based reasoning (CBR). The use of clustering depicts its utility for target aggregation for Integrated Air Defence in Network Centric Warfare (NCW). The role of CBR highlights to create Common Operational Picture (COP), which is further assessed by commanders for threat evaluation and for taking preventive measures in the battlefield scenario. From the simulation results is can be concluded that proposed approach can be effectively applied for various application in situation awareness. The proposed approach will be applied in the dynamic scenario as a future work.

Acknowledgments. We are highly indebted to Shri H.V. Srinivasa Rao, Director, ISSA for his active support and guidance. We are also thankful to Smt. Sanchita Malik, Scientist E, ISSA and our colleagues for sharing their valuable information in this field.

References

1. Aamodt, A.: Explanation-driven case-based reasoning, Topics in Case-based reasoning, pp. 274–288. Springer, Heidelberg (1994)
2. Ahlberg, S., et al.: An information fusion demonstrator for tactical intelligence processing in network-based defense. Information Fusion 8, 84–107 (2007)
3. Endsley, M.R.: Toward a theory of situational awareness in dynamic systems. Human Factors 37, 32–64 (1995)
4. Feng, H., et al.: Modelling situation awareness for Context-aware Decision Support. Expert Systems with Applications 36, 455–463 (2009)
5. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika 32(3), 241–254 (1967)
6. Kolodner, J.L.: Maintaining Organization in a Dynamic Long-Term Memory. Cognitive Science 7(4), 243–280 (1983)
7. Looney, C.G., et al.: Cognitive situation and threat assessments of ground battlespaces. Information Fusion 4, 297–308 (2003)
8. Looney, C.G.: Exploring fusion architecture for a common operational picture. J. Information Fusion (article in press)
9. Schank, R.: Dynamic memory: a theory of reminding and learning in computers and people. Cambridge University Press, UK (1982)
10. Steinberg, A.N., et al.: Revisions to the JDL data fusion model. In: Proceedings of the SPIE. Sensor Fusion: Architectures, Algorithms and Applications, pp. 430–441 (1999)

Algorithms to Automate Estimation of Time Codes for Captioning Digital Media

Daniel P. Harvey II and Peter Ping Liu

Eastern Illinois University, Charleston, IL 61920, USA

Abstract. Procedures were developed to partially automate the captioning process by estimating caption time codes using plain-text transcripts and audio recordings. Signal analysis is performed on the audio to measure pause location and duration, zero crossing rate (ZCR), and obtain frequency domain data. Algorithms were developed to match pauses in the audio to the ends of sentences in the transcript based on the observation that pause durations are greater at ends of sentences than within sentences. We have observed that ZCR peaks correspond to consonants in speech and that continuous wavelet transforms (CWT) work well for distinguishing between groups of consonants. These measurements will be used to develop algorithms to match selected phonemes in the audio to text in the transcript to supplement the pause matching results.

1 Introduction

The presentation of information on the World Wide Web relies increasingly on multimedia technologies. The audio component of these media presentations on the web remains largely inaccessible to persons who are Deaf or Hard of Hearing [1]. Captioning provides accessibility to media resources for deaf and hearing-impaired persons. However, captioning with currently available captioning software is time and labor intensive because it requires manually generating a text transcript and manually determining time codes. The goal of this study is to develop ways to provide synchronized captions more effectively and efficiently by extracting and analyzing signal data from the audio and using this information to automatically align text captions with audio recordings. In order to simplify implementation as much as possible, we are attempting to accomplish alignment without speech recognition in order to circumvent the need to compile a vocabulary for the system and so that users will not need to train the system for individual speakers.

2 Methods

2.1 Recordings and Transcripts

The recordings used in this study were of professional speakers/readers from American English radio and television broadcasts and of non-professional speakers reading text from a novel. The data set used in this study consisted of seven

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 585–590, 2009.

audio files that represented actual cases of media requiring captioning. The text consisted of 7140 words in 341 sentences. The transcripts were manually generated because commercially available speech recognition is currently not accurate enough to automatically generate the transcripts from recorded speech.

2.2 Signal Analysis

WAV audio files were read as binary data using a program written in C. Amplitude was measured. It was found that the RMS amplitude of each audio file performs adequately as a threshold between silence and speech for captioning. The locations and durations of pauses in the audio recordings were compiled by scanning for segments of the recording below the amplitude threshold. Statistically, pause durations at the ends of sentences are significantly greater than those within sentences. This observation was used to match the ends of sentences in the text to pauses in the audio track using the algorithms developed in this study.

Zero Crossing Rate (ZCR) was measured using 20 ms windows. The ZCR was plotted as a function of time. The ZCR peaks were manually matched to what phoneme was occurring at that time.

The times of the ZCR peaks were tabulated. In windows of 20 ms centered on the ZCR peaks, wavelet analysis was performed using the FAWAVE software package. Scalograms were computed using a complex Gabor wavelet [2] with width 0.125 and frequency 16. The magnitudes were graphed using 6 octaves and 16 voices. For comparison purposes, Fourier spectrograms were computed for the 20 ms samples.

2.3 Global Averaging Method

In the Global Averaging Method, the number of characters in each sentence and in the entire transcript are counted. The algorithm differentiates between punctuation points at the ends of sentences (periods, question marks, and exclamation points) and punctuation points within sentences (commas, semi-colons, colons, and dashes). All characters, except for punctuation marks, are assumed to represent equal amounts of time. This algorithm has an additional assumption that the pauses with the longest durations correspond to punctuation marks in the transcript. The total duration of each caption ($t_{caption}$) is the sum of the duration of the pauses (t_{pauses}) and the duration of the speech segments, or articulation time, (t_{speech}) in that caption as expressed in Equation 1.

$$t_{caption} = t_{pauses} + t_{speech} \tag{1}$$

The duration of the articulation time for each sentence was estimated from the number of characters in each sentence as a proportion of the total recording time. Character Count Weighting (CCW) assumes that each non-punctuation character, including blank spaces, in the text transcript corresponds to an equal amount of time. This assumption is based on the observation that one of the factors in syllable timing is the number of phonemes it includes [3]. For every caption, the number of spaces, the number of punctuation points within sentences, the number of punctuation points at the ends of sentences, and the number of alphanumeric characters are tabulated. The articulation time of each caption (t_{speech}) is estimated by calculating the ratio of the number of non-punctuation characters in each caption ($cc_{caption}$) to the number of non-punctuation characters in the entire transcript (cc_{total}) and multiplying that ratio by the clip time (t_{total}) as shown in Equation 2.

$$t_{speech,caption} = t_{speech,total}(cc_{caption}/cc_{total}) \tag{2}$$

To improve upon the accuracy of the algorithm, factors are added to account for pauses corresponding to punctuation. This is done by measuring the longest pauses detected in the audio track, calculating an average pause duration associated with punctuation marks within sentences and at the ends of sentences, and adding a pause time based on the number of punctuation marks in each caption.

The estimated timeline of the captions is constructed by:

1) estimate articulation time for a caption using CCW
2) add averaged pause time based on number of punctuation marks
3) repeat process to the end of the transcript.

2.4 Local Maxima Method

In the Local Maxima Method, the text is parsed into captions using end-of-sentence punctuation as break points. The number of end-of-sentence punctuation points (n) equals the number of captions. A total articulation time is estimated by subtracting the total duration of the n longest pauses from the total cliptime. The duration of each caption articulation time is estimated with Character Count Weighting (CCW). The estimated timeline of the captions is then constructed by:

1) moving out the estimated caption duration for sentence 1($t_{speech,caption1}$)
2) querying pauses in an area centered on $t_{speech,caption1}$
3) using a range that is defined as a percentage of the estimated duration of the preceding caption with a set minimum range
4) finding the longest pause starting within that range
5) if there are multiple candidate pauses in the search range, obtaining manual feedback from user to locate end of sentence (optional)
6) adding the chosen pause duration to the estimated caption duration
7) moving out the estimated caption duration for sentence 2 ($t_{speech,caption2}$)
8) and then repeating the process to the end of the transcript

The Local Maxima Method constructs a timeline without using average values for pauses at ends of sentences, but rather uses actual pause duration values.

When matching estimated times from text analysis to the location of pauses measured from the audio file, the local maxima method limits the pause pool by limiting the time range searched; whereas the global threshold method limits the pause pool by a pause duration criterion.

3 Results

3.1 Pause Matching

Table 1 shows the maximum and the average errors for the three types of speech tested in this study. The maximum error and the average error are much less for the Local Maxima Method than for the Global Averaging Method. In order to successfully distinguish between within-sentence pauses and end-of-sentence pauses on the basis of duration, it is advantageous to utilize data over localized portions of a file rather than over the entire audio file. The incorporation of manual feedback with the alignment algorithm resulted in a greater reduction in error. The overall rate of manual feedback requests for the files tested was approximately one request for every ten captions. For the files tested in this study, the Local Maxima Method with Manual Feedback accurately estimated the timing of 96% of the captions within 0.5 seconds. Comparable accuracy has been achieved in ongoing testing with actual media captioning projects using a web application that incorporates the Local Maxima Method with Manual Feedback.

3.2 Zero Crossing Rate and Frequency Analysis

ZCR peaks were distinctive and easily detected. We found a total of 14 consonants that corresponded to ZCR peaks. These consonants are listed in Table 2. The number of peaks within a given amount of time (90/min) is too large to be helpful in matching them to the transcript. We then analyzed the audio data near the ZCR peaks using spectrograms from Fourier analysis and continuous wavelet transforms (CWT).

The magnitude peaks of sections of CWT scalograms taken at ZCR peaks are distinctive and occur at different octaves for different phonemes. The results of these analyses are summarized in Table 2. Based on these preliminary observations, the octave at which the maximum peak in the scalogram slice can be used to distinguish four phonemes from the set of 14 consonants that correspond to ZCR peaks. The rate of occurrence for those four phonemes in the recordings tested was seven per minute. This should facilitate the process of matching these phonemes to their occurrence in the transcript. The approach of combining ZCR and wavelet analysis has been used to classify segments of speech signals into broad phonetic categories including silence, voiced, unvoiced, and plosive release [4]. The results of the CWT were more straightforward and easier to interpret than the spectrogram results. This is consistent with a previous study using wavelet analysis to distinguish between classes of phonemes [5].

Table 1. Error in Pause Matching Algorithms

Audio Type	Cases	Algorithm	Max. Error (s)	Average Error (s)
novel reading	2	GAM	24.83	3.66
		LMM	15.14	3.26
		LMM with FB	5.58	0.10
radio broadcasts	2	GAM	10.56	3.24
		LMM	10.11	1.36
		LMM with FB	5.06	0.42
scripted narrations	3	GAM	6.84	0.53
		LMM	0.00	0.00
		LMM with FB	0.00	0.00

Table 2. Octave and Magnitude of maximum peak in CWT scalogram of phonemes corresponding to ZCR peaks

Phoneme	Octave	Magnitude
ch	4.83	2.69E+10
j	4.94	1.17E+10
sh	4.84	2.04E+10
zh	4.63	4.15E+09
d	0.69	4.58E+10
f	3.19	3.64E+08
g	0.63	2.18E+11
h	3.81	6.78E+08
k	3.00	1.21E+10
p	2.81	2.70E+08
s	2.41	1.07E+10
t	2.69	6.89E+09
th	2.00	3.08E+08
z	1.88	1.74E+09

4 Conclusions and Future Work

The level of accuracy achieved in this study indicates that the method proposed for aligning text to audio using pauses is sufficient for the task of estimating the timing of captions for media. However, future work will include testing the algorithms on a larger data set taken from a standard corpus such as the TIMIT speech database [4, 5] in order to verify these results and to facilitate comparison to standard methods of text alignment. It is also planned to test the pause matching algorithms with other languages such as Hindi and Spanish.

The use of ZCR and wavelet analysis appears to be a promising approach for distinguishing a relatively small subset of phonemes. Because the feature matching process relies primarily on time domain parameters and because the wavelet analysis is distinguishing consonants that have a component of turbulent airflow, we are avoiding the use of signal features that are strongly influenced by the size and shape of the vocal tract and thereby again avoiding the need for training the system to individual speakers. However, testing on a larger data set will be necessary to verify these preliminary results. Based on the results of the pause matching algorithms, the approach that will be taken to match these four phonemes is to use character count weighting to approximate the temporal location of the candidate phonemes found in the transcript and then match them to the appropriate audio events based on the location of the ZCR peaks. These results will be used to supplement the pause matching results in order to improve the accuracy of the alignment process for a wider variety of speech recordings and to reduce the need for manual feedback in the alignment process.

References

1. Canadian Network for Inclusive Cultural Exchange: Online enhanced captioning guidelines (2004), http://cnice.utoronto.ca/guidelines/caption.pdf (retrieved September 22, 2004)
2. Walker, J.S.: Primer on wavelets and their scientific applications. CRC Publishing, Boca Raton (1999)
3. Campbell, W.N.: Syllable-based segmental duration. In: Bailly, G., Benoit, C., Sawallis, T.R. (eds.) Talking machines: Theories, models, and designs, pp. 211–224. Elsevier Science, Amsterdam (1992)
4. Pernkopf, F., Van Pham, T., Bilmes, J.A.: Broad phonetic classification using discriminative Bayesian networks. Speech Communication 51, 151–166 (2009)
5. Tan, B.T., Fu, M., Spray, A., Dermody, P.: The Use of Wavelet Transforms in Phoneme Recognition. In: Proceedings of ICSLP 1996 (1996)

Disclosing Patterns in IT Project Management - A Rough Set Perspective

Georg Peters[1] and M. Gordon Hunter[2]

[1] Munich University of Applied Sciences
Department of Computer Science and Mathematics
80335 Munich, Germany
georg.peters@hm.edu
[2] University of Lethbridge
Faculty of Management
Lethbridge, Alberta T1K 3M4, Canada
ghunter@uleth.ca

Abstract. Information technology has become one of the most important infrastructure components of virtually any organization. Although information technology has a crucial impact on the success of organizations it is reported that IT projects have rather high failure rates. Therefore, it is vitally important for organizations to improve the performance and success rates of IT projects. However, the reasons for failures are versatile and an ongoing very active fields of research especially in information systems and management. An established approach to evaluate IT projects is to define relevant so called *critical success factors* and analyze IT projects according to these criteria. This analysis is often of a qualitative nature. The objective of our paper is to enrich the analysis of critical success factors by alternative methods in particular rough set theory. We motivate the usage of rough sets to further improve the analysis of critical success factors with the goal to better manage IT projects and increase their success rate.

Keywords: Uncertainty, Soft Computing, Rough Sets, IT Project Management, Critical Success Factors.

1 Introduction

Although the management of IT projects has a long history of many decades it still is a challenging and risky venture to implement software in organizations - no matter whether it is individual software or ERP systems. There have been countless reports on spectacular IT project failures and figures about failure rates go up as high as up to 60 percent. Therefore, very active research in information systems concentrates on the evaluation of IT projects and concepts of how to successfully manage them.

Due to the nature of these projects research in this area is often of a qualitative nature. Since, rough set theory provides very useful concepts to deal with qualitative data it can enrich the analysis of IT projects.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 591–596, 2009.

Therefore, the objective of the paper is to evaluate the potentials of rough set theory for the analysis of critical success factors in IT projects, particularly the implementation of ERP software.

The remaining paper is organized as follows. In Section 2 we discuss main challenges and give an overview of critical success factor for IT projects. In the following section we present a rough set approach to analyze the critical success factors. The paper concludes with a summary in Section 4.

2 Challenges in IT Project Management

2.1 Challenges

Challenges and failures of IT projects have been widely discussed in the literature (e.g. [1,2]). Even after roughly two decades of experience in IT project management these projects are of high risk. Besides failing IT projects often exceed time and budget constraints.

2.2 Critical Success Factors

To address the challenges of IT projects critical success factors have been suggested and evaluated to give guidelines how to set up and run a successful IT project [3,4,5,6,7,8]. Due to the complexity of IT projects much of the research on critical success factors is of a qualitative nature rather than based on quantitative research methods.

The critical success factors as investigated by Somers and Nelson [9] have gained reasonable attention in literature. Somers and Nelson evaluated the importance of the given critical success factors by interviewing IT project managers. The critical success factors ordered in decreasing relevance are shown in Tab. 1.

We use these critical success factors in our research to disclose IT project patterns in the following Section.

Table 1. Critical Success Factors [9]

Top management support
Project team competence
Interdepartmental cooperation
Clear goals and objectives
Project management
Interdepartmental communication
Management of expectations
Project champion
Vendor support
Careful package selection
Data analysis and conversion
Dedicated resources
Use of steering committee
User training on software
Education on new business processes
Business Process Reengineering
Minimal customization
Architecture choices
Change management
Partnership with vendor
Use of vendors' tools
Use of consultants

3 Rough Set Based Analysis of Critical Success Factors

3.1 Preliminaries

Rough sets were introduced by Pawlak [10,11,12] in the early eighties of the last century. The core idea of rough sets is to describe a set by two approximations, its lower and upper approximation. While the objects in the lower approximation surely belong to the set, the membership of the objects in the upper approximation is unclear - they may or may not belong to the set. Obviously, to obtain this ambiguous status, these objects have to be members of at least two different upper approximations simultaneously.

Applying this concept to the analysis of critical success factors in IT project management we define the two sets *green* and *red* which describe the project status[1]. Projects assigned to the lower approximation of the green set are considered to run well and therefore 'just' need the normal attention of the management. IT projects assigned to the lower approximation of the red set are definitely at risk so they need management action to be taken immediately.

The status of the projects in the upper approximations of the set red as well of the green set is unclear. They may run well or they may be at risk. To obtain a traffic light style project monitoring system (see Fig. 1) they are labeled with the color yellow to indicate that they require a more detailed analysis to identify their current project status. Therefore, summarized, we obtain the following traffic light style project monitoring system:

- *Red*: The project is at risk.
- *Yellow*: The project needs further investigations before a decision on its status can be made.
- *Green*: The project runs well.

The underlying rough set based analysis consists of the following two major steps (see Fig. 1):

1. *Classifier Design*[2] *- Creating the Rough Decision Table.* The classifier design mainly consists of defining the rough decision table which will be used to evaluate the status of the project.
2. *Classification - Deciding on the Project Status.* The decision on the project status and definition of any consequent action is often within the responsibility of so called steering committees. Based on the rough decision table the steering committee can be provided with a traffic light system that classifies the IT projects.

These two steps, classifier design and classification, will be discussed in more detail in the following sections.

[1] Further research may lead to a refined structure of sets.

[2] Note, we use the terms classify and classifier in the sense that new data are assigned to the cluster obtained by a cluster analysis. Both terms are also used in supervised learning and have a different meaning there.

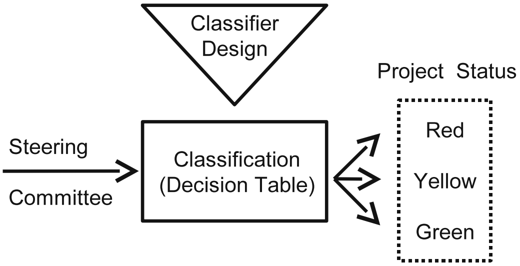


Fig. 1. Mile Stone Decision of a Steering Committee

3.2 Classifier Design - Creating the Rough Decision Table

For the design of the classifier, the rough decision table, we propose to interview experienced project managers or evaluate completed IT projects according to predefined critical success factors. For each IT project a record in the decision table will be created.

Due to the space limitations of the paper we use a simple sample decision table consisting of six (fictitious) interviews and a reduced set of four critical success factors as depicted in Tab. 2.

Table 2. A Sample Decision Table for IT Projects

#	Top Management Support	Project Team Competence	User Training on Software	Use of Consultants	*Decision:* Project Status
1	moderate	high	extensive	no	red
2	low	weak	moderate	yes	red
3	strong	high	moderate	yes	green
4	moderate	weak	low	no	red
5	moderate	high	extensive	no	green
6	moderate	high	extensive	yes	green

The records 2 and 4 are unambiguous members of the set red while the records 3 and 6 clearly belong to the set green. However, the records 1 and 5 have identical arguments but lead to different decisions. So, their memberships are equivocal. In rough set terms they belong to the upper approximations of both, the green and the red, sets [13].

Note that the design of the decision table can easily be enriched by missing attribute values as suggested by Grzymala-Busse [14].

3.3 Classification - Deciding on the Project Status

Based on the rough decision table as developed in the previous section the classification of IT projects is straight forward. In an on-going project the relevant

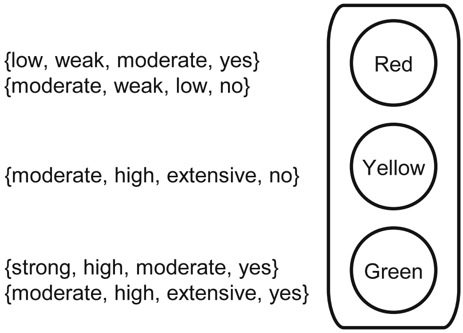

{low, weak, moderate, yes}
{moderate, weak, low, no}
{moderate, high, extensive, no}
{strong, high, moderate, yes}
{moderate, high, extensive, yes}
Red
Yellow
Green

Fig. 2. Traffic Light Style IT Project Monitoring System

attribute values have to be determined and compared to the attribute values of the rough decision table. The classification rules according to the traffic light system are depicted in Fig. 2.

After each classification the rough decision table has to be verified again. For example, an IT project is classified as ‘green’ but eventually fails. So, a new record has to be inserted into the rough decision table with the project status ‘red’. Consequently the attribute values lead to contradicting project status now (‘green’ and ‘red’). Therefore the traffic light has to be updated and must indicate ‘yellow’ for this set of attribute values.

In the case of a new set of attribute values the IT project has to be classified whether it is regarded as successful or not. Based on this evaluation the rough decision table has to be extended. Finally, the rough decision table has to be analyzed again to identify any changes in the upper and lower approximations.

4 Conclusion

In this paper we presented a potential application of rough sets for the analysis of critical success factors in IT projects, particulary ERP projects. We showed that - based on the evaluation of their critical success factors - IT projects can be group into classes.

We suggested a traffic light system indicating whether an IT project runs well, is at risk, or has an unclear status. Such a traffic light system is well accepted in management since it provides easily assessable status information within a company. While a green status project does not require more management attention, projects with a yellow status need further investigation, and red projects require immediate action.

Our future research will include an extensive case study to further evaluate and validate the proposed method and to develop a more advanced rough decision table of the critical success factors in IT projects.

References

1. Avital, M., Vandenbosch, B.: SAP implementation at Metalica: an organizational drama in two acts. Journal of Information Technology 15, 183–194 (2000)
2. Chen, C., Law, C., Yang, S.: Managing ERP implementation failure: A project management perspective. IEEE Transactions on Engineering Management 56(1), 157–170 (2009)
3. Akkermans, H., van Helden, K.: Vicious and virtuous cycles in ERP implementation: A case study of interrelations between critical success factors. European Journal of Information Systems 11(1), 35–46 (2002)
4. Al-Mashari, M., Al-Mudimigh, A., Zairi, M.: Enterprise resource planning: A taxonomy of critical factors. European Journal of Operational Research 146(2), 352–364 (2003)
5. Hsu, L.L., Lai, R.S.Q., Weng, Y.T.: Understanding the critical factors effect user satisfaction and impact of erp through innovation of diffusion theory. International Journal of Technology Management 43(1-3), 30–47 (2008)
6. Nah, F.F., Lau, J.L., Kuang, J.: Critical factors for successful implementation of enterprise systems. Business Process Management Journal 7(3), 285–296 (2001)
7. Plant, R., Willcocks, L.: Critical success factors in international ERP implementations: A case research approach. Journal of Computer Information Systems 47(3), 60–70 (2007)
8. Kappelman, L.A., McKeeman, R., Zhang, L.: Early warning signs of it project failure: The dominant dozen. Information Systems Management 23(4), 31–36 (2006)
9. Somers, T., Nelson, K.: The impact of critical success factors across the stages of enterprise resource planning implementations. In: Proceedings of the 34th Hawaii International Conference on System Sciences, Maui, Hawaii (2001); (on CD)
10. Pawlak, Z.: Rough sets. Report 431, Institute for Computer Science. Polish Academy of Sciences (1981)
11. Pawlak, Z.: Rough sets. International Journal of Computer and Information Science 11, 341–356 (1982)
12. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
13. Grzymala-Busse, J.: Introduction to rough set - theory and applications. In: Tutorial at KES 2004 - 8th International Conference on Knowledge Based Intelligent Information & Engineering Systems, Wellington, New Zealand, pp. 2004–2008 (2004)
14. Grzymala-Busse, J.: Three approaches to missing attribute values: A rough set perspective. In: Lin, T., Xie, Y., Wasilewska, A., Liau, C. (eds.) Data Mining: Foundations and Practice. Studies in Computational Intelligence (SCI), vol. 118. Springer, Heidelberg (2008)

Comparison of GARCH, Neural Network and Support Vector Machine in Financial Time Series Prediction

Altaf Hossain[1], Faisal Zaman[2], M. Nasser[1], and M. Mufakhkharul Islam[3]

[1] Department of Statistics, Rajshahi University, Rajshahi-6205, Bangladesh
rasel_stat71@yahoo.com, mnasser.ru@gmail.com
[2] Department of System Design and Informatics, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, 820-8502 Fukuka, Japan
faisal.zaman80@gmail.com
[3] Department of Computer Science & Engineering, BUET, Dhaka-1000, Bangladesh
nasif192@yahoo.com

Abstract. This article applied GARCH model instead AR or ARMA model to compare with the standard BP and SVM in forecasting of the four international including two Asian stock markets indices.These models were evaluated on five performance metrics or criteria. Our experimental results showed the superiority of SVM and GARCH models, compared to the standard BP in forecasting of the four international stock markets indices.

Index Terms-Generalized Autoregressive Conditional Heteroskedastic (GARCH), Neural Network (NN), Back Propagation (BP), Artificial BPNN (BPANN), Support Vector Machine(SVM).

1 Introduction

There is a powerful method, available in the literature for univariate time series forecasting, known as Box-Jenkins [2] Autoregressive Moving Average (ARMA) approach on stationary time-series. In such conventional econometric models, the variance of the disturbance is assumed to be constant. But many economic and financial time series such as exchange rates, indices, market returns, inflation rate etc exhibit periods of unusual large volatility, followed by periods of relative tranquility. In such circumstance, Engle [5] showed that it is possible to model the mean and the variance of a series simultaneously by a model named Autoregressive Conditional Heteroskedastic (ARCH). In latter, Bollerslev [1] extended Engle's original work by developing a technique that allows the conditional variance to be an ARMA process and that extended process is known as the GARCH process.The classical methods are based on some specific assumptions, such as linearity; or on error distributions, such as normality. In this circumstance, ANNs were developed to meet the increasing demand that can predict, detect, classify and summarize the structure of variables and define the relationships between them - without relying too much on such assumptions.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 597–602, 2009.

Not all relationships in economics and finance are direct. But the hidden layer of an ANN can capture all non-direct relationships between input and output variables. On the other hand, the SVM, originally developed as an implementation of Vapnik's Structural Risk Minimization (SRM) principle (Vapnik [12]), is now being used to solve a variety of learning, classification and prediction problems. In many ways, an SVM performs the same function as an ANN. It has the following advantages over ANN: (1)it can obtain the global optimum and (2) the overfitting problem can be easily controlled.

Steiner and Wittkemper [8] investigated the performance of several ANN models that forecast the return of a single stock. Kamruzzaman and Sarker [7] modeled and forecasted currency of exchange rates using three ANNs and a comparison was made with traditional ARIMA model. Tay and Cao [9] examined the feasibility of SVM in financial time series forecasting by comparing it with a multilayer BPNN and they showed that SVM outperforms the BPNN. Kim and Han [14] showed that SVM provides a promising alternative to stock market prediction comparing it with BPNNs. In most of studies like the above, they showed that SVM outperformed both BPNN and traditional statistical models. Again the simple neural learning procedures, such as BP algorithms easily outperformed the best practice of traditional statistical models. Chen and Wu [3] compared SVMs and BPs taking Autoregressive (AR) as a benchmark in forecasting the six major Asian stock markets. They showed that both the SVMs and BPs outperform the traditional models, ARs. They did prediction of transformed data, but not level data.

In this paper, our main contribution is to examine the capacity of the GARCH model comparing with the standard BPNN and simple SVM in forecasting financial time series at level data. We are going to use the monthly data in our empirical study for its Aggregational Gaussianity property.

The methodology including two prediction methods in this paper is discussed in Section 2. The results and discussion are presented in Section 3. Finally, our conclusion is in Section 4.

2 Methodology

2.1 The GARCH Modeling

Engle [5] showed that the serial correlation in squared returns, or conditional heteroskedasticity, can be modeled using an ARCH model. An important extension of the ARCH model proposed by Bollerslev [1] replaces the AR (p) representation in the variance series of ARCH process with an ARMA (p,q) formulation of the form

$$y_t = E_{t-1}[y_t] + \epsilon_t \tag{1}$$

$$\epsilon_t = z_t \sigma_t \tag{2}$$

$$\sigma_t^2 = a_0 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2 + \sum_{i=1}^{q} a_i \epsilon_{t-i}^2 \tag{3}$$

where $E_{t-1}[.]$ represents expectation conditional on information available at time $t-1$ and z_t is a sequence of iid random variables with mean zero and unit variance. In the GARCH model z_t is assumed to be iid standard normal. And the coefficients $a_i(i=0,...,q)$ and $b_j(j=1,...,p)$ are all assumed to be positive to ensure that the conditional variance σ_t^2 is always positive. The model in (3) together with (1)-(2) is known as the generalized ARCH or GARCH (p,q) model.

Hansen and Lunde [6], provided compelling evidence that it is difficult to find a volatility model that outperforms the simple GARCH (1,1). So we use GARCH(1,1) in this paper. For testing ARCH/GARCH effects, assessing the fit, diagnostic checks for model adequacy and estimation see Enders [4].

2.2 Support-Vector Regression (SVR)

The SVM deals with the classification and regression problems by mapping the input data into the higher-dimensional feature spaces. Our problem is regression.Firstly, we discuss the simplest model of SVR, a linear regression in the feature space. The SVR algorithm tries to construct a linear function such that training points lie within a distance ε(i.e., ε-insensitive zone in the error loss function).

Given a set of training data $\{(X_1,Y_1),...,(X_l,Y_l)\} \subset X \times \mathbb{R}$, where X denotes the space of the input patterns, the goal of SVR is to find a function $f(x)$ that has at most ε deviation from the targets y_i for all the training data and, at the same time, is as flat as possible. Let the linear function f takes the form:

$$f(x) =< w, x > +b; w \in X, b \in \mathbb{R} \tag{4}$$

The optimal regression function is given by the minimum of the functional,

$$\Phi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i^- + \xi_i^+) \tag{5}$$

where C is pre-specified value, and ξ_i^-, ξ_i^+ are slack variables representing upper and lower constraints on the outputs of the system. Flatness in (5) means a smaller $\|w\|$. Using ϵ -insensitive loss function, we get our required solutions.

To enable the SVR to predict a non-linear situation, we map the input data into a feature space. By using a kernel function, it is possible to compute the SVR without explicitly mapping in the feature space. See Scholkopf and Smola [13], and Chen and Wu [3] for more details. We use the radial basis kernel function $k(x,y) = \exp(-\gamma|x-y|^2)$ from several typical kernel functions as it performs well under general smoothness assumptions.

3 Results and Discussion

3.1 Data Source, Nature and Preprocessing of the Models

For analysis we have used the monthly data of the four international stock market indices collected from the Financial Forecast Center Home Page. We have used

Table 1. Description of data sets

INDEX	*TIMEPERIOD*	*N*	*HIGH*	*LOW*	*MEAN*	*SD*
Nikkei 225 (NK)	1984.01- 2006.01	265	38920	7837	18020	6643.7
Hang Seng (HS)	1986.12- 2006.01	230	17410	2138	8977	4411.6
FTSE 100 (FT)	1984.04- 2006.01	262	6930	1009	3559	1612.3
DAX (DX)	1990.11- 2006.01	183	7645	1398	3637	1690.4

Here N indicates the number of realizations for each market.

here the monthly data for its Aggregational Gaussianity property. The names of the stock markets are Japan (Nikkei 225 from January, 1984 to January, 2006), Hong Kong (Hang Seng from December, 1986 to January, 2006), U.K. (FTSE 100 from April, 1984 to January, 2006), and Germany (DAX from November, 1990 to January, 2006). The time periods used and the indices' statistical data are listed in Table 1. The time periods cover many important economic events, which we believe are sufficient for the training models.

We transform the original closing price into white noise series differencing two times at lag one for each series except the market NK, where we take one time difference. According to Thomason [10], [11] and Tay and Cao [9], this transformation has many advantages. The modification is known, so we can reverse it to compare the forecasts with the original series. Each of the four data sets is partitioned into three subsets according to the time sequence in the ratio 95: 2.5: 2.5. The first part is used for training; the second is a validating set that selects optimal parameters for the GARCH and SVR, and prevents the overfitting found in the BP neural networks selecting its number of inputs and units for the hidden layer, and the last is used for testing.

3.2 Comparison and Discussion

For each market, we also design different ordered ARMA models as to show the necessity of GARCH model. The Autocorrelation Function (ACF) of squared residuals for each fitted ARMA indicates the presence of GARCH effects. The forecasting results of the ARMA, GARCH, BP and SVM for the test set are collated in Table 2. According to Tay and Cao [9] and Thomason [10] [11], the prediction performance is evaluated using the following statistics: Mean Square Error (MSE), Normalized Mean Squared Error (NMSE), Mean Absolute Error (MAE), Directional Symmetry (DS) and Weighted Directional Symmetry (WDS). See Chen and Wu [3] for more details. In general, GARCH, BP and SVM models outperform the ARMA model in the deviation performance criteria except only one market, HS. However, in the either direction and weighted direction performance criteria, the ARMA model sometimes shows better results comparing with BP and SVM models, possibly because SVMs and BPs are trained in terms deviation performance; the former to find the error bound and the latter to minimize MSE.The results of ARMA model are not added in the Table due to space constrain.

Table 2. Comparison of the results of GARCH, NN AND SVM models on the test set

		$SQRTMSE$	MAE	$NMSE$	DS	WDS
NK	GARCH	824.82	686.75	0.21	85.7	0.00
	BP	900.91	765.64	0.25	71.4	0.42
	SVM	792.57	659.18	0.91	85.7	0.00
HS	GARCH	656.85	525.59	1.88	0.00	0.00
	BP	778.94	864.11	2.65	33.3	1.18
	SVM	632.15	505.45	1.74	16.6	5.06
FT	GARCH	196.91	170.64	1.04	42.8	1.23
	BP	236.81	220.12	1.51	28.5	5.87
	SVM	158.47	147.05	0.67	57.1	0.45
DX	GARCH	234.99	229.32	0.55	40.0	0.66
	BP	191.05	156.14	0.36	60.0	0.79
	SVM	218.31	204.54	0.47	40.0	0.50

The deviation criteria, such as NMSEs are less than 1, except for the HS index in the SVMs. The NMSEs, both of GARCH and BP are greater than 1 for the two markets, HS and FT. It is remarkable that the NMSE of the BP is not only greater than 1 but also 2 for the market, HS. In the NK and HS data sets, the SVM models perform better than the GARCH and BP models. On the other hand, the GARCH method performs better than SVM and BP in only one market, FT; the BP models also perform better than the GARCH and SVM in only one market, DX. If we make comparison only between GARCH and NN, then GARCH shows better results than NN in three markets out of four.

4 Conclusion

In this article, we examine the feasibility of applying two Artificial Intelligence (AI) models, SVM and standard BP, and one classical statistical model, GARCH to financial time-series forecasting. Our experiments demonstrate that both models, SVM and BP perform better than the ARMA model in the deviation measurement criteria. Previous research also claims that the NN method is superior to the classical statistical method, ARMA or AR. If we consider the classical statistical method GARCH (just extended version of ARMA process), it is superior to the NN method except in the only one market DX of indices. It is not only superior to the NN method but also it shows close performance to the SVM method. The classical methods, ARMA and GARCH, require large number of sample size for better forecasting and these drastically reduce the original sample size when the high order model is fitted. The SVMs and GARCH models are more parsimonious than the NN models since the SVMs and GARCH models take smaller number of parameters for better forecasting. The SVMs and GARCH models are also more interpretable than that of NN because the ANN structures contain hidden layers which are not interpretable. In general, the GARCH, BP and SVM models perform well in the prediction of indices behavior in terms of the deviation criteria. Our experiments also show that both the statistical (GARCH) and the AI techniques can assist the stock market trading and the development of the financial decision support systems. But if we consider predictability along with interpretability, parsimonious and intrinsic properties, then we recommend the GARCH models with SVMs in financial time series prediction.

References

1. Bollerslev, T.: Generalized autoregressive conditional heteroscedasticity. Journal of Econometric 31, 307–327 (1986)
2. Box, G.E.P., Jenkins, G.M.: Time Series Analysis Forecasting and Control. Holden-Day, San Francisco (1976)
3. Chen, W.-H., Shih, J.-Y., Wu, S.: Comparison of support-vector machines and back propagation neural networks in forecasting the six major Asian stock markets. Int. J. Electronic Finance 1(1), 49–67 (2006)
4. Enders, W.: Applied Econometric Time Series. John Wiley & Sons, Inc., Chichester (2004)
5. Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50, 987–1007 (1982)
6. Hansen, P., Lunde, A.: A Forecast comparison of volatility models: does anything beat a GARCH (1,1) model? Journal of Applied Econometrics 20, 873–889 (2004)
7. Kamruzzaman, J., Sarker, R.: Forecasting of currency exchange rates using ANN: a case study. In: 2003 Proc. IEEE Intl. Conf. on Neur. Net. & Sign. Process. (ICNNSP 2003), China (2003)
8. Steiner, M., Wittkemper, H.G.: Neural Networks as an alternative stock market model. Neural Networks in the Capital Markets, 135–147 (1995)
9. Tay, F.E.H., Cao, L.: Application of support vector machines in financial time-series forecasting. Omega 29, 309–317 (2001)
10. Thomason, M.: The practitioner method and tools: a basic neural network-based trading system project revisited (parts 1 and 2). Journal of Computational Intelligence in Finance 7(3), 36–45 (1999a)
11. Thomason, M.: The practitioner method and tools: a basic neural network-based trading system project revisited (parts 3 and 4). Journal of Computational Intelligence in Finance 7(3), 35–48 (1999b)
12. Vapnik, V.N.: The Nature of Statistical Learning Theory, 2nd edn. Springer, New York (1995)
13. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, Massachusetts (2002)
14. Kim, K.-S., Han, I.: The cluster-indexing method for case-base reasoning using self-organizing maps and learning vector quantization for bond rating cases. Expert Systems with Applications 21, 147–156 (2001)

Mean-Entropy-Skewness Fuzzy Portfolio Selection by Credibility Theory Approach

Rupak Bhattacharyya[1], Mohuya B. Kar[2], Samarjit Kar[1],
and Dwijesh Dutta Majumder[3]

[1] Department of Mathematics, National Institute of Technology, Durgapur 713209, India
[2] Department of C.S.E., Heritage Institute of Technology, Kolkata 107, India
[3] Electronics & Communication Science Unit, Indian Statistical Institute, Kolkata 108, India
mathsrup@gmail.com, mohuya_kar@yahoo.com, kar_s_k@yahoo.com,
ddmdr@hotmail.com

Abstract. In this paper fuzzy mean-entropy-skewness models are proposed for optimal portfolio selection. Entropy is favored as a measure of risk as it is free from dependence on symmetric probability distribution. Credibility theory is applied to evaluate fuzzy mean, skewness and entropy. Hybrid intelligence algorithm is used for simulation. Numerical examples are given in favor of each of the models.

Keywords: Fuzzy portfolio selection problem, Credibility theory, Entropy, Skewness, Mean- entropy- skewness model, Hybrid intelligence algorithm.

1 Introduction

Different authors like Philppatos and Wilson [1], Philippatos and Gressis [2], Nawrocki and Harding [3], Simonelli [4], Huang [5], Qin et al. [6] used entropy as an alternative measure of risk to replace variance proposed by Markowitz [7]. Entropy is used as risk in the sense that uncertainty causes lose and so investors dislike uncertainty and entropy is a measure of uncertainty. Entropy is more general than variance as an efficient measure of risk because entropy is free from reliance on symmetric probability distributions and can be computed from non-metric data.

Skewness measures the asymmetry of a distribution. Portfolio returns are generally asymmetric and investors would prefer a portfolio return with larger degree of asymmetry when the mean value and risk are same. There will be three goals in portfolio optimization, maximizing the mean and the skewness and minimizing the entropy. Consideration of skewness in portfolio selection was started by Lai [8] and then continued by Konno et al. [9], Chunhachinda et al. [10], Liu et al. [11], Briec et al. [12] etc . All these literatures assume that the security returns are random variables. There are many non-stochastic factors that affect stock markets and they are improper to deal with probability approaches. So Ramaswamy [13], Inuiguichi et al. [14], Li et al. [15] etc. studied fuzzy portfolio selection. But till date none has considered mean-entropy- skewness model for portfolio selection problem.

In this paper, in section 2, the definition of fuzzy entropy by Li and Liu [16] is discussed and also credibility theory is used to evaluate the mean and skewness of the

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 603–608, 2009.

fuzzy return. In section 3 portfolio selection models are constructed. In section 4 fuzzy simulation integrated genetic algorithm is used to solve the proposed models. In section 5 numerical results are given to illustrate the method followed by conclusions in section 6.

2 Mean, Entropy, Skewness by Credibility Theory

In this section we will discuss credibility theory (c.f. Liu [17]) to have the mean, skewness and entropy of a fuzzy variable.

Definition 1: Suppose that ξ be a fuzzy variable with membership function μ_ξ . Then for any set $B \subset \Re$ the credibility of ξ ∈ B is defined as:

$$Cr\{\xi \in B\} = \frac{1}{2}\left(\sup_{x \in B} \mu_\xi(x) + 1 - \sup_{x \in B^c} \mu_\xi(x) \right).$$

Definition 2: The expected value of ξ is defined as

$$E[\xi] = \int_0^\infty Cr\{\xi \geq r\}dr - \int_{-\infty}^0 Cr\{\xi \leq r\}dr.$$

Definition 3: Suppose that ξ be a fuzzy variable with finite expected value. The skewness of ξ is defined as

$$S[\xi] = E[(\xi - E[\xi])^3].$$

Definition 4: Let ξ be a continuous fuzzy variable and let $T(t) = - t.ln(t) - (1-t).ln(1-t)$. Then the entropy of ξ is defined by

$$H[\xi] = \int_{-\infty}^{\infty} T(Cr\{\xi = r\})dr.$$

Definition 5: Let ξ be a discrete fuzzy variable taking values in $\{x_1, x_2, \ldots\}$ and let $T(t) = - t.ln(t) - (1-t).ln(1- t)$. Then the entropy of ξ is defined by

$$H[\xi] = \sum_{i=1}^{\infty} T(Cr\{\xi = x_i\}).$$

Example 1: The expected value, skewness and entropy of a triangular fuzzy number ξ = (a, b, c) are respectively

$$E[\xi] = \frac{a + 2b + c}{4}, S[\xi] = \frac{(c - a)^2(c - 2b + a)}{32} \text{ and } H[\xi] = \frac{c - a}{2}.$$

Theorem 1: Let $\tilde{r}_i = (a_i, b_i, c_i)$ (i =1,2,…,n) be independent triangular fuzzy numbers. Then

$$E[\tilde{r}_1 x_1 + \tilde{r}_2 x_2 + ... + \tilde{r}_n x_n] = \frac{1}{4}\sum_{i=1}^{n}(a_i + b_i + c_i),$$

$$H[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]=\frac{1}{2}\sum_{i=1}^{n}(c_i-a_i),$$

$$S[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]=\frac{1}{32}\left(\sum_{i=1}^{n}(c_i-a_i)x_i\right)^2\sum_{i=1}^{n}(c_i+a_i-2b_i).$$

Proof: Since $\tilde{r}_i=(a_i,b_i,c_i)$ are triangular fuzzy numbers, by Extension Principle of Zadeh it follows that

$$\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n=\left(\sum_{i=1}^{n}a_ix_i\ ,\ \sum_{i=1}^{n}b_ix_i\ ,\ \sum_{i=1}^{n}c_ix_i\right),$$

which is also a triangular fuzzy number. Combining this with the results obtained in example 1 we are with the theorem.

3 Mean-Entropy-Skewness Model

Let $\tilde{r}_i$ be a fuzzy number representing the return of the ith security. Let x_i be the portion of the total capital invested in security i, i = 1, 2, …, n.

Then $\tilde{r}_i=\dfrac{p_i^{'}+d_i-p_i}{p_i}$, where p_i is the closing price of the ith security at present, $p_i^{'}$ is the estimated closing price in the next year and d_i is the estimated dividends in the next year.

Now when minimum expected return (α) and maximum risk (γ) are known, the investor will prefer a portfolio with large skewness. It can be modeled as:

$$\begin{cases}\text{maximize } S[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]\\ \text{subject to}\\ E[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]\ \geq\ \alpha\\ H[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]\ \leq\ \gamma\\ x_1+x_2+....+x_n=1\\ x_i\ \geq\ 0,\ i=1,2,....,n.\end{cases} \qquad \text{.......................................} (1)$$

When expected return (α) and skewness (β) are both not less than some given target values, the investor would aim to minimize the risk; which can be modeled by II.

$$\begin{cases}\text{minimize } H[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]\\ \text{subject to}\\ E[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]\ \geq\ \alpha\\ S[\tilde{r}_1x_1+\tilde{r}_2x_2+....+\tilde{r}_nx_n]\ \geq\beta\\ x_1+x_2+....+x_n=1,\ x_i\ \geq\ 0, i=1,2,....,n\ .\end{cases} \qquad \text{.............................} (2)$$

When minimum skewness (β) and maximum risk (γ) is known, the investor would aim to maximize the expected return. It can be modeled as:

$$\begin{cases} \text{maximize } E[\tilde{r}_1 x_1 + \tilde{r}_2 x_2 + ... + \tilde{r}_n x_n] \\ \text{subject to} \\ H[\tilde{r}_1 x_1 + \tilde{r}_2 x_2 + ... + \tilde{r}_n x_n] \le \gamma \\ S\,[\tilde{r}_1 x_1 + \tilde{r}_2 x_2 + ... + \tilde{r}_n x_n] \ge \beta \\ x_1 + x_2 + ... + x_n = 1,\ x_i \ge 0,\ i = 1, 2, ..., n. \end{cases} \qquad \text{.................................} (3)$$

4 Hybrid Intelligence Algorithm

To find the optimal portfolio, we integrate fuzzy simulation into the genetic algorithm. The genetic algorithm procedure has been introduced in detail in [18]. Here, we sum up the hybrid intelligent algorithm as follows:

Step 1) In the GA, a solution $x = (x_1, x_2, \ldots, x_n)$ is represented by the chromosome $C = (c_1, c_2, \ldots, c_n)$, where the genes $c_1, c_2, \ldots, c_n$ are in the interval [0, 1]. The matching between the solution and the chromosome is through $x_i = c_i/(c_1 + c_2 + \cdots + c_n)$, $i = 1, 2, \ldots, n$, which ensures that $x_1 + x_2 + \cdots + x_n = 1$ always holds.
Randomly generate a point C from the hypercube $[0, 1]^n$. Use fuzzy simulation to calculate the entropy value, skewness and expected value. Then check the feasibility of the chromosomes. Take the feasible chromosomes as the initialized chromosomes.

Step 2) Calculate the objective values for all chromosomes by fuzzy simulation. Then, give the rank order of the chromosomes according to the objective values. For model-III, the greater the expected value is, the better the chromosome is, and the smaller the ordinal number the chromosome has. For model-II, the smaller the entropy value is, the better the chromosome is, and the smaller the ordinal number the chromosome has. Next, compute the values of the rank-based evaluation function of the chromosomes and the fitness of each chromosome according to the rank-based-evaluation function.

Step 3) Select the chromosomes by the roulette wheel selection method, which is fitness proportional.

Step 4) Update the chromosomes by crossover and mutation operations. Check the feasibility of the chromosomes in a similar way as the initial step.

Step 5) Repeat the second to fourth steps for a given number of cycles.

Step 6) Choose the best chromosome as the solution of portfolio selection. For model-III the chromosome with the maximum expected value is the best chromosome. For model-II the chromosome with the minimum entropy value is the best chromosome. For model-I the chromosome with the maximum skewness is the best chromosome.

5 Numerical Example

We apply our mean-entropy-skewness models to the data from Huang [19]. We take the first seven securities. The returns are triangular fuzzy numbers.

Security	Return($\tilde{r}_i$)	Security	Return($\tilde{r}_i$)
I	(-0.3, 1.8, 2.3)	V	(- 0.7, 2.4, 2.7)
II	(-0.4, 2.0, 2.2)	VI	(- 0.8, 2.5, 3.0)
III	(-0.5, 1.9, 2.7)	VII	(- 0.6, 1.8, 3.0)
IV	(-0.6, 2.2, 2.8)		

Example 1. Considering the minimum expected return and the bearable maximum risk to be as 1.6 and 1.8, we judge the model-I and the solution of the above model is obtained as:

Securities							Skewness
I	II	III	IV	V	VI	VII	
0	0	0	0	0	0.3077	0.6923	12.6875

Example 2. Considering the minimum skewness and the bearable maximum risk to be as 12.5 and 1.8, we judge the model-III and the solution of the above model are obtained as:

Securities							Mean
I	II	III	IV	V	VI	VII	
0	0	0	0	0	0.5549	0.4451	1.6803

Example 3. : Considering the maximum skewness and the minimum expected return to be as 12.5 and 1.8, we judge the model-II and the solution of the above model is obtained as:

Securities							Risk
I	II	III	IV	V	VI	VII	
0	0	0	0	0.0813	0.2451	0.6736	1.772

6 Conclusion

Since skewness is incorporated in the portfolio selection model, the model has become more sensible. Entropy is used as the measure of risk. The smaller the entropy value is, the more concentrative the portfolio return is and the safer the investor is. A hybrid intelligence algorithm is designed and numerical results are given to show the effectiveness of the method. Works are going on to apply the proposed approach in Indian stock market.

References

1. Philippatos, G.C., Wilson, C.J.: Entropy, market risk and selection of efficient portfolios. Applied Economics 4, 209–220 (1972)
2. Philippatos, G.C., Gressis, N.: Conditions of equivalence among E–V, SSD, and E–H portfolio selection criteria: The case for uniform, normal and lognormal distributions. Manag. Sci. 21, 617–625 (1975)
3. Nawrocki, D.N., Harding, W.H.: State-value weighted entropy as a measure of investment risk. Appl. Econ. 18, 411–419 (1986)
4. Simonelli, M.R.: Indeterminacy in portfolio selection. Eur. J. Oper. Res. 163, 170–176 (2005)
5. Huang, X.: Mean-entropy models for fuzzy portfolio selection. IEEE Transactions on Fuzzy Systems 16(4), 1096–1101 (2008)
6. Qin, Z., Li, X., Ji, X.: Portfolio selection based on fuzzy cross-entropy. Journal of Computational and Applied Mathematics 228(1), 188–196 (2009)
7. Markowitz, H.: Portfolio selection. J. Finance 7, 77–91 (1952)
8. Lai, T.: Portfolio selection with skewness: a multiple – objective approach. Review of the Quantitative Finance and Accounting 1, 293–305 (1991)
9. Konno, H., Suzuki, K.: A mean-variance-skewness optimization model. Journal of the Operations Research Society of Japan 38, 137–187 (1995)
10. Chunhachinda, P., Dandapani, P., Hamid, S., Prakash, A.J.: Portfolio selection and skewness: evidence from international stock markets. Journal of Banking and Finance 21, 143–167 (1997)
11. Liu, S.C., Wang, S.Y., Qiu, W.H.: A mean- variance- skewness model for portfolio selection with transaction costs. International Journal of System Science 34, 255–262 (2003)
12. Briec, W., Kerstens, K., Jokung, O.: Mean-variance- skewness portfolio performance gauging: a general shortage function and dual approach. Management Science 53, 135–149 (2007)
13. Ramaswamy, S.: Portfolio selection using fuzzy decision theory. Working paper of Bank for International Settlements 59 (1998)
14. Inuiguchi, M., Ramik, J.: Possibilistic linear programming: a brief review of fuzzy mathematical programming and a comparison with stochastic programming in portfolio selection problem. Fuzzy Sets and Systems 111, 3–28 (2000)
15. Li, X., Qin, Z., Kar, S.: Mean-variance-skewness model for portfolio selection with fuzzy returns. European Journal of Operational Research (2009), doi:10.1016/j.ejor.2009.05.003
16. Li, P., Liu, B.: Entropy of credibility distributions for fuzzy variables. IEEE Transactions for Fuzzy Systems 16(1), 123–129 (2008)
17. Liu, B.: Uncertainty Theory, 3rd edn., http://orsc.edu.cn/liu/ut.pdf
18. Huang, X.: Fuzzy chance-constrained portfolio selection. Applied Mathematics and Computation 177, 500–507 (2006)
19. Huang, X.: Mean-semivariance models for fuzzy portfolio selection. Journal of Computational and Applied Mathematics 217, 1–8 (2008)

Late Blight Forecast Using Mobile Phone Based Agro Advisory System

Arun Pande[1], Bhushan G. Jagyasi[1], and Ravidutta Choudhuri[2]

[1] TATA Consultancy Services, TCS Innovation Labs Mumbai, Thane, India
arun.pande@tcs.com, bhushan.jagyasi@tcs.com
[2] Department of Mathematics, Indian Institute of Technology Kharagpur,
Kharagpur, India
ravidutta@iitkgp.ac.in

Abstract. The late blight disease is the most common disease in potato, which is caused by the pathogen *Phytopthora infestans*. In this paper, a novel method to collect symptoms of the disease, as observed by the farmers, using a mobile phone application has been presented. A cumulative composite risk index (CCRI) obtained from more than one existing disease forecast models is validated from the actual late blight queries received from the farmers. The main contribution of the paper is a protocol that combines the symptoms based diagnostic approach along with the plant disease forecasting models resulting in detection of Potato late blight with higher accuracy. This in turn reduces the disease risk along with avoiding the unnecessary application of fungicide.

1 Introduction

This paper focuses on timely detection of the Late Blight in Potato crop under the Indian climatic conditions. About 80% of the potato farming in India is done in winter season, a time quite ideal for the pathogen *phytophthora infestans*, the agent responsible for Late Blight, to infest. Several models [1], [2], [3], [4], [5], [6], [7] have been developed for the prediction of a Late Blight attack based on weather parameters. Some of these models take into consideration atmospheric parameters such as temperature, humidity and rainfall, to decide the likelihood of a late blight attack. The *Ullrich* [4] and *Fry* [1] models utilizes hourly temperature and hourly humidity in order to compute a risk value. A cumulated risk value is computed in Ullrich's model, by summing the risk value for each day provided the daily risk is more than a particular threshold. The prediction for late blight is then made when this cumulated risk value crosses an assigned threshold. To calculate its 'blight unit', Fry's model classifies the temperatures broadly and then declares a unit based on the number of hours the humidity levels cross the threshold. *Winstel* model [5] issues a warning whenever the average temperature is within a certain range and hours of relative humidity greater than 90% cross a threshold of 10 hours. *Wallin* model [2], [3] provides a severity value based on hourly temperature and humidity measurements and suggests

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 609–614, 2009.

that the disease risk is high on 7-14 days after the cumulative severity value had exceeded the assigned threshold.

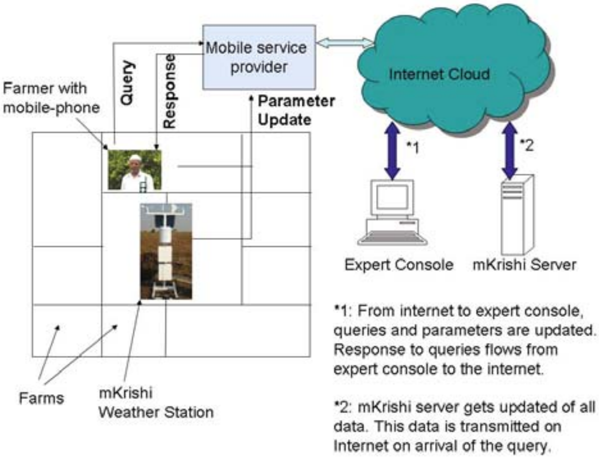
On the other hand, the diagnostic method to detect the late blight is by observing the symptoms [8], personally, on the crop.

In this paper we present a disease forecast protocol based on a mobile phone based agro advisory (mKRISHI) system. The key idea is to combine the disease risk obtained from various models and the knowledge of symptoms obtained from farmers input on mobile phone, to determine the appropriate doses of the fungicide. The proposed protocol minimizes the disease risk along with the farming losses.

2 The mKRISHI Based Late Blight Forecast and Diagnostic Protocol

Mobile phone based agro-advisory system (mKRISHI) [9] can provide expert's advice related to various domains such as farming, education, finance etc., to the rural masses. A simplified schematic of mKRISHI is shown in Fig. 1. A farmer can ask any query to the expert in the form of voice, text and images using the mKRISHI mobile client application on his mobile phone. Further, atmospheric, soil and plant related parameters, observed by the weather station deployed in the farm, are also made available to an expert for detailed investigation.

Here, we present a novel protocol to combine the existing forecast models and diagnostic techniques using the mKRISHI framework.



*1: From internet to expert console, queries and parameters are updated. Response to queries flows from expert console to the internet.

*2: mKrishi server gets updated of all data. This data is transmitted on Internet on arrival of the query.

Fig. 1. A Simplified Schematic of mKRISHI

2.1 The Proposed Protocol

Cumulative Composite Risk index. Some of the existing Late Blight disease forecast models were discussed in Section 1. Each disease forecast model m $(m = 1, 2, \cdots, M)$ has been assigned a weight 'W_m'. Let r_i^m be the risk value shown by the model m on the ith day, and r_{max}^m be the maximum possible risk value corresponding to the model m. The composite risk index R_i for ith day, considering all M models is given by

$$R_i = \sum_{k=1}^{k=M} \eta_i^k W_k \tag{1}$$

where, $\eta_i^m = r_i^m / r_{max}^m$. The composite risk index is then accumulated over last N days to obtain a Cumulative Composite risk index (CCRI). So, for any ith day, the Cumulative Composite Risk Index C_i is given by

$$C_i = \sum_{j=i-N+1}^{j=i} R_j \tag{2}$$

where, N depends on models selected to obtain composite risk index and the region of application.

Protocol. The CCRI (C_i) is compared with lower and higher thresholds T_L^C and T_H^C, respectively, $(T_H^C \geq T_L^C)$ to determine the appropriate doses of fungicide.

- **High Disease Risk** $(C_i > T_H^C)$
  If the high disease risk is observed and if the fungicide has not been applied in the past T_{ED} days, then an alert message is sent to the farmer, prompting for the application of fungicide. Here, T_{ED} is the duration for which the applied drug is effective on the farms. According to the standard treatment procedures of Late Blight, this period is usually atleast $5 - 6$ days [6], [1]. In our field trials, the agriculture experts were found to recommend the application of fungicide after 10 days from the date of application of the previous dose. Reuse of fungicide in T_{ED} days could be *harmful* in terms of higher residues and *unnecessary* in terms of cost.
- **Moderate Disease Risk** $(T_L^C < C_i < T_H^C)$
  If the moderate disease risk is observed and if the fungicide has not been applied in the past T_{ED} days, then after a period of δ days, the farmer is prompted on his mobile phone seeking the symptoms of Late Blight [8], as observed by him on the farms. Here, δ is the time interval between the risk shown by the models and the appearance of the symptoms. An estimate of δ can be obtained by using an adaptive scheme discussed in the following part of this section. The queries asked to the farmers to obtain the knowledge of symptoms of the disease may have binary answers. This enhances the usability and minimizes the communicated overhead. If the responses from the farmer confirms the attack of late blight, then the farmer is prompted for

the spray of appropriate fungicide. Otherwise, the queries are sent again on daily basis for the next few days. The number of days for which farmer should be queried depends on the value of δ chosen and the models considered for evaluating composite risk index.

- **Low disease Risk** ($C_i < T_L^C$) If the disease risk indicated by C_i is low, then no action needs to be taken.

Adaptive Evaluation of Parameters
Here, we provide a description of the adaptability of the scheme with respect to unknowns used in the proposed protocol. The detection of symptoms, is tentatively $3-7$ days after inoculation ([10]). For the first time, the queries to observe the symptoms can be sent 3 days ($\delta = 3$) after the moderate disease risk is indicated by C_i. Further, δ can be adapted based on the received responses from the farmer. Let γ be the percentage of queries with positive responses (symptoms confirming the disease) from a farmer. For each farmer, 'γ_{min}' and 'γ_{max}' are assigned, such that if there is a positive response to more than γ_{min} % of the queries and less than γ_{max} % of the queries, then the value of $\delta = 3$ is kept unchanged. If more than γ_{max} % of positive responses are received then next time the queries shall be asked a day earlier ($\delta = \delta - 1$). If less than γ_{min} % of the queries come in with a positive response then the queries are sent with a delay of one day ($\delta = \delta + 1$). However the time interval δ is not altered if doing so causes its value to exceed the stipulated limits. The limits are applied to make sure that the value of δ does not become too large to prevent the early detection or too small to increase the overload.

A weight W_m assigned to each model m, required to evaluate composite risk index, can be adapted over seasons of farming considering the difference between the composite risk values with the actual disease occurrences on the field. The values of N, T_L^C and T_H^C can also be fixed based on the results obtained by this scheme for the previous seasons and comparing the prediction dates with time of attacks.

3 Evaluation of Models

In this Section, we present the results obtained from disease forecast models, Ullrich and Schrodter [4], Wallin [3], [2], Fry [1] and Winstel [5], when applied to the data obtained from the mKRISHI weather station for the last farming season at Aligarh district, India. The hourly data of temperature and relative humidity from 16th to 27th of November and 1st to 8th of January, has been used to evaluate the Late Blight disease risk obtained from these models. Fig. 2 presents the binary disease risks (after quantization) by all models.

In the month of November, all the four models show strong correlation in disease risk indication, particularly on the 20th and 27th of November. Further, on 16th and 21st of November, Fry and Winstel indicated a disease risk. It was observed that Fry and Winstel gave indications of viable days of late blight inoculation for more number of times than Ullrich and Wallin models in November. In the first week of January, Ullrich-Schrodter, Wallin and Fry models depicted

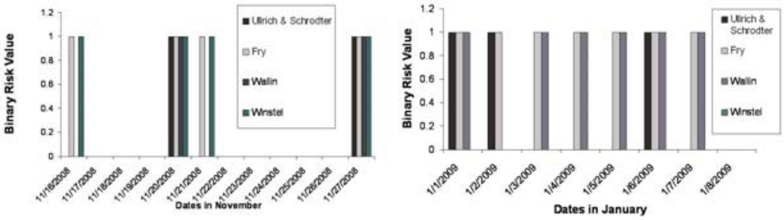


Fig. 2. Binary Late blight disease risks by all four models and Late blight disease queries asked by farmers in the month of November and January

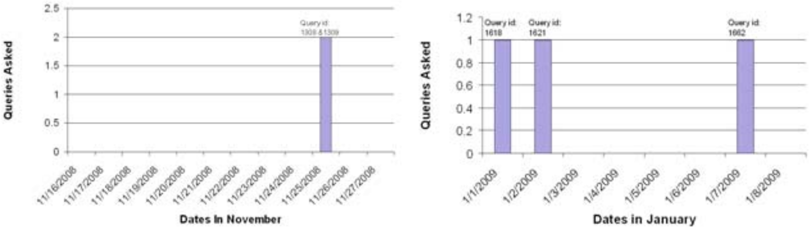


Fig. 3. Late blight disease queries asked by farmers in the month of November and January

the disease risks consistently. In Winstel's model, only the first phase has been implemented to obtain the likelihood of the initial attack. Thus no attack has been observed for Winstel model in the month of January. Fig. 3 presents the number of queries asked by the farmers using the mKRISHI application in the months of November and December. It shows that there were queries concerning late blight attack during the 3^{rd} week of November 2008 and 1^{st} week of January 2009. This validates risk indications obtained by the models in those weeks. Fig. 4 shows images of the crop sent by the farmers along with the queries which further confirm a Late Blight attack.

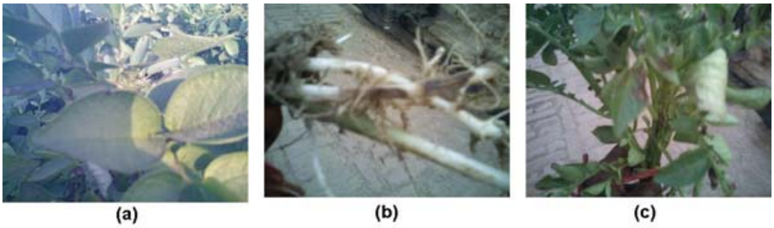


Fig. 4. Images received from the farmers indicating Late Blight attacks. (a) Query id: 1621, Dated: 2nd Jan 2008 (b) and (c) Query id: 1309, Dated: 25th Nov 2008.

4 Conclusion

The mKRISHI based Potato Late Blight forecast protocol has been proposed by making a conjunctive use of existing disease forecasting models and symptoms based disease diagnostic procedures. Towards this end, we have implemented four late blight disease models in Indian scenarios by using the data obtained from mKRISHI weather station. The results obtained were then validated using the actual *Late Blight queries* (in the form of text, voice and images) that were sent by the farmers using mKRISHI application. We proposed a protocol, which confirms the disease symptoms from the farmers by asking them simple questions having 'yes/no' (binary) responses. This minimizes false alarm and hence reduces the amount of fungicide application if the computed disease risk is moderately high. Further, the losses are reduced by alerting the farmers immediately if the risk indicated by the models crosses certain higher (critical) threshold.

Acknowledgments. We acknowledge B. B. Singh from Tata Chemicals Limited, Sunil, Sanjay, Pankaj, Ajay, Ankita and Dinesh, along with their team members from TCS Innovation Labs-Mumbai, for their support in conceptualization and development of the mKRISHI application.

References

1. Fry, W.E., Apple, A.E., Bruhn, J.A.: Evaluation of potato late blight forecasts modified to incorporate host resistance and fugicide weathering. Phytopathology 73, 1054–1059 (1983)
2. Wallin, J.R.: Summary of recent progress in predicting the late blight epidemics in unitedstates and canada. American Potato Journal 19, 306–312 (1962)
3. Wallin, J.R.: Forecasting tomato and potato late blight in the northcentral region. Phytopathology 41, 37 (1951)
4. Ullrich, J., Schrödter, H.: Das Problem der Vorhersage des Auftretens der Kartoffelkrautfäule (*Phytophthora infestans*) und die Möglichkeit seiner Lösung durch eine "Negativprognose". Nachrichtenblatt Dt. Pflanzenschutzdienst(Braunschweig) 18, 33–40 (1966)
5. Winstel, K.: Kraut- und Knollenfäule der Kartoffel – Eine neue Prognosemöglichkeit sowie Bekämpfungsstrategien. Med. Fac. Landbouww. Univ Gent 58/32b (1993)
6. Forrer, H.R., Gujer, H.U., Fried, P.M.: A comprehensive information and decision support system for late blight in potatoes. In: Workshop on Computer-based Decision Support System (DSS) in Crop Protection, Parma, Italy (November 1993)
7. Forsund, E.: Late blight forecasting in norway 1957-1980. EPPO Bull. 13, 255–258 (1983)
8. Henfling, J.W.: Late blight of potato: *Phytophthora infestans*. In: Technical Information Bulletin 4, International Potato Centre Lima, Peru (1987)
9. Pande, A., Jagyasi, B.G., Kimbahune, S., Doke, P., Mittal, A., Singh, D., Jain, R.: Mobile phone based agro-advisory system for agricultural challenges in rural india. In: IEEE Conference on Technology for Humanitarian Challenges (August 2009)
10. Seaman, A., Loria, R., Fry, W., Zitter, T.: What is late blight? In: Integrated Pest Management Program, Cornell University (2008), http://www.nysipm.cornell.edu/publications/blight/ (last access September 2008)

Evolutionary and Iterative Crisp and Rough Clustering I: Theory

Manish Joshi[1] and Pawan Lingras[2]

[1] Department of Computer Science, North Maharashtra University
Jalgaon, Maharashtra, India
joshmanish@gmail.com

[2] Department of Mathematics and Computing Science, Saint Mary's University
Halifax, Nova Scotia, B3H 3C3, Canada
pawan@cs.smu.ca

Abstract. Researchers have proposed several Genetic Algorithms (GA) based clustering algorithms for crisp and rough clustering. In this two part series of papers, we compare the effect of GA optimization on resulting cluster quality of K-means, GA K-means, rough K-means, GA rough K-means and GA rough K-medoid algorithms. In this first part, we present the theoretical foundation of the transformation of the crisp clustering K-means and K-medoid algorithms into rough and evolutionary clustering algorithms. The second part of the paper will present experiments with a real world data set, and a standard data set.

Keywords: Crisp Clustering, Rough Clustering, GA Rough K-means, GA Rough K-medoid.

1 Introduction

Clustering is an unsupervised learning process that partitions physical or abstract objects into groups based on some optimality criterion (e.g. similarity). The typical partition based clustering algorithms, K-means and K-medoids, categorize an object into precisely one cluster. Whereas, fuzzy clustering [1][13] and rough set clustering [5][12][14][16] provide ability to specify the membership of an object to multiple clusters, which can be useful in real world applications.

Deterministic local search converges to the nearest local optimum from a starting position of the search, hence K-means result largely depends on the initial cluster centers. On the contrary, stochastic search heuristics inspired by evolution and genetics have the ability to cope with local optima by maintaining, recombining and comparing several candidate solutions simultaneously. Use of GAs for clustering is proposed by [2][9][11]. GA guided evolutionary algorithms are also proposed for rough set and fuzzy clustering [8][12][15].

In this paper we discuss various aspects of crisp, rough set based, and evolutionary clustering algorithms. We discuss and present appropriate modifications required to the basic K-means and K-medoid algorithms so that these algorithms adapt to rough and evolutionary clustering. In particular, we explain K-means,

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 615–620, 2009.

GA K-means, rough K-means, GA rough K-means, and GA rough K-medoid algorithms and their corresponding fitness functions.

Section 2 describes how rough set theory is embedded with a K-means algorithm. Section 3 discusses the inclusion of GA to K-means, rough K-means, and rough K-medoid algorithms. We also discuss the formulation of a fitness function for GA based crisp and rough clustering in this section, followed by the conclusion in section 4. Due to space restrictions, an elaborate experimental comparison is presented separately in the second part of the series.

2 Rough Clustering

In addition to clearly identifiable groups of objects, it is possible that a data set may consist of several objects that lie on the fringes. The conventional clustering techniques mandate that such objects belong to precisely one cluster. Such a requirement is found to be too restrictive in many data mining applications [4]. In practice, an object may display characteristics of different clusters. In such cases, an object should belong to more than one cluster, and as a result, cluster boundaries necessarily overlap. Fuzzy set representation of clusters, using algorithms such as fuzzy C-means, makes it possible for an object to belong to multiple clusters with a degree of membership between 0 and 1 [13]. In some cases, the fuzzy degree of membership may be too descriptive for interpreting clustering results. Rough set based clustering provides a solution that is less restrictive than conventional clustering and less descriptive than fuzzy clustering.

Lingras and West [5] provided an alternative based on an extension of the K-means algorithm [3][10]. Incorporating rough sets into K-means clustering requires the addition of the concept of lower and upper bounds. The rough K-means approach has been a subject of further research. Peters [14] discussed various refinements of Lingras and West's original proposal [5]. These included calculation of rough centroids and the use of ratios of distances as opposed to differences between distances similar to those used in the rough set based Kohonen algorithm described in [6]. The rough K-means and its various extensions [12], [14] have been found to be effective in distance based clustering. However, there is no theoretical work that proves that rough K-means explicitly finds an optimal clustering scheme. Moreover, the quality of clustering that is maximized by the rough clustering is not precisely defined. We compare crisp and rough clustering algorithm results and present our observations in the second part of this paper.

Rough K-means Algorithm. We represents each cluster $c_i, 1 \leq i \leq k$, using its lower $\underline{A}(c_i)$ and upper $\overline{A}(c_i)$ bounds. All objects that are clustered using the algorithm follow basic properties of rough set theory such as:

(P1) An object $\boldsymbol{x}$ can be part of at most one lower bound

(P2) $\boldsymbol{x} \in \underline{A}(\boldsymbol{c}_i) \Longrightarrow \boldsymbol{x} \in \overline{A}(\boldsymbol{c}_i)$

(P3) An object $\boldsymbol{x}$ is not part of any lower bound

$$\Updownarrow$$

$\boldsymbol{x}$ belongs to two or more upper bounds.

```
Input:
    $k$: the number of clusters,
    $D(n, m)$: a data set containing n objects where each object has m dimensions,
    $p$: a threshold value (1.4),
    $w\_lower$: relative importance assigned to lower bound (0.75),
    $w\_upper$: relative importance assigned to upper bound (0.25),
Output:
    A set of clusters.Each cluster is represented by the objects in the lower region
    and in boundary region (upper bound)
Steps:
    arbitrarily choose $k$ objects from $D$ as the initial cluster centers (centroids);
    repeat
        (re)assign each object to lower/upper bounds of appropriate clusters by
            determining its distance from each cluster centroid,
        update the cluster centroids using number of objects assigned and relative
            importance assigned to the lower bound and upper bound of the cluster;
    until no change;
```

Fig. 1. The K-means algorithm for rough clustering

Fig. 1 depicts the general idea of the algorithm. An object is assigned to lower and/or upper bound of one or more clusters. For each object vector, $\boldsymbol{v}$, let $d(\boldsymbol{v}, \boldsymbol{c}_j)$ be the distance between itself and the centroid of a cluster $\boldsymbol{c}_j$. Let $d(\boldsymbol{v}, \boldsymbol{c}_i) = \min_{1 \le j \le k} d(\boldsymbol{v}, \boldsymbol{c}_j)$. The ratios $d(\boldsymbol{v}, \boldsymbol{c}_i)/d(\boldsymbol{v}, \boldsymbol{c}_j)$, $1 \le i, j \le k$, are compared with a cut-off value to determine the membership of an object $\boldsymbol{v}$. This parameter is called as a *threshold*. Let $T = \{j : d(\boldsymbol{v}, \boldsymbol{c}_i)/d(\boldsymbol{v}, \boldsymbol{c}_j) \le \mathit{threshold}$ and $i \ne j\}$.

1. If $T \ne \emptyset$, $\boldsymbol{v} \in \overline{A}(\boldsymbol{c}_i)$ and $\boldsymbol{v} \in \overline{A}(\boldsymbol{c}_j), \forall j \in T$. Furthermore, $\boldsymbol{v}$ is not part of any lower bound. The above criterion guarantees that property (P3) is satisfied.
2. Otherwise, if $T = \emptyset$, $\boldsymbol{v} \in \underline{A}(\boldsymbol{c}_i)$. In addition, by property (P2), $\boldsymbol{v} \in \overline{A}(\boldsymbol{c}_i)$.

It should be emphasized that the approximation space A is not defined based on any predefined relation on the set of objects. The lower and upper bounds are constructed based on the criteria described above.

The values of p(a threshold), *w_lower*, *w_upper* are finalized based on the experiments described in [7].

3 Evolutionary Clustering Algorithms

This section contains some of the basic concepts of genetic algorithms as described in [2]. A genetic algorithm is a search process that follows the principles of evolution through natural selection. The domain knowledge is represented using a candidate solution called a *chromosome*. Typically, a chromosome is a single *genome* represented as a vector of length n:

$$g = (g_i \mid 1 \le i \le n), \tag{1}$$

where g_i is called a *gene*.

A group of chromosomes is called a *population*. Successive populations are called *generations*. A generational GA starts from initial generation $G(0)$, and each generation $G(t)$ generates a new generation $G(t+1)$ using genetic operators such as *mutation* and *crossover*. The mutation operator creates new genomes by changing values of one or more genes at random. The crossover operator joins segments of two or more genomes to generate a new genome.

The evaluation process of a genome i.e. evaluate $G(t)$, is a combination of two steps. The first step determines membership of all objects to corresponding clusters. As described in earlier section, appropriate calculations for crisp and rough clustering figure out the members of each cluster. In the next step, fitness of the genome is determined. The intuitive distance measure is used to decide the fitness of the genome. Obviously, there is a difference between the fitness calculations for crisp clustering and rough clustering. The fitness formulas used for both clustering are described below.

Genome Fitness Function for Crisp Clustering. The fitness is calculated based on the allocation of all objects to the clusters. It is given by:

$$Fitness \ = \ \sum_{i=1}^{k} \sum_{u \in c_i} d(u, x_i). \tag{2}$$

$Fitness$ is the sum of the Euclidean distances for all objects in the cluster; u is the point in space representing a given object; and x_i is the centroid/medoid of cluster c_i (both u and x_i are multidimensional). The function d computes distance between any two vectors u and v using following equation.

$$d(u, v) \ = \ \sqrt{\sum_{j=1}^{m} (u_j \ - \ v_j)^2}. \tag{3}$$

Here, the value of m indicates the total number of dimensions.

Genome Fitness Function for Rough Clustering. The Fitness function has to change to adapt to the rough set theory by creating lower and upper versions of the Fitness as:

$$\underline{Fitness} \ = \ \sum_{i=1}^{k} \ \sum_{u \in \underline{A}(c_i)} d(u, x_i), \tag{4}$$

$$\overline{Fitness} \ = \ \sum_{i=1}^{k} \ \sum_{u \in \overline{A}(c_i)} d(u, x_i), \tag{5}$$

where $\underline{A}(c_i)$ and $\overline{A}(c_i)$ represents lower and upper bound of cluster c_i. The distance function d does not change. The $Fitness$ value for the rough clustering is calculated as

$$Fitness \ = \ w\_lower \times \underline{Fitness} + w\_upper \times \overline{Fitness}. \tag{6}$$

where *w_lower* and *w_upper* are relative importances assigned to lower and upper bound of the clusters.

Thus evolutionary algorithms for clustering differ mostly in the genome representation and the fitness calculation. The variations of GA based algorithms for crisp and rough clustering are described in next subsections.

3.1 GA K-means

The crisp K-means algorithm is modified to adapt the principles of GA. The chromosome has a total of $k \times m$ genes. A batch of every m genes represents centroid of a corresponding cluster. The population size and generation values are input parameters.

3.2 GA Rough K-means

The proposed genome for the evolutionary algorithm has a total of $k \times m$ genes, where k is the desired number of clusters and m is the number of dimensions used to represent objects and centroids. The first m genes represent the first centroid. Genes $m+1, \ldots, 2 \times m$ give us the second centroid, and so on. Finally, $((k-1) \times m) + 1, \ldots, k \times m$ corresponds to the k^{th} centroid. The rough fitness measure given by Eq. 6 is minimized.

3.3 GA Rough K-medoid

Unlike K-means algorithm where mean value is used as a centroid of a cluster, in K-medoid algorithm actual object is used as a reference point of a cluster [15]. A medoid is the most centrally located object in a given cluster. For k clusters, we have k medoids. A genetic algorithm is used to search for the most appropriate k medoids. The genome has k genes, each corresponding to a medoid. This reduces the size of a genome from $k \times m$ by a factor of m to k. The steps in this algorithm are similar to that of GA rough K-means. The major difference is the use of medoids instead of centroids of the clusters.

The gene values for rough K-medoids are discrete values corresponding to object IDs as opposed to continuous real variables used for centroids in the rough K-means evolution. This results in a restricted search space for the proposed rough K-medoids leading to further increase in the chances of convergence. We have tested this hypothesis in our second part of the paper.

4 Conclusion

In this paper - the first part of a two part series - we discuss how crisp clustering K-means and K-medoid algorithms adapt to rough and evolutionary clustering. The paper describes algorithmic steps for rough K-means, GA K-means, GA rough K-means and GA rough K-medoid. We also discuss the fitness function for rough and evolutionary algorithms to determine the quality of clustering. The second part of this paper will elaborate on an experimental comparison with a real world data set, and a standard data set.

References

1. Bezdek, J.C., Hathaway, R.J.: Optimization of Fuzzy Clustering Criteria using Genetic Algorithms (1994)
2. Buckles, B.P., Petry, F.E.: Genetic Algorithms. IEEE Computer Press, Los Alamitos (1994)
3. Hartigan, J.A., Wong, M.A.: Algorithm AS136: A K-Means Clustering Algorithm. Applied Statistics 28, 100–108 (1979)
4. Joshi, A., Krishnapuram, R.: Robust Fuzzy Clustering Methods to Support Web Mining. In: Proc. ACM SIGMOD Workshop Data Mining and Knowledge Discovery, June 1998, pp. 1–8 (1998)
5. Lingras, P., West, C.: Interval Set Clustering of Web Users with Rough K-Means. Journal of Intelligent Information Systems 23, 5–16 (2004)
6. Lingras, P., Hogo, M., Snorek, M.: Interval Set Clustering of Web Users using Modified Kohonen Self-Organizing Maps based on the Properties of Rough Sets. Web Intelligence and Agent Systems: An International Journal 2(3), 217–230 (2004)
7. Lingras, P.: Precision of rough set clustering. In: Chan, C.-C., Grzymala-Busse, J.W., Ziarko, W.P. (eds.) RSCTC 2008. LNCS (LNAI), vol. 5306, pp. 369–378. Springer, Heidelberg (2008)
8. Lingras, P.: Evolutionary rough K-means Algorithm. In: Proc. The Fourth International conference on Rough Set and Knowledge Technology, RSKT2009 (2009)
9. Lu, Y., Lu, S., Fotouhi, F., et al.: FGKA: A Fast Genetic K-Means Clustering Algorithm. In: Proc. ACM Symposium on Applied Computing (2004)
10. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
11. Maulik, U., Bandyopadhyay, S.: Genetic Algorithm-Based Clustering Technique. Pattern Recognition 33, 1455–1465 (2000)
12. Mitra, S.: An Evolutionary Rough Partitive Clustering. Pattern Recognition Letters 25, 1449 (2004)
13. Pedrycz, W., Waletzky, J.: Fuzzy Clustering with Partial Supervision. IEEE Trans. on Systems, Man and Cybernetics 27(5), 787–795 (1997)
14. Peters, G.: Some Refinements of Rough k-Means. Pattern Recognition 39, 1481–1491 (2006)
15. Peters, G., Lampart, M., Weber, R.: Evolutionary rough k-medoid clustering. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets VIII. LNCS, vol. 5084, pp. 289–306. Springer, Heidelberg (2008)
16. Peters, J.F., Skowron, A., Suraj, Z., et al.: Clustering: A Rough Set Approach to Constructing Information Granules, pp. 57–61 (2002)

Evolutionary and Iterative Crisp and Rough Clustering II: Experiments

Manish Joshi[1] and Pawan Lingras[2]

[1] Department of Computer Science, North Maharashtra University,
Jalgaon, Maharashtra, India
joshmanish@gmail.com
[2] Department of Mathematics and Computing Science, Saint Mary's University,
Halifax, Nova Scotia, B3H 3C3, Canada
pawan@cs.smu.ca

Abstract. In this second part of the paper, we compare the cluster quality of K-means, GA K-means, rough K-means, GA rough K-means and GA rough K-medoid algorithms. We experimented with a real world data set, and a standard data set using total within cluster variation, precision and execution time as the measures of comparison.

Keywords: Rough Clustering, Crisp Clustering, GA based Clustering, Cluster Quality.

1 Introduction

In this second part of our two part series, we discuss and compare results of crisp, rough set based, iterative and evolutionary clustering algorithms [3]. In particular, we compare performance of K-means, GA K-means, rough K-means, GA rough K-means, and GA rough K-medoid algorithms using appropriate evaluation metrics.

Latiff et al. [2] compared GA based algorithm with K-means and PSO in wireless sensor networks domain. Małyszkoa et al. [5] compared K-means and GA K-means algorithms for image segmentations. These two studies, like most of the other comparative studies, concentrated on crisp clustering. This paper compares evolutionary and iterative algorithms for both crisp as well as rough clustering using K-means and K-medoid clustering strategies.

Section 2 describes the evaluation metrics used for comparison. Comparative results for two different types of data set are presented and discussed in two subsequent sections, followed by the conclusions in section 5.

2 Evaluation Metrics

In order to compare the results obtained using evolutionary and iterative versions of crisp and rough clustering we use a *total within cluster variation* measure. The measure accumulates distances between a cluster center (cenroid/medoid) and objects assigned to a cluster.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 621–627, 2009.

$$Fitness = \sum_{i=1}^{k} \sum_{u \in c_i} d(u, x_i) . \tag{1}$$

Fitness is the sum of the Euclidean distances for all objects in the cluster; u is the point in space representing a given object; and x_i is the centroid/medoid of cluster c_i (both u and x_i are multidimensional). The function d provides the distance between two vectors. The distance $d(u, v)$ is given by:

$$d(u, v) = \sqrt{\sum_{j=1}^{m} (u_j - v_j)^2} . \tag{2}$$

Here the value of m indicates the total number of dimensions.

The *Fitness* function has to adapt to the rough set theory by creating lower and upper approximations of the *Fitness* as:

$$\underline{Fitness} = \sum_{i=1}^{k} \sum_{u \in \underline{A}(c_i)} d(u, x_i) , \tag{3}$$

$$\overline{Fitness} = \sum_{i=1}^{k} \sum_{u \in \overline{A}(c_i)} d(u, x_i) , \tag{4}$$

where $\underline{A}(c_i)$ and $\overline{A}(c_i)$ represent lower and upper bounds of cluster c_i. The *Fitness* value for the rough clustering is calculated as

$$Fitness = w\_lower \times \underline{Fitness} + w\_upper \times \overline{Fitness} . \tag{5}$$

where *w_lower* and *w_upper* are relative importance assigned to lower and upper bound of the clusters. Smaller the value of fitness, better the cluster quality.

3 Comparative Results – Real World Data Set

We used the data obtained from a public library to compare the results of various algorithms. The data consist of books borrowed by members. The objective is to group members with similar reading habits. Information about how many times a member borrows books of a particular category is collected. The data is normalized in the range of 0 to 1 to reduce the effect of outliers. In order to visualize the data set, it is restricted to two dimensions. Data of 1895 members shows propensity of a member to borrow a book of a certain category. It will be interesting to see how all the algorithms carve out reasonable clusters from such scattered data.

The Gas used the crossover probability of 70% and mutation probability of 10%. We carried experiments with different population sizes and generations. Table 1 shows the results obtained using K-means and rough K-means algorithms. Clusters generated by rough K-means have less intra-cluster variation than for the clusters generated by K-means algorithm. Hence we can conclude that the rough K-means results in better cluster quality than the crisp K-means algorithm. We have observed similar trends for the synthetic data set.

Table 1 also shows the comparison of K-means against GA K-means and rough K-means against GA rough K-means algorithms. The results include average *Fitness* from five trials. For a normal range of population size and generations the GA K-means does not outperform the K-means. But for population size of 500 and 500 generations the average *Fitness* of GA K-means for 3 clusters is less than K-means. This performance improvement requires 90 seconds of computation. For five clusters the GA K-means (population size of 500 and 500 generations) could outperform K-means at the cost of 111.2 seconds of processing time.

As the data set size increases the population size and generations of GA should be increased to obtain improved *Fitness*. In some cases, the higher computational cost of GA K-means may not be justified by slight increase in accuracy.

Table 1 shows that GA rough K-means improves the results of rough K-means. Moreover, GA rough K-means does this with far less population size and generations than GA K-means.

Both GA rough K-medoid and GA rough K-means results are better than the K-means results. The GA rough K-medoid algorithm is faster than the GA rough K-means in surpassing the K-means results (see Table 2). GA rough K-means requires

Table 1. Comparison between crisp and evolutionary algorithms for Library data

Average K-means Fitness	Population size, Generations	GA K-means		Average Rough K-means Fitness	GA Rough K-means	
		Average Fitness	Avg. Time (Sec.)		Average Fitness	Avg. Time (Sec.)
3 Clusters						
7.9764	10, 20	9.1223	1	6.9517	8.094	1.2
	20, 20	8.6642	1		7.82	2.4
	20, 30	8.3182	1.4		7.12	2.4
	30, 50	8.0323	1.6		**6.91**	**4.4**
	50, 50	8.0261	2.8		6.9001	6
	100, 100	7.9780	6.8		6.9006	19
	500, 500	**7.9761**	**90**		-	-
5 Clusters						
5.9020	10, 20	8.9841	1	5.2061	7.1033	1.4
	20, 20	6.7930	1.8		6.7505	2.8
	20, 30	7.4541	1.8		6.2648	3
	30, 50	6.3575	2.8		5.8794	4.6
	50, 50	6.4889	3.2		5.2801	8
	100, 100	5.9527	16.4		**5.0730**	**25.8**
	500, 500	**5.9002**	**111.2**		**-**	**-**

Table 2. Comparison of K-means, GA rough K-means and GA Rough K-medoid for Library Data set

K-means Fitness	Population size, Generations	GA Rough K-means		GA Rough K-medoid	
		Fitness (Best / Average)	Avg. Time (Sec.)	Fitness (Best / Average)	Avg. Time (Sec.)
3 Clusters					
7.9764	10, 20	8.094	1.2	**7.7042**	**1.2**
	20, 20	7.82	2.4	7.1390	2.2
	20, 30	7.12	2.4	7.2670	2.8
	30, 50	**6.91**	**4.4**	7.0719	4
	50, 50	6.9001	6	7.0154	6.4
	100, 100	6.9006	19	7.0465	19
5 Clusters					
5.9020	10, 20	7.1033	1.4	6.0141	2
	20, 20	6.7505	2.8	**5.3114**	**3**
	20, 30	6.2648	3	5.3502	3.2
	30, 50	**5.8794**	**4.6**	5.1660	5
	50, 50	5.2801	8	5.1770	8.6
	100, 100	5.0730	25.8	5.1034	21.4

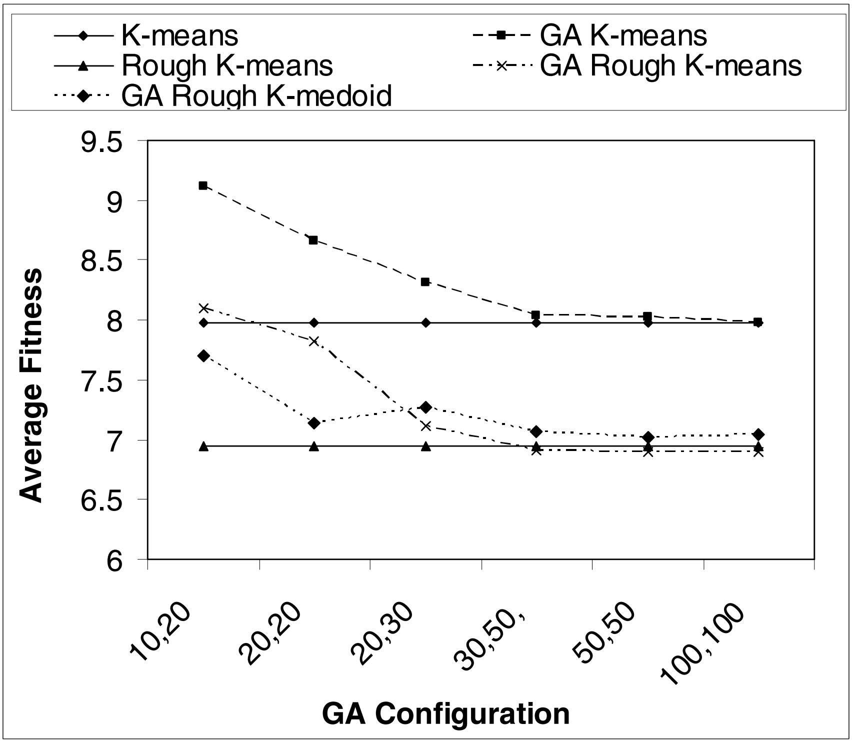


Fig. 1. Comparison chart of average fitness obtained by different algorithms

4.6 seconds with 30 population size and 50 generations to get better solution than K-means. Whereas GA rough K-medoid with 20 population size and 20 generations surpasses the K-means results in 3 seconds.

Fig. 1 shows the performance of GAs for different configurations. GA K-means performs better than K-means for large number of population size and generations. GA rough K-medoid algorithm promptly outdoes the K-means results, but for higher population size and generations GA rough K-means generates optimal results.

4 Comparative Results – Standard Data Set

We used *Letter Recognition* [1] data from the University of California Irvine machine learning repository. The data set contains 20,000 events representing character images of 26 capital letters in the English alphabet based on 20 different fonts. Each event is represented using 16 primitive numerical attributes (statistical moments and edge counts) that are scaled to fit into a range of integer values from 0 through 15.

In order to crosscheck the results and compare the cluster quality we decided to consider limited number of characters. We prepared a data set that consists of 8 distinctly different characters A, H, L, M, O, P, S, and Z. This data set consists of 6114 events.

Besides using the *Fitness* measure for comparison we calculated the average precision for each cluster. Each cluster is labelled using the character that appears most frequently in the cluster. Precision of the cluster is defined as the number of events that match the cluster label divided by the total number of events in the cluster. Average precision is the average of all cluster precisions.

Table 3 and Table 4 show the comparison of results obtained using K-means and rough K-means algorithms for an experimental data set.

Rough K-means clustering generates better quality clusters than the crisp K-means algorithm. This conclusion is supported by reduced *Fitness* and increase in average precision. Moreover, only one letter 'H' is not prominently identified by rough K-means whereas two letters 'H' and 'S' are not prominently identified by any clusters generated using K-means algorithm.

Table 3. K-means algorithm clustering for an experimental set of characters

Cluster No	Cluster Label Character	Frequency	Precision	Average Precision	*Fitness*
0	M	412/1119	0.37	0.60	406.23
1	Z	276/880	0.31		
2	Z	378/734	0.51		
3	P	607/639	0.95		
4	A	387/660	0.59		
5	A	311/514	0.61		
6	L	328/328	1.0		
7	O	600/1240	0.48		

Table 4. Rough K-means algorithm clustering for an experimental set of characters

Cluster No	Cluster Label Character	Frequency	Precision	Average Precision	*Fitness*
0	L	335/335	1.0	0.82	372.23
1	O	553/684	0.81		
2	Z	536/705	0.76		
3	A	183/255	0.72		
4	A	358/362	0.99		
5	S	96/294	0.33		
6	M	292/306	0.95		
7	P	553/555	1.0		

GA based algorithms however generate poor results for this standard data set. For crisp as well as for rough clustering, most of the objects are grouped into few clusters. The remaining clusters are left empty. Further investigations are necessary to determine the reasons for failure of GAs for the character data set.

5 Conclusions

In this second part of a two part series, we provide experimental comparison of results obtained by K-means, GA K-means, rough K-means, GA rough K-means and GA rough K-medoid algorithms. We applied all algorithms to a synthetic data set, a real world data set, and a standard data set. A simple and intuitive measure of total within cluster variation (Fitness) is used for the evaluation.

The rough K-means algorithm seems to provide better cluster quality in terms of the Fitness and average precision than the crisp K-means algorithm.

For sufficiently high population size and generations, GA K-means can improve average performance of K-means. As the size of data set increases, higher population size and generations are required in GA K-means algorithms to outperform the K-means results. Execution time increases when GAs with higher population size and generations are used. There is a trade off between execution time and better cluster quality when the GA K-means algorithm is used.

GA rough K-medoid converges faster and surpasses the K-means results with smaller populations and fewer generations than GA rough K-means. But for larger population and more generations the GA rough K-means results are superior to all other algorithms.

GA effect on rough clustering is more promising than that on crisp clustering. For both small as well as large data sets, the GA rough K-means and the GA rough K-medoid generate better clustering in reasonable amount of execution time.

Unexpectedly though, the GA version could not cope up with the data set that has 12 dimensions. We shall test by initializing the genome of the GA with the centroids generated by the basic K-means result. A better starting point may help GAs to reproduce optimal result.

Acknowledgments. This work is partially supported by an ACE-Net post-doctoral fellowship and NSERC, Canada.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, Irvine, CA (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Latiff, N.M.A., Tsimenidis, C.C., Sharif, B.S.: Performance Comparison of Optimization Algorithms for Clustering in Wireless Sensor Networks (2007)
3. Lingras, P.: Applications of Rough Set Based K-Means, Kohonen SOM, GA Clustering. In: Peters, J.F., Skowron, A., Marek, V.W., Orłowska, E., Słowiński, R., Ziarko, W.P. (eds.) Transactions on Rough Sets VII. LNCS, vol. 4400, pp. 120–139. Springer, Heidelberg (2007)
4. Lingras, P., Chen, M., Miao, D.: Rough Multi-category Decision Theoretic Framework. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 676–683. Springer, Heidelberg (2008)
5. Małyszkoa, D., Wierzchoń, Sławomir, T.: Standard and Genetic k-means Clustering Techniques in Image Segmentation (2007)

Author Index