



#### **Areas of Research Strength**

Intelligent Informatics (I<sup>2</sup>)

Agents, Machine-learning, and Data-mining

Networking and

Recognition

#### Focusing on

Advanced Theories, Methods, and Algorithms of Information Processing and Communications for Adaptive Interfaces, Systems, and Networks

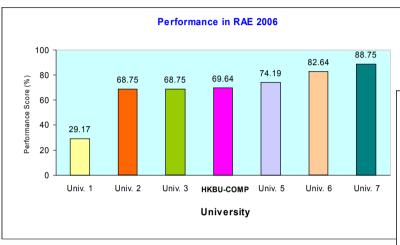
#### Research & Development Laboratories

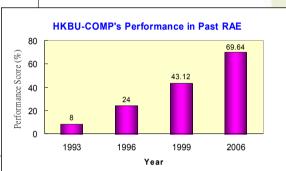
Laboratories	Expertise
1. Agents, Machine-learning, and Data-mining (AMD)	<ul> <li>Autonomous Agents and Multi-agent Systems</li> <li>Data Mining and Machine Learning</li> <li>Intelligent Software Tools and Systems</li> <li>On-line Media and Network Analysis</li> <li>Web Intelligence</li> </ul>
2. Networking and Multimedia	<ul> <li>All-optical Networks</li> <li>IP Networks</li> <li>Mobile Computing</li> <li>Multimedia Communications and Systems</li> <li>Visual Information Systems</li> <li>Wireless Communications</li> </ul>
3. Pattern Recognition	<ul> <li>Image Processing</li> <li>Vision-based Human Understanding Technology</li> <li>Wavelet Analysis and its Applications</li> </ul>



#### Research Assessment Exercise (RAE) 2006

The Department has achieved very good research performance in the past several years, as confirmed in the Research Assessment Exercise 2006 conducted by the UGC. Our research performance scores among local universities in the 2006 exercise and in the past exercises are presented as follows:





#### **Research Outputs**

Number of Refereed International Journal Papers

	No. of Papers	Average No. of Papers per Faculty
2005-2006	38	2.714
2004-2005	42	3.50
2003-2004	32	2.285

### Monographs, Books, & Proceedings Written or Edited, 2003-2006

Monographs	5
Edited Books & Proceedings	12
Total	17









#### **Editors-in-Chief:**

Annual Review of Intelligent Informatics (World Scientific)

International Journal of Wavelets, Multiresolution and Information Processing (World Scientific)

The IEEE Intelligent Informatics Bulletin (IEEE Computer Society)

Web Intelligence and Agent Systems (IOS Press, The Netherlands)

#### **Editorial Boards:**

IEEE Transactions on Knowledge and Data Engineering

IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews

International Journal of Applied Systematic Studies (Inderscience)

International Journal of Document Analysis and Recognition (Springer)

International Journal of Pattern Recognition and Artificial Intelligence (World Scientific)

Journal of Data Warehousing and Mining (IGI Global)

Journal of Embedded Computing (IOS Press)

Journal of Mobile Multimedia (Rinton Press)

Journal of Multimedia Tools and Applications (Springer)

Journal of Pervasive Computing and Communications (Troubador)

Journal of Ubiquitous Computing and Intelligence (American Scientific Publishers)

Knowledge and Information Systems (Springer)

Microprocessor and Microsystems (Elsevier)

Pattern Recognition (Elsevier)

Real-Time Systems: The International Journal of Time-Critical Computing Systems (Kluwer)

#### ... in High-Impact Journals

Artificial Intelligence

Communications of the ACM

**IEEE Computer** 

**IEEE Intelligent Systems** 

**IEEE Internet Computing** 

IEEE Journal of Lightwave Technology

IEEE Journal of Selected Areas in Communications

IEEE Multimedia

IEEE Transactions on Circuits and Systems for Video Technology

**IEEE Transactions on Communications** 

**IEEE Transactions on Computers** 

IEEE Transactions on Evolutionary Computation

IEEE Transactions on Image Processing

IEEE Transactions on Knowledge and Data Engineering

IEEE Transactions on Medical Imaging

IEEE Transactions on Mobile Computing

IEEE Transactions on Multimedia

**IEEE Transactions on Neural Networks** 

IEEE Transactions on Parallel and Distributed Computing

IEEE Transactions on Pattern Analysis and Machine Intelligence

IEEE Transactions on Signal Processing

IEEE Transactions on Systems, Mans and Cybernetics

**IEEE/ACM Transactions on Networking** 

Journal of Systems and Software

Pattern Recognition

Real-Time Systems Journal



**10**5



#### Major Conferences Organized/Co-organized, 2003-2006

#### Including ...

**Computer Science** 

- 2006 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology ((WI 2006 & IAT 2006)
- 2006 IEEE International Conference on Data Mining (ICDM 2006)
- 18th International Conference on Pattern Recognition (ICPR 2006)
- 2005 International Conference on Computational Intelligence and Security (CIS 2005)
- 2005 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2005 & IAT 2005)
- 11th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2005)
- 2005 IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE 2005)
- 2004 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2004 & IAT 2004)
- 2004 IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE 2004)
- 2003 IEEE/WIC International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2003 & IAT 2003)
- 4th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2003)



Enroute to Excellence



#### Centre for e-Transformation Research (CTR)

(An affiliated research centre of Web Intelligence Consortium)

#### Mission

To be a major contributor to Hong Kong's knowledge base in researching the new fundamental roles and practical impacts of Artificial Intelligence (AI) and advanced Information Technology (IT) on the next generation of Web-empowered products, systems, services, and activities

#### Milestones

- 2003: Established under the Faculty of Science
- 2004: Funded by HK Research Grants Council (first IT group research project funded)
- 2007: Advanced e-Transformation Technology Research becomes a University's strategic research area

#### Research Team

#### **Director:**

Jiming Liu, Head and Professor Computer Science Department, HKBU

#### **Area Leaders:**

- Qiang Yang, Professor
   Computer Science Department, HKUST
- C. J. Tan, Director and Chair Professor
   E-Business Technology Institute (ETI), HKU

#### **Coordinator:**

William K. Cheung, Associate Professor Computer Science Department, HKBU

#### **Key Members:**

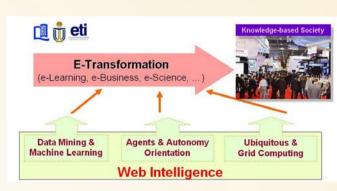
From HKBU, HKUST, and ETI-HKU

#### Research Focus

# - Agents and Autonomy Oriented Computing - Data Mining and Machine Learning - Grid and Service Oriented Computing - Grid and Service Oriented Computing - Community Analysis in the Networked Society - Decision Technologies for Supply Chain Management - Services Management for Computation, Data Analysis, and Knowledge Inference

#### **Publications**

- Communications of the ACM
- IEEE Internet Computing
- IEEE Transactions on Knowledge and Data Engineering
- Journal of Data Mining and Knowledge Discovery
- SIGKDD Exploration
- Proceedings of prestigious international conferences, such as AAAI, ICDM, ICML, ISCAI, and WI/IAT



Centre URL: http://www.comp.hkbu.edu.hk/~ctr

### Research Centre for Ubiquitous Computing



#### Mission

- To excel in some focused areas related to ubiquitous/pervasive computing
- To coordinate and develop research activities towards ubiquitous/pervasive computing

 To construct test-beds based on different positioning technologies for testing & benchmarking and to develop software which becomes the common basic building blocks for ubiquitous/ paragive applications

Self-Organizing Map (SOM)

pervasive applications

#### Research Team

#### **Principal Investigator:**

Joseph Kee-Yin Ng, Professor Computer Science Department, HKBU

#### Co-Investigator:

Lionel Ni, Head and Professor Computer Science Department, HKUST

#### **Key Members:**

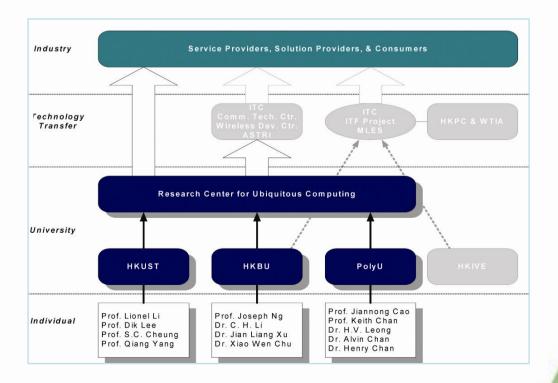
From HKBU, HKUST, and PolyU

# the nearest Chinese seafood restaurant? Stelly Stell

# Location System by use of Cellphones and Fleet Management Neural Network

#### Research Focus

- Location-aware computing with a focus on mobile and wireless technology
- RFID and sensor networks with a focus on positioning and data access
- The design and construct of API / Middleware for ubiquitous computing
- Software development on providing location-based services





For more information about the Research Centre for Ubiquitous Computing, please contact Prof. Joseph Ng at jng@comp.hkbu.edu.hk or (852)-3411-7864

### Awards



**Prof. Liu Jiming** 

President's Award
for Outstanding Performance 2007
in recognition of his Outstanding Performance
in Scholarly Work





**Prof. Tang Yuan-yan** 

First Class of Natural Science Award of Technology Development Centre Ministry of Education of the People's Republic of China in 2005

Elected Fellow of the IEEE in January 2004 and Fellow of IAPR in August 2004

. . . . . . . . . . .







President's Award
for Outstanding Performance 2003
in recognition of his Outstanding Performance
in Scholarly Work



Dr. Feng Jian

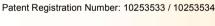
Most Cited Paper Award for the Journal of Visual Communication and Image Representation 2004-2006





Prof. Ng Kee-yin, Joseph

Together with his two German partners, Prof. Wolfgang Halang and Dr. Thomas Erdner and his graduate student Mr. Stephen Ka Chun Chan, received two **patents** issued by the German Patent Office for their research work entitled "Clock Synchronization in Nodes on a Ring Bus", and "Real Time Capable and Fault Tolerant Data Transfer" granted by The Germany Patent Office, Munchen, 9 February 2006 and 22 March 2007, respectively.





IEEE Computer Society
Meritorious Service Award 2004

Together with his two students Mr. Kan Ka-ho Kenny,
Mr. Chan Ka-chun Stephen, received
"AINA2003 Excellent Paper Award"





**Mr. Tong Tat Ming** 

under the supervision of Prof. Tang Yuan-yar

First-grade Award
in the Pan-Pearl River Delta Region
Amway Universities IT Project
Competition 2007





Mr. Chan Ka-ho Mr. Chan Kai-kin Mr. Yeung Man-chung

under the supervision of Prof. Ng Kee-yin, Joseph

2nd Runners-up in the Microsoft Imagine Cup 2007





Mr. Yeung Man-chung

under the supervision of Prof. Ng Kee-yin, Joseph

Distinction Award
in Intervarsity Internet Technologies Exposition
and Conference 2005

Certificate of Merit (Post-Secondary) of the 7th IT Excellence Awards by Hong Kong Computer Society in 2005



### S Q $\boldsymbol{\omega}$ ರ 0 ರಾ 0

### Agents, Machine-learning, and Data-mining

coin 001 COIN.cls February 8, 2005 19:54

Computational Intelligence, Volume 21, Number 2, 2005

ON KNOWLEDGE GRID AND GRID INTELLIGENCE: A SURVEY

WILLIAM K. CHEUNG AND JIMING LIU Computer Science Department, Hong Kong Baptist University, Kowloon Tong, Hong Kong

The next generation Web Intelligence (WI) aims at enabling users to go beyond the existing online information search and knowledge queries functionalities and to gain, from the Web, 1 practical wisdom for problem solving. To search and knowledge queries functionalities and to gain, from the web, 'practical wisdom for problem solving. To support such a Wisdom Web, we envision that a grid-like computing infrastructure with intelligent service agencies is needed, where these agencies can interact, self-organize, learn, and evolve their course of actions, identities, and interrelationships for new knowledge creation, as well as scientific and social evolution. In this paper, we first provide an overview of recent development in WI and Semantic/Knowledge Grid. Then, the fundamental capabilities of the Wisdom. Web as well as the conceptual architecture of an intelligent Grid for supporting it are described. Technical challenges for realizing Grid Intelligence are highlighted and the recent advancements in related research areas are reviewed.

Key words: Wisdom Web, Grid Intelligence, Knowledge Grid, autonomy-oriented computing.

1.1. Web Intelligence and Wisdom Web

The Web has irrevocably revolutionized the world we live in. This impact is inevitable due to the facts that the Web connectivity rapidly increases and that the online information astronomically explodes. In order not only to live with such a change but also to benefit from the information infrastructure that the Web has empowered, we have witnessed the fast development as well as applications of many Web Intelligence (WI) techniques and technologies (Zhong, Liu, and Yao 2003), which cover:

1. INTRODUCTION

1. Internet-level communication, infrastructure, and security protocols. The Web is tegarded as a computer-networked system. WI techniques for this level include, for instance, Web data-prefetching systems built upon Web-surfing patterns to resolve the issue of Web latency. The intelligence of the Web prefetching comes from adaptive learning based on observations of user-surfing behavior.

2. Interface-level multimedia presentation standards. The Web is regarded as an interface for human-Internet interaction. WI techniques for this level are used to develop the intelligent Web interfaces in which the capabilities of adaptive cross-language processing, personalized multimedia representation, and multimodel data processing are required. Knowledge-level information processing and management tools. The Web is regarded as a distributed data/knowledge base. We need to develop semantic markup languages to represent the semantic contents of the Web available in machine-understandable for-

mats for agent-based computing, such as searching, aggregation, classification, filtering, managing, mining, and discovery on the Web (Berners-Lee, Hendler, and Lassila 2001). Application-level ubiquitous computing and social intelligence environments. The Web is regarded as a basis for establishing social networks that contain communities for establishing social networks that contain communities of people (or organizations or other social entities) connected by social relationships, such as friendship, coworking, or information exchange with common interests. They are Web-supported social networks or virtual communities. The study of WI concerns the important issues central to social

Address correspondence to William K. Cheung, Computer Science Department, Hong Kong Baptist University, Kowloon Tong, Hong Kong; e-mail: william@comp.hkbu.edu.hk

1 Here, the notion of "Web" should be taken in a broader sense.

© 2005 Blackwell Publishing, 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford OX4 2DQ, UK.



### Service-Oriented **Distributed Data Mining**

Data mining research currently faces two great challenges; how to embrace data mining services with just-in-time and autonomous properties and how to mine distributed and privacy-protected data. To address these problems, the authors adopt the Business Process Execution Language for Web Services in a serviceoriented distributed data mining (DDM) platform to choreograph DDM component services and fulfill global data mining requirements. They also use the learning-from-abstraction methodology to achieve privacy-preserving DDM. Finally, they illustrate how localized autonomy on privacy-policy enforcement plus a bidding process can help the service-oriented system self-organize.

ost data mining algorithms assume market. This real-time objective imposes

William K. Cheung, Xiao-Feng Zhang, Ho-Fai Wong, and Jiming Liu Hong Kong Baptist University Zong-Wei Luo

E-Business Technology Institute,

University of Hong Kong

and Frank Tong

that data analysts will aggregate additional requirements on distributed data extracted from production data mining (DDM), including providing systems at a server for subsequent com- on-demand and self-adaptive services so putationally intensive data-crunching that companies can cope with heteroprocesses. However, issues such as data geneities in data sources, with respect to privacy concerns (with respect to cus- data privacy requirements, which aren't tomer information stored in bank servers, always known in advance. for example) and limits on data transmis- We can address these challenges in two sion bandwidth (affecting terabytes of sci- ways: a distributed computing architecture entific data generated from remote lab can support seamless provision, integra-

instruments or supercomputers) demon- tion, and coordination of just-in-time and strate that aggregating data for central- autonomous data mining services, and a ized mining simply isn't possible in a privacy-conscious DDM methodology can growing number of cases. Instead, it's work on top of this architecture. become necessary to develop methodolo- In this article, we describe our recent gies for mining distributed data that must efforts to create a novel DDM methodol-

remain private. In addition, being able to ogy known as learning from abstraction get the right information at the right time on a service-oriented platform in which (with respect to real-time business intelli- the underlying processes are specified in gence, for example) is an important busi- the Business Process Execution Lanness strategy in today's highly dynamic guage for Web Services (BPEL4WS). We

JULY . AUGUST 2006

Published by the IEEE Computer Society 1089-7801/06/\$20.00 © 2006 IEEE

IEEE INTERNET COMPUTING

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 34, NO. 1, JANUARY 2004

#### Correspondence\_

**Extended Latent Class Models for Collaborative** Recommendation

Kwok-Wai Cheung, Kwok-Ching Tsui, and Jiming Liu

Abstract—With the advent of the World Wide Web, providing just-in-time personalized product recommendations to customers now Just-In-time personance product recommender systems utilize correlation between customer preference ratings to identify "like-minded" customers and predict their product preference. One factor determining the success of the recommender systems is the prediction accuracy, which in many cases is limited by lacking adequate ratings (the sparsity problem). Recently, the use of latent class model (LCM) has been proposed to alleviate this problem. In this paper, we first study how the LCM can be extended to handle customers and products outside the training set. In tion, we propose the use of a pair of LCMs (called dual latent class model—DLCM), instead of a single LCM, to model customers' likes and dislikes separately for enhancing the prediction accuracy. Experimental results based on the EachMovie dataset show that DLCM outperforms available ratings are sparse.

Index Terms—Collaborative fitering, latent class models (LCMs), per-

I. Introduction

Product recommendation is one of the most important business activities for attracting customers. With the advent of the World Wide Web, online companies can now recommend products to their customers on a one-to-one basis in real time, and more importantly, at a much lower cost. Different recommender systems have been proposed in the literature [1], [2] and related products/services have also been released in the market (e.g., Andromedia.com, Netperception.com). Based on the uation. It was found that DLCM outperforms LCM and a conventional underlying technology, recommender systems can be broadly categorized as content-based or collaborative.

Content-based recommender systems match customer interest pro

files (e.g., revealed by their highly rated products) with the product attributes (or features) when making recommendations. Different machine learning [3], [4] and information retrieval [5], [6] algorithms have been proposed for profile representation and ratings prediction. One successful application of the content-based approach is personalized Web pages recommendation (e.g., Letizia [7]). In order for the approach to be effective, sufficiently rich and accurate product information as well as personal profiles should be available. Besides, the product attributes have to be carefully chosen for the product and profile. Bad choices of features result in recommender systems with either low discriminating power (the shallow-analysis problem) or bias in reflecting the customer interest (the over-specialization problem) [8]. Collaborative recommender systems are based on the similarity between customer preference ratings for computing recommendations.

Manuscript received February 27, 2001; revised September 27, 2002 and where  $r_{i,y}$  denotes the recorded ratings of customer i for product  $y, \bar{r}_i$ July 10, 2003. This work was jointly supported by Hong Kong Baptist University via Faculty Research under Grant FRG/99-00/II-36P and by RGC Grant denotes the expected rating of customer i over all the products, w(x, i) denotes the expected rating of customer i over all the products, w(x, i) denotes the expected rating of customer i over all the products, w(x, i) denotes the expected rating of customer i over all the products, w(x, i)HKBU/2090/01E. The authors are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: william@comp.hkbu.edu.hk; tsuikc@comp.hkbu.edu.hk; jiming@ comp.hkbu.edu.hk). Digital Object Identifier 10.1109/TSMCA.2003.818877

two problems of the content-based approach and thus has widely been used for recommending products where product descriptions are either lacking or found to be too specific to be useful. Many different techniques have been proposed for collaborative recommendation, including the most original correlation-based methods [9], [10], latent semantic indexing (LSI) [11], [12], Bayesian learning [13], [14], etc. Successful application domains include recommendation of Usenet articles [9], musics [10], etc. In order for collaborative recommendation to be accurate, a large enough number of customers willing to provide preference ratings for the products are required, and the product coverage of their ratings should have significant overlaps. However, this may not be the case in reality because of either lacking such a large customer pool or new products being encountered (the sparsity problem). Applying simple clustering or some statistical cluster models to the preference ratings has been demonstrated to be able to improve the local density of the ratings and is considered to be a promising remedy for the spar-

As the approach does not rely on product contents, it is free from the

sity problem [15], [16]. In this paper, we first describe a statistical cluster model—the latent class model (LCM), originally proposed by Hofmann et al. for collaborative filtering [15], and study how a properly trained LCM can also be used to handle customers and products outside the training set for recommendation. Also, we argue that the LCM is limited in terms of correctly modeling like and dislike ratings and propose a dual latent class model (DLCM) which is trained using two sets of data converted from the original ratings, one with ratings for liked items and another with those for disliked ones. This modification allows the groupings of customers with similar likes and dislikes to be captured separately and thus improve the overall predictive power of the model. Experiments based on the EachMovie dataset were conducted for performance eval-

II. COLLABORATIVE RECOMMENDER SYSTEMS

The concept of collaborative recommendation (also called the word-of-mouth approach) was first used in Goldberg et al.'s e-mail filtering system [17]. The idea was then quickly pursued for product recommendation. In this section, we further elaborate the sparsity problem and briefly survey some existing methods proposed in the literature for alleviating it.

Most of the pioneering collaborative systems use the correlationbased approach for recommendation prediction. For example, in [9],

 $p_{x,y} = \bar{r}_x + \kappa \sum w(x,i)(r_{i,y} - \bar{r}_i)$ 

the predicted rating of customer x for product y is computed as

denotes the Pearson correlation coefficient (P-Corr) between the ratings of customers x and i, given as

 $w(x,i) = \frac{\sum_{j} (r_{x,j} - \bar{r}_x)(r_{i,j} - \bar{r}_i)}{\sqrt{(r_{i,j} - \bar{r}_i)}}$ 

 $\sqrt{\sum_{j}(r_{x,j}-\bar{r}_x)^2\sum_{i}(r_{i,j}-\bar{r}_i)^2}$ 

00,00000000

0,000 0,000

0,0,0,0,0

000000000

in the grid is associated with a parametric real vector called model or

weight that has the same dimension as the input vectors. The task of

SOM is to learn those models so that the similar high-dimensional input

data are mapped into one-dimensional (1-D) or 2-D output space with

the topology as unchanged as possible. That is, the similar data in the

input space under a measurement are placed physically close to each

other on the map. This topology-reserved feature of SOM leads it to

the wide applications to data visualization [12], [6], data clustering [3],

In general, a conventional adaptive SOM needs to initialize a

learning rate and gradually reduces its value over time to ensure the

convergence of the map. Usually, a small initial value of learning rate

is prone to make the models stabilized at some locations of input space

in an early training stage. As a result, the map is not well established

Hence, by rule of a thumb, the learning rate is often initialized at a

relatively large value, and then gradually reduced over time using a

slowly, the map can learn the topology of inputs well with the small

quantization error, but the map convergence needs a large number of

iterations and becomes quite time-consuming. On the other hand, if we

reduce the learning rate too quickly, the map will be likely trapped into

error. To the best of our knowledge, it is a nontrivial task to select an

appropriate learning rate, in particular its associated monotonically

rough topological structure of the training data quickly. In general, the

structure may not be well established. In the second phase, a much

smaller learning rate is utilized to the trained map from the first phase,

which aims at the fine-tuning topological map to ensure the map

convergence. Nevertheless, the performance of the training algorithm

In this letter, we therefore propose a new rival-model penalized

self organizing map (RPSOM) learning algorithm inspired by the

idea of the rival penalized competitive learning (RPCL) [15] and

resulting map in the first phase is prone to error and the topological

monotonically decreasing function. If we reduce the learning rate very

image analysis [10], data mining [13], and so forth.

1083-4427/04\$20.00 © 2004 IEEE

#### Grounding Collaborative Learning in Semantics-Based Critiquing

William K. Cheung<sup>1</sup>, Anders I. Mørch<sup>2</sup>, Kelvin C. Wong<sup>1</sup>, Cynthia Lee<sup>3</sup>, Jiming Liu<sup>1</sup>, and Mason H. Lam<sup>1</sup>

Department of Computer Science, Hong Kong Baptist University, Hong Kong {william, kokwong, jiming, mason}@comp.hkbu.edu.hk <sup>2</sup> InterMedia, University of Oslo, Norway arders.morch@intermedia.uio.no

<sup>3</sup> Language Centre, Hong Kong Baptist University, Hong Kong

#### ABSTRACT

In this paper we investigate the use of Latent Semantic Analysis (LSA), Critiquing Systems, and Knowledge Building to support computer-based teaching of English composition. We have built and tested an English Composition Critiquing System that makes use of LSA to analyze student essays and compute feedback by comparing their essays with teacher's model essays. LSA values are input to a critiquing component to provide a user interface for the students. A software agent can also use the critic feedback to coordinate a collaborative knowledge building session with multiple users (students and teachers). Shared feedback provides seed questions that can trigger discussion and extended reflection about the next phase of writing. We present the first version of a prototype we have built, and report the results from three experiments. We end the paper by describing our plans for future work.

Keywords: Semantic Matching; Web-based Learning; Critiquing Systems; Knowledge Building; Essay Writing.

#### INTRODUCTION

English is the preferred second language for many people and learning it occurs in many ways. For example, young people are quite apt in learning spoken English phrases when watching TV, browsing the Internet and communicating with peers on mobile phones (e.g. SMS). However, previous studies have shown these influences may have negative effect on vocabulary development (Rice et al 1990; Weizman & Snow 2001). As a consequence, students' reading and writing skills do not keep pace with listening, viewing and speaking. Furthermore, English composition is primarily taught in the classroom and practiced in homework assignments, supported by qualified teachers and parents. These are important but scarce resources, creating an im-

#### A Novel Orthogonal NMF-Based Belief Compression for POMDPs

LIXIN@COMP.HKBU.EDU.HK William K. W. Cheung WILLIAM@COMP.HKBU.EDU.HK Jiming Liu JIMING@COMP.HKBU.EDU.HK VINCENT@COMP.HKBU.EDU.HK Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, HK

Abstract

High dimensionality of POMDP's belief state space is one major cause that makes the underlying optimal policy computation intractable. Belief compression refers to the methodology that projects the belief state space to a low-dimensional one to alleviate the problem. In this paper, we propose a novel orthogonal non-negative matrix factorization (O-NMF) for the projection. The proposed O-NMF not only factors the belief state space by minimizing the reconstruction error, but also allows the compressed POMDP formulation to be efficiently computed (due to its orthogonality) in a valuedirected manner so that the value function will take same values for corresponding belief states in the original and compressed state spaces. We have tested the proposed approach using a number of benchmark problems and the empirical results confirms its effectiveness in achieving substantial computational cost saving in policy computation.

#### 1. Introduction

Partially Observable Markov Decision Process (POMDP) models how an agent acts in a stochastic environment given partial observations and feedback from the environment for a better average reward in the long run. Due to the partial observability, it is common to represent the belief state of a POMDP as a probability mass function defined over the true states. Upon each action-taking and then observation arrival, the belief state is re-estimated using Bayesian updating. The complete set of belief states spans Appearing in Proceedings of the 24<sup>th</sup> International Conference on Machine Learning, Corvallis, OR, 2007. Copyright 2007 by the author(s)/owner(s).

a |S|-1 dimensional continuous hypercube (also called belief space). Solving a POMDP is equivalent to computing its optimal policy (a mapping between belief states and actions) which is known to be computationally challenging. Even though there exist related computational shortcuts, the complexity bound for obtaining the optimal one with t steps ahead considered has been shown to be  $O(\zeta_{t-1}^{|Z|})$ (Cassandra, 1998) where Z is the set of possible observations and Ci is the space complexity of the value function at the ith iteration. So, the higher the dimension of the belief space, the larger will be the value of  $\zeta_i$  and thus the overall complexity.

Belief compression refers to the methodology which projects the high-dimensional belief space to a lowdimensional one for approximation in order to cut down the policy computation cost. In the literature, two main approaches have been proposed. One explores the belief space's sparsity by analyzing belief samples (N. Roy and G. Gordon & Thrun, 2005), and another derives the POMDP formulation in the compressed space in a valued-directed manner so that the value function will take same values for corresponding belief states in the original and compressed spaces (Poupart & Boutilier, 2003).

In this paper, we propose to combine the strengths of the two approaches via a novel orthogonal nonnegative matrix factorization (O-NMF). The proposed belief compression approach has a number of advantages, including: (1) O-NMF guarantees all the elements of the low-dimensional belief states to be nonnegative, which is important as belief states are by themselves probability distributions; (2) O-NMF ex plores sparsity in belief space; (3) The value-directed property can be maintained as far as possible; (4) The overhead computation needed for getting the compressed POMDP formulation is carefully designed to avoid solving LP problems; and (5) The high-

dimensional  $\alpha$  vectors (characterizing the value fund

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005

#### IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 18, NO. 1, JANUARY 2001

- [15] S. W. Kim and B. J. Oommen, "On utilizing search methods to select subspace dimensions for kernel-based nonlinear subspace classifiers," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 1, pp. 136–141,
- [16] M. J. D. Powell, A tolerant algorithm for linearly constrained optimiza-tions calculations Univ. Cambridge, Cambridge, U.K., DAMTP Rep
- NA17, 1988.
   M. J. D. Powell, TOLMIN: A Fortran package for linearly constrained optimizations calculations Univ. Cambridge, Cambridge, U.K., DAMTP Report NA2, 1989.
   C. L. Blake and C. J. Merz, UCI Repository of Machine Learning Data-
- base (1998) [Online]. Available: http://www.ics.uci.edu/mlearn/ML
- [19] T. Golub, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531-537, 1999.

  gular neighborhood. (b) Hexagonal neighborhood. The number n next to the dashed lines denotes the n-neighborhood of the neuron in grey. The number of neurons in one-neighborhood in a k-polygonal neighborhood is k, i.e., the rect-
- pp. 531-537, 1999.

  [20] K. C. Chou and Y. D. Cai, "Predicting protein localization in budding yeast," Bioinf., vol. 21, pp. 944-950, 2005.

  [21] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea, "Cilobal analysis of protein localization in budding yeast," Nature, vol. 425, pp. 686-691, 2003.

  [22] K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines suing compositions of amino acid and amino acid paris," Bioinf., vol. 19, pp. 1656-1663, 2003.

  [23] F. A. Thabtah, P. Cowling, and Y. Peng, "MMAC: A new multi-class, multi-label associative classification approach." in Proc. 4th IEEE Int.
- munt-tabet associative classification approach," in *Proc. 4th IEEE Int. Conf. Data Mining*, 2004, pp. 217–224.

#### Rival-Model Penalized Self-Organizing Map

Yiu-ming Cheung and Lap-tak Law

Abstract—As a typical data visualization technique, self-organizing map (SOM) has been extensively applied to data clustering, image analysis, dimension reduction, and so forth. In a conventional adaptive SOM, it needs to choose an appropriate learning rate whose value is monotonically reduced over time to ensure the convergence of the map, meanwhile being kept large enough so that the map is able to gradually learn the data topology. Otherwise, the SOM's performance may seriously deteriorate. In general, it is nontrivial to choose an appropriate monotonically decreasing unction for such a learning rate. In this letter, we therefore propose a novel rival-model penalized self-organizing map (RPSOM) learning algorithm that, for each input, adaptively chooses several rivals of the best-matching unit (BMU) and penalizes their associated models, i.e., those parametric real vectors with the same dimension as the input vectors, a little far away from the input. Compared to the existing methods, this RPSOM utilizes a constant learning rate to circumvent the awkward selection of a a local suboptimal solution and finally led to the large quantization monotonically decreased function for the learning rate, but still reaches a robust result. The numerical experiments have shown the efficacy of our

decreasing function. In the literature, a two-phase training of the SOM  $\label{local_equation} \emph{Index Terms} — \textbf{Constant learning rate, rival-model penalized self-organizing map (RPSOM), self-organizing map (SOM).}$ has been further proposed by [14], which utilizes two learning rates to solve the previous selection problem in training the SOM. In the first phase, it keeps a large learning rate that aims at capturing the

I. INTRODUCTION

Self-organizing map (SOM) [9] and its variants, e.g., see [5], [7], [8], and [17], are one of the popular data visualization techniques that

Manuscript received September 5, 2004; revised June 20, 2006. This work was supported by the Research Grant Council of the Hong Kong SAR, China, under Project HKBU 2156/04E and Project HKBU 210306 and by the Faculty is still sensitive to the time-varied learning rate. Research Grant of Hong Kong Baptist University, Hong Kong, under Project The authors are with the Department of Computer Science, Hong Kong Bapits recently improved variant, named rival penalization controlled tist University, Hong Kong, SAR, China (e-mail: ymc@comp.hkbu.edu.hk).
Color versions of one or more of the figures in this paper are available online at http://iecexplore.ieee.org.
Digital Object Identifier 10.1109/TNN.2006.885039

competitive learning (RPCCL) approach [1], [2]. For each input, the RPSOM adaptively chooses several rivals of the best-matching unit (BMU) and penalizes their associated models a little far away from

1045-9227/\$20.00 @ 2006 IEEE

### ELSEVIER

Available at www.ComputerScienceWeb.com POWERED BY SCIENCE (D) DIRECTS Pattern Recognition Letters 24 (2003) 2883-2893

Pattern Recognition Letters www.elsevier.com/locate/patre

 $k^*$ -Means: A new generalized k-means clustering algorithm

Yiu-Ming Cheung \* Department of Computer Science, Hong Kong Baptist University, 7IF Sir Run Run Shaw Building, Kowloon Tong, Hong Kong

Received 23 July 2002; received in revised form 11 April 2003

This paper presents a generalized version of the conventional k-means clustering algorithm [Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, I. University of California Press, Berkeley, 1967, p. 281]. Not only is this new one applicable to ellipse-shaped data clusters without dead-unit problem, but also performs correct clustering without pre-assigning the exact cluster number. We qualitatively analyze its underlying mechanism, and show its outstanding performance through the experiments. © 2003 Elsevier B.V. All rights reserved.

Keywords: Clustering analysis; k-Means algorithm; Cluster number; Rival penalization

#### 1. Introduction

Clustering analysis is a fundamental but important tool in statistical data analysis. In the past, the clustering techniques have been widely applied in a variety of scientific areas such as pattern recognition, information retrieval, microbiology analysis, and so forth.

In the literature, the k-means (MacQueen, 1967) is a typical clustering algorithm, which aims to partition N inputs (also called data points interchangeably)  $x_1, x_2, \dots, x_N$  into  $k^*$  clusters by assigning an input x, into the jth cluster if the indicator function  $I(j|x_i) = 1$  holds with

\*This work was supported by a Faculty Research Grant of Hong Kong Baptist University with the project code: FRG/02-Tel.: +852-3411-5155; fax: +852-3411-7892. E-mail address: ymc@comp.hkbu.edu.hk (Y.-M. Cheung).

 $I(j|x_t) = \begin{cases} 1 & \text{if } j = \arg\min_{1 \leqslant r \leqslant k} ||x_r - m_r||^2; \\ 0 & \text{otherwise.} \end{cases}$ 

Here,  $m_1, m_2, \ldots, m_k$  are called seed points or units that can be learned in an adaptive way as follows: Step 1. Pre-assign the number k of clusters, and initialize the seed points  $\{m_i\}_{i=1}^k$ . Step 2. Given an input  $x_i$ , calculate  $I(j|x_i)$  by Eq. Step 3. Only update the winning seed point  $m_{\nu}$ , i.e.,  $I(w|x_t) = 1$ , by  $m_{\omega}^{\text{new}} = m_{\omega}^{\text{old}} + \eta (x_i - m_{\omega}^{\text{old}}),$ 

where  $\eta$  is a small positive learning rate. The above Step 2 and Step 3 are repeatedly implemented for each input until all seed points

0167-8655/\$ - see front matter @ 2003 Elsevier B.V. All rights reserved.

#### Maximum Weighted Likelihood via Rival Penalized EM for Density Mixture Clustering

Yiu-ming Cheung, Member, IEEE

with Automatic Model Selection

Abstract—Expectation-Maximization (EM) algorithm [10] has been extensively used in density mixture clustering problems, but it is

weighted likelihood. Under a specific weight design, we give out a Rival Penalized Expectation-Maximization (RPEM) algorithm, which makes the components in a density mixture compete each other at each time step. Not only are the associated parameters of the winner updated to adapt to an input, but also all rivals' parameters are penalized with the strength proportional to the corresponding posterior density probabilities. Compared to the EM algorithm [10], the RPEM is able to fade out the redundant densities from a density mixture during the learning process. Hence, it can automatically select an appropriate number of densities in density mixture clustering. We experimentally demonstrate its outstanding performance on Gaussian mixtures and color image segmentation problem. Moreover, a simplified version of RPEM generalizes our recently proposed RPCCL algorithm [8] so that it is applicable to elliptical clusters as well with any input proportion. Compared to the existing heuristic RPCL [25] and its variants, this generalized RPCCL (G-RPCCL) circumvents the difficult preselection of the so-called delearning rate. Additionally, a special setting of the G-RPCCL not only degenerates to RPCL and its Type A variant, but also gives a guidance to choose an appropriate delearning rate for them. Subsequently, we propose a stochastic version of RPCL and its Type A variant, respectively, in which the difficult selection problem of delearning rate has been novelly circumvented. The experiments show the promising results of this stochastic implementation.

Index Terms-Maximum weighted likelihood, rival penalized Expectation-Maximization algorithm, generalized rival penalization controlled competitive learning, cluster number, stochastic implementation.

#### 1 INTRODUCTION

A s a statistical tool, clustering analysis has been widely applied to a variety of scientific areas such as vector quantization [7], [15], data mining [18], image processing [14], [22], statistical data analysis [6], [13], and so forth. In the past, the clustering analysis has been extensively studied along two directions: 1) hierarchical clustering and 2) nonhierarchical clustering. In this paper, we concentrate on the latter only, which aims to partition N inputs (also called observations or data points interchangeably measure, but those in different clusters are not. Conventionally, the k-means [16] is a popular clustering algorithm that updates the seed points, i.e., those learnable at least two major drawbacks:

1. It implies that the shapes of data clusters are spherical because it performs clustering based on the Euclidean distance only. 2. It needs to preassign the cluster number k, which is an estimate of  $k^*$ . When k is exactly equal to  $k^*$ , the

• The author is with the Department of Computer Science, Rm. 709, 7/F, Sir Run Run Shaw Building, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, P.R. China. E-mail: ymc@comp.hkbu.edu.hk. Manuscript received 19 Mar. 2004; revised 28 Sept. 2004; accepted 13 Dec. 2004; published online 20 Apr. 2005. For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0075-0304.

k-means algorithm can correctly find out the clustering centers as shown in Fig. 1b. Otherwise, it will lead to an incorrect clustering result as depicted in Figs. 1a and 1c, where some of the seed points are not located at the centers of the corresponding clusters. Instead, they are at some boundary points between different clusters or at points far away from the cluster centers. In the literature, a broad view of the clustering problem has

hereinafter), written as  $x_1, x_2, \dots, x_N$ , into  $k^*$  true clusters so been formulated within the framework of density estimates that the inputs within the same cluster are similar under a [5], [17], [19], [21], in which the probability density of inputs is represented by a finite mixture model. Each mixture component represents the density distribution of a data cluster. Consequently, clustering can be viewed as identifypoints in the input space representing the cluster centers, ing the dense regions of the input densities and therefore via minimizing a mean-square-error function. The k-means named density mixture clustering. In particular, it is called has been widely used in a variety of applications, but it has Gaussian mixture clustering when each mixture component is a Gaussian density. Often, the Expectation-Maximization (EM) algorithm [10], [12] provides a general solution for the parameter estimate in a density mixture model. Unfortunately, it also needs to preassign an appropriate number of densities analogous to the k-means algorithm. Otherwise, the EM will mostly give out a poor estimate result. In the past decades, some works have been done toward

determining the correct number of clusters or densities along two major lines. The first one is to formulate the cluster number selection as the choice of component number in a finite mixture model. Consequently, there have been some criteria proposed for model selection, such as AIC [2], [3], CAIC [4], and SIC [21]. However, these conventional criteria may overestimate or underestimate the cluster number due to the difficulty of choosing an 1041-4347/05/\$20.00 © 2005 IEEE Published by the IEEE Computer Society

#### **Concise Papers**

Likelihood and an Extended EM Algorithm Zhenyue Zhang and

On Weight Design of Maximum Weighted

Yiu-ming Cheung, Senior Member, IEEE

Abstract—The recent Maximum Weighted Likelihood (MWL) [18], [19] has provided a general learning paradigm for density-mixture model selection and aight design, however, is a key issue. This paper will therefor explore such a design, and through which a heuristic extended Expectation-Maximization (X-EM) algorithm is presented accordingly. Unlike the EM algorithm [1], the X-EM algorithm is able to perform model selection by fading the redundant components out from a density mixture, meanwhile estimating the model parameters appropriately. The numerical simulations demonstrate the efficacy of

Index Terms-Maximum weighted likelihood, weight design, extended expectation-maximization algorithm, model selection.

#### 1 INTRODUCTION

As a general solution for the parameter estimate in a density mixture model, the Expectation-Maximization (EM) algorithm [1] and its variants, e.g., see [2], [3], [4], [5], [6] have been extensively Bayesian network [8], hidden Markov model [9], bioinformatics [10], and so forth. Nevertheless, the EM is unable to make a model selection, i.e., to determine the appropriate number of model components in a density mixture. If the number of components is not correctly assigned, the EM will generally lead to a poor estimate of the model parameters. In general, it is a nontrivial task to determine such a number.

In the literature. Bayes forther Markov model [9], bioinformatics independently and identically distributed from a mixture distribution of  $k^*$  probability densities, denoted as  $p(\mathbf{x}|\mathbf{\theta}^*)$ , the Maximum Likelihood (ML) estimate of true model parameters set  $\mathbf{\theta}^*$ , denoted as  $\mathbf{\Theta}$ , can be obtained via maximizing the following cost function  $\ell(\mathbf{\Theta}) = \sum_{i=1}^{N} \ln n(\mathbf{x}_i|\mathbf{\Theta})$ to determine such a number.

In the literature, Bayes factors [11] have provided a general with

framework for model comparisons, which gives a systematic means of selecting not only the parameterization of the model, but also the number of components. From Bayes factors, there are some typical model selection criteria, e.g., AIC [12], [13], BIC [14], CAIC [15], and SIC [14]. Generally, the EM algorithm and these criteria are used separately for the parameter estimation and model selection. Alternatively, another interesting way is to learn the parameter estimation and model selection jointly in a single the parameter estimation and model selection jointly in a single paradigm. One example is the Reversible Jump Markov Chair Monte Carlo (R]MCMC) method proposed by Green [16], which is essentially a random sweep Metropolis-Hastings method. The R]MCMC algorithm has been successfully applied to the Gaussian EM (CEM) [5]. Another example is the Competitive EM (CEM) [5]. Analogous to the SMEM algorithm [4], the CEM utilizes a heuristic split-and-merge mechanism to either split the model components in an underpopulated region or merge the components in an overpopulated region iteratively so that the algorithm is able to avoid local solutions. Furthermore, the CEM also exploits a heuristic component annihilation mechanism to determine the number of model components. The experiments in [5] have shown the promising results of this method. Nevertheless, Θ\* is then:

 Z. Zhang is with the Department of Mathematics, Zhejiang University, 2. Chang is with the Department of Pathermanics, England China. E-mail: yyzhang@zju.edu.cn.
 Y.-m. Cheung is with the Department of Computer Science, Rm. 709, 7/f. Sir Run Run Shaw Building, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, P.R. China.
 E-mail: ymc@comp.hkbu.edu.hk.
 However, if k\* is miss-estimated, the EM algorithm almost always reads to a poor estimate of Θ\*. For example, if k > k\*, the EM will regard a true density as a mixture of two or more densities, and has no mechanism to push those redundant α<sub>j</sub>s toward zero. As a

Manuscript received 21 Mar. 2005; revised 6 Sept. 2005; accepted 16 May result, the estimate of Θ\* is poor.

1041-4347/06/520 00 to 2006 IEEE Published by the IEEE Computer Society

#### IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 10, OCTOBER 2006

the computations of the CEM are rather heavier than the EM. In particular, its split-and-merge mechanism introduces two new meters, which have close relations with the performance of the CEM. To the best of our knowledge, how to determine their values

is still an open problem from the theoretical viewpoint.

Recently, the second author of this paper has proposed a new novel learning framework, namely, Maximum Weighted Likelihood (MWL) [18], [19]. With a specific weight design, it has been shown that the MWL is able to select the models automatically during the parameter learning process. Nevertheless, the MWL leaves two open questions: 1) How to design the weights so that an MWL algorithm is able to perform model selection? 2) What are the general convergence properties of an MWL algorithm, although the convergence of a specific MWL algorithm proposed in [18], [19] has been guaranteed?
In this paper, we will concentrate on the first question only to

explore the weight design, and through which a heuristic extended EM (X-EM) algorithm is presented accordingly. Keeping the number of model components unchanged, this new algorithm updates the parameters analogous to the EM [1], and is able to perform the model selection by fading the redundant components out from a density mixture during the parameter learning process. Experimental results have shown the efficacy of our algorithm.

#### 2 THE FRAMEWORK OF MWL LEARNING

 $p(\mathbf{x}_t|\Theta) = \sum_{i=1}^{k} \alpha_j p(\mathbf{x}_t|\theta_j), \sum_{i=1}^{k} \alpha_j = 1, \text{ and } \alpha_j \ge 0 \text{ for } \forall j,$  (2)

where k is an estimate of the true mixture number k\*, and

 $\Theta = \{\alpha_j, \theta_j\}_{j=1}^k$ . Hereinafter, we suppose that  $p(\mathbf{x}|\Theta)$  is an identifiable density with respect to  $\Theta$ , i.e., for any two possible values of  $\Theta$ , denoted as  $\Theta_1$  and  $\Theta_2$ ,  $p(\mathbf{x}|\Theta_1) = p(\mathbf{x}|\Theta_2)$  if and only if  $\Theta_1 = \Theta_2$ .

 $h(j|\mathbf{x}_{t}, \Theta^{(n)}) = \frac{\alpha_{j}^{(n)} p(\mathbf{x}_{t}|\theta_{j}^{(n)})}{p(\mathbf{x}_{t}|\Theta^{(n)})}, \text{ with } j = 1, 2, ..., k.$  (3) M-Step: Fixing all  $h(j|x_l, \Theta^{(n)})$ s, the new improved estimate of

 $\Theta^{(n+1)} = \arg \max_{\Theta} \sum_{t=1}^{N} \sum_{i=1}^{k} h(j|\mathbf{x}_{t}, \Theta^{(n)}) \ln[\alpha_{j} p(\mathbf{x}_{t}|\theta_{j})].$  (4)

2006; published online 18 Aug. 2006.

For information on obtaining reprints of this article, please send e-mail to: Ikde@computer.org, and reference IEEECS Log Number TKDE-0101-0305.

## Agents, Machine-learning, and Data-mi

The Development of Successful On-Line Communities

Karen S.K. Cheung, Fion S.L. Lee, Rachael K.F. Ip, and Christian Wagner

> Department of Information Systems City University of Hong Kong

Abstract

communities.

1. Introduction

communities play an increasingly important role in economic, been witnessing the emergence of information, and emotional exchanges. geographically unbound communities, with Furthermore, some of them have become so vast economic power and impact. For large that their social and economic impact example, by the end of 2001 AOL reached a can be considerable. At the same time, not worldwide membership of over 34 Million all virtual communities seem to flourish, and subscribers (8 million outside the US, with some perform better than others. This study one million in France and 750,000 in Latin seeks to examine the main design criteria America), seven million more than the year underlying successful communities, and to before. With 34 Million subscribers, the identify principles for the design of AOL "economy" has a larger number of successful communities. The article is based "citizens" than Canada, Morocco, or Peru, on a review of the literature, and the study of and as a country would rank 34th in several communities. As such, it is population size (http://blue.census.gov/cgiexploratory in nature. All of the principles bin/ipc/idbrank.pl). will require further in-depth research to test its validity and generalizability. Among the studies of the interesting observations are behavioral changes, such as switching from the (almost) non-sustainability of hub-type the Internet Explorer Browser to Netscape, it (1:N) communities, and the attempt of many has a huge impact on the popularity of such communities to reposition themselves to take technologies. If they make economic

Over the last several years, we have

If these 34 million citizens make advantage of community interaction and the decisions, such as purchasing decisions, they social capital that exists in tightly integrated can have a large financial impact. If 34 million subscribers decide to support a particular cause, they present a formidable

International Journal of The Computer, the Internet and Management Vol. 13#1 (January – April, 2005) pp 71 - 89

Fion S. L. Lee Department of Information Systems City University of Hong Kong Tel: (852) 2784-7538 Email: isfion@is.cityu.edu.hk

Virtual Community Informatics: What We Know and What We Need to Know

Douglas Vogel Department of Information Systems City University of Hong Kong Tel: (852) 2784-7560 Email: isdoug@is.cityu.edu.hk

Moez Limayem Department of Information Systems City University of Hong Kong Tel: (852) 2784-8530 Email: ismoez@is.cityu.edu.hk

Abstract

The virtual community has just recently emerged with divergent opinions on the basic understanding of it. This study aims at collecting different definitions and classifications in the virtual community, and offers a working definition. It also addresses research conducted in the field by referring to Information Systems journals. The With the exponential growth of the virtual community, more and more studies have been conducted on how virtual communities affect living standards by providing functions for relationship building and knowledge sharing (Siwolop 1997, Brown 2000, Bieber et. al. 2001, Blase 2000, lack of research available on the topic.

with others. The current practice is to build web sites and allow people to register as members who can then share information or feelings virtually. However, it is doubtful whether the tools that support virtual community web sites assist in relationship building and knowledge sharing.

This paper aims at comparing the different definitions and classifications of the virtual community to achieve a more compromising agreement on these basic concepts of the virtual community. Existing research on the virtual community is identified and future research topics are proposed. A survey is conducted on internet tools used in virtual community web sites, and suggestions are provided for how these tools can provide support. It is intended that these guidelines will support the virtual community

research categorizes the different stages in virtual community growth to show the transition of research in this area. A survey is also conducted on the extent of the adoption of informatics in virtual community web sites.

Introduction

Baranski 1997). Nevertheless, among these studies, little consensus has been reached on basic concepts such as definitions and classifications of the virtual community. Without such underlying concepts, researchers show various meanings for the same terms used. There is also a The virtual community provides access for engaging in common activities, sharing feelings, or discussing ideas Definition of a Virtual Community

A generally agreed upon definition of a virtual community would be a good starting point. What we need is a working definition of the virtual community, a consensus found in the major stream of literature, a definition that understood by most of people. To achieve this goal, definitions of the virtual community proposed by various authors are compared in Table I. Similar items found in definitions are then extracted in order to build up a working definition.

ward (1993) Social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal

JOURNAL OF INFORMATION TECHNOLOGY THEORY AND APPLICATION

**VIRTUAL COMMUNITY INFORMATICS:** A REVIEW AND RESEARCH AGENDA

FION S. L. LEE, City University of Hong Kong Department of Information Systems, Kowloon Tong, Tel: (852) 2788-7538, Email: isfion@is.citywedu.hk

DOUGLAS VOGEL, City University of Hong Kong Department of Information Systems, Kowloon Tong, Tel: (852) 2788-7560, Email: isdoug@is.cityu.edu.hk

MOEZ LIMAYEM, City University of Hong Kong Department of Information Systems, Kowloon, Tel: (852) 2788-8530, Email: ismoez@is.cityu.edu.hk

ABSTRACT

Divergent opinions exist on the basic understanding of the concept, virtual community. This study offers a working definition by examining different definitions, and proposes adoption of virtual community classifications. It also includes a summary of research conducted in the field. The research categorizes the different stages in virtual community growth to show the transition of research in this area. The results illustrate a paucity of technology development studies. We also investigate the extent of the adoption of informatics in these communities using a survey 200 virtual communities. The results indicate that discussion forum is the most popular tool adopted in virtual communities. The integration of the research review and tool adoption survey contributes to the generation of an agenda to direct future virtual community research.

Ken Peffers acted as senior editor for this paper.

Lee, F.S.L., D. Vogel, and M. Limayem, "Virtual Community Informatics: A Review and Research Agenda", The Journal of Information Technology Theory and Application (JITTA), 5:1, 2003, 47-61.

Available online at www.sciencedirect.com

ScienceDirect

**OURNAL** OF COMPUTER ND SYSTEM SCIENCES

www.elsevier.com/locate/jcss

Journal of Computer and System Sciences ••• (••••) •••-•••

Topological analysis of AOCD-based agent networks and experimental results

Hao Lan Zhang\*, Clement H.C. Leung, Gitesh K. Raikundalia

School of Computer Science and Mathematics, Victoria University, PO Box 14428, Melbourne City, MC 8001, Australia Received 15 June 2006; received in revised form 31 October 2006

Abstract

Topological analysis of intelligent agent networks provides crucial information about the structure of agent distribution over a network. Performance analysis of agent network topologies helps multi-agent system developers to understand the impact of topology on system efficiency and effectiveness. Appropriate topology analysis enables the adoption of suitable frameworks for specific multi-agent systems. In this paper, we systematically classify agent network topologies and propose a novel hybrid topology for distributed multi-agent systems. We compare the performance of this topology with two other common agent network topologies—centralised and decentralised topologies—within a new multi-agent framework, called Agent-based Open Connectivity for DSS (AOCD). Three major aspects are studied for estimating topology performance, which include (i) transmission time for a set of requests; (ii) waiting time for processing requests; and (iii) memory consumption for storing agent information. We also conduct a set of AOCD topological experiments to compare the performance of hybrid and centralised agent network topologies and illustrate our experimental results in this paper. © 2007 Elsevier Inc. All rights reserved.

Keywords: Agent network topology; Agent network performance; Intelligent agents; AOCD

1. Introduction

Applications of multi-agent systems have been arising in many areas. Agent-based systems provide a way of conceptualising complex software applications. These applications often face problems involving multiple and distributed sources of knowledge [1]. However, current research work with respect to topological theory in the area of intelligent agents is inadequate. This situation has led to a set of important research problems concerning how an agent network should be designed to perform efficiently and effectively. The agent network topology issue becomes one of the key issues in solving problems of agent cooperation and communication. An appropriate agent network topology design helps to enhance agent communication efficiency and network mobility.

An agent network topology represents the information of agent distribution over an agent network, which incorpo-

rates agent mobility and intelligence aspects into the process of arranging and configuring an agent network. The term,

\* Corresponding author.

E-mail addresses: haolan@sci.vu.edu.au (H.L. Zhang), clement.leung@vu.edu.au (C.H.C. Leung), gitesh.raikundalia@vu.edu.au (G.K. Raikundalia).

0022-0000/\$ - see front matter © 2007 Elsevier Inc. All rights reserved doi:10.1016/j.jcss.2007.04.006

Please cite this article in press as: H.L. Zhang et al., Topological analysis of AOCD-based agent networks and experimental results, J. Comput. System Sci. (2007), doi:10.1016/j.jcss.2007.04.006

Research Roadmap Information Information High Perform. HTML DTD heterogeneity overflow computing manageable ... interoperable ... on-demand ... Agent Tech. / Grid Data Mining Semantic Web Computing Web Intelligence MIDDLEWARE DAML-S WSDL SOAP/XML HTTP TCP/IP

#### Feature Selection with Transductive Support Vector Machines

Zhili Wu<sup>1</sup> and Chunhung Li<sup>2</sup>

Department of Computer Science, Hong Kong Baptist University vincent@comp.hkbu.edu.hk

Department of Computer Science, Hong Kong Baptist University chli@comp.hkbu.edu.hk

Summary. SVM-related feature selection has shown to be effective, while feature selection with transductive SVMs has been less studied. This paper investigates the use of transductive SVMs for feature selection, based on three SVM-related feature selection methods: filtering scores | SVM wrapper, recursive feature elimination(RFE) and multiplicative updates(MU). We show transductive SVMs can be tailored to feature selection by embracing feature scores for feature filtering, or acting as wrappers and embedded feature selectors. We conduct experiments on the feature selection competition tasks to demonstrate the performance of Transductive SVMs in feature selection and classification.

#### 1 Introduction

SVMs have been studied to work with different divisions of feature selection methods, like wrappers, embedded methods, and filters. For example, the SVM as a wrapper can be used to select features, while the feature set quality is indicated by the performance of the trained SVM (Yu and Cho, 2003). Feature selection can also be conducted during the process of training SVMs. This induces some embedded feature selection methods (Brank et al., 2002, Guyon et al., 2002, Weston et al., 2003, 2000, Guyon et al., 2003, Perkins et al., 2003). SVMs are also useful for filter methods. Although filters often score or rank features without utilizing a learning machine, they can utilize a SVM as the final predictor for classification. Moreover, filters are often integrated as a preprocessing step into wrappers and embedded methods, such that they can help feature selection of wrappers or embedded methods.

The development of SVM-related feature selection methods motivates us to explore the role of Transductive SVMs (TSVMs) in feature selection. Transductive SVMs are a type of transductive learning (Vapnik, 1998) machine. They aim to build SVM models upon both labeled and unlabeled data. Transductive SVMs, as generalized from inductive SVMs, inherit the advantages of inductive SVMs such as the large margin (Boser et al., 1992) and regularization formulation as well as kernel mapping. Besides, TSVMs are suitable for the tasks we consider here, which are characterized

Agent-Based Load Balancing on

Homogeneous Minigrids: Macroscopic

Modeling and Characterization

Jiming Liu, Senior Member, IEEE, Xiaolong Jin, and Yuanshi Wang

Abstract—In this paper, we present a macroscopic characterization of agent-based load balancing in homogeneous minigrid

environments. The agent-based load balancing is regarded as agent distribution from a macroscopic point of view. We study two

quantities on minigrids; the number and size of teams where agents (tasks) queue. In macroscopic modeling, the load balancing mechanism is characterized using differential equations. We show that the load balancing we concern always converges to a steady

state. Furthermore, we show that load balancing with different initial distributions converges to the same steady state gradually. Also,

we prove that the steady state becomes an even distribution if and only if agents have complete knowledge about agent teams on

minigrids. Utility gains and efficiency are introduced to measure the quality of load balancing. Through numerical simulations, we

discuss the utility gains and efficiency of load balancing in different cases and give a series of analysis. In order to maximize the utility

gain and the efficiency, we theoretically study the optimization of agents' strategies. Finally, in order to validate our proposed agent-

based load balancing mechanism, we develop a computing platform, called Simulation System for Grid Task Distribution (SSGTD).

Through experimentation, we note that our experimental results in general confirm our theoretical proofs and numerical simulation

results from the proposed equation system. In addition, we find a very interesting phenomenon, that is, agent-based load balancing

Index Terms.—Homogeneous minigrids, load balancing, task distribution, agents, macroscopic modeling, steady states, convergence,

Applied Intelligence 22, 37-46, 2005 © 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands

Guided Cluster Discovery with Markov Model\*

C.H. LI Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong chli@comp.hkbu.edu.hk

Abstract. Cluster discovery is an essential part of many data mining applications. While cluster discovery process is mainly unsupervised in nature, it can often be aided by a small amount of labeled data. A probabilistic model on the clustering structure is adopted and a novel unified energy equation for clustering that incorporates both labeled data and unlabeled data is introduced. This formulation is inspired by a force-field model integrating labeling constraint on labeled data and similarity information on unlabeled data for joint estimation. Experimental results show that good clusters can be identified using small amount of labeled data.

Keywords: clustering semi-supervised learning, Markov model

1. Introduction

In machine learning for classification problems, there are two distinct approaches to learning or classifying data: the supervised learning and un-supervised learning. The supervised learning deals with problem where a set of data are labeled for training and another set of data would be used for testing. The un-supervised learning deals with problem where none of the labels of the data are available. Unsupervised clustering can be broadly classified into whether the clustering algorithm is hierarchical or non-hierarchical. Hierarchical methods often model the data to be clustered in the form of a tree, or a dendrogram [1]. The lowest level of the tree is usually each datum as a cluster. A dissimilarity measure is defined for merging clusters at a lower level to form a new cluster at a higher level in the tree. The hierarchical methods are often computationally intensive for large number of samples and is difficult to analyze if there is no logical hierarchical structure in the data. Non-hierarchical methods divide the samples into a fixed number of groups using some measure of opti-

\*This work is partially support by the Faculty research grant of the

mality. The most widely used measure is the minimiza-

tion of the sum of squared distances from each sam-

ple to its cluster center. The k-means algorithm, also known as Forgy's method [2] or MacQueen [3] algorithm is a classical algorithm for non-hierarchical unsupervised clustering. However, the k-means algorithm tends to cluster data into even populations and rare abnormal samples in medical problems cannot be properly extracted as individual clusters. Recent progress in clustering includes the modeling of proximity structure [4], the dynamic programming approach to hierarchical clustering using graphs [5] and spectral method to clustering [6]. However, these methods do not make use of prior knowledge on dataset such as possible labels or possible structures within the dataset. In recent years, important data mining tasks have

emerged with enormous volume of data. The labeling of a significant portions of the data for training is either infeasible or impossible. Sufficient labeled data for training are often unavailable in data mining, text categorization and web page classification. A number of approaches have been proposed to combine a set of labeled data with unlabeled data for improving the classification rate. The co-training approach has been proposed to solve the problem of web page classification where the web pages can be represented by two independent representations [7]. The drawback of this co-training approach is that not all data have two independent representations and the algorithm is thus

IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 8, NO. 3, JUNE 2006

#### A Novel Fractal Image Watermarking

Ming Hong Pi, Chun Hung Li, Member, IEEE, and Hua Li, Member, IEEE

Abstract—A novel watermarking method is proposed to hide a Discrete Fourier Transform (DFT), and Discrete Wavelet Transbinary watermark into image files compressed by fractal block form (DWT). Cox et al. [4] asserted that a watermark should be coding. This watermarking method utilizes a special type of transform is determined by the range block mean and contrast orthogonalization fractal coding method where the fractal affine scaling. Such orthogonalization fractal decoding is a mean-in- was added into the 1000 largest coefficients of the DCT (except riant iteration. In contrast, the fractal parameters of classical DC value). Lin et al. [5] proposed a watermarking algorithm ro fractal compression are very sensitive to any change of domain bust to rotation, scaling and translation, where the watermark is block pool and to common signal and geometric distortion. Hence, embedded into a one-dimensional signal, a cumulative function it is impossible to directly place a watermark in fractal parameters. The proposed watermark embedding procedure inserts a permutated pseudo-random binary sequence into the quantized range block means. The watermark is detected by computing and multiresolution detection can be allowed because of wavelet the correlation coefficient between the original and the extracted watermark. Experimental results show that the proposed fractal watermarking scheme is robust against common signal and geometric distortion such as JPEG compression, low-pass filtering, rescaling, and clipping.

Index Terms-Detector response, fractal block coding, orthogonalization fractal transform, watermarking.

I. INTRODUCTION

TTH THE extensive distribution of multimedia data such as text, image, video, and audio, there is a strong demand for ownership management and copyright protection. Digital watermarking is a process of hiding a watermark in multimedia object without perceptual degradation so that the watermark can signature into the fractal codes with a redundancy U, based on be detected or extracted later for copyright ownership identification. Early watermarking techniques are directly implemented in the pixel domain. For instance, Schyndel et al. [1], [2] proposed inserting a watermark into the least significant bits (LSB) of an image by bit-plane manipulation of LSBs or adding the watermark and the image. The watermark is detected by computing the correlation coefficient between the original m-sequence and the watermarked image. Matsui and Tanaka [3] applied linear predictive coding into watermarking, and hided a watermark to make the watermark resemble quantization noise. However, in general, a digital watermark embedded in the LSBs is highly sensitive to noise and susceptible to be destroyed. Many watermarking techniques embed a watermark in the transform domain, such as Discrete Cosine Transform (DCT),

Manuscript received September 5, 2004; revised July 6, 2005. This work was supported in part by FRG0203/II-60 and NSERC.. The associate editor coordi nating the review of this manuscript and approving it for publication was Prof. M. H. Pi is with the Department of Computing Science, University of Alberta, Edmonton, AB T6G 2V4 Canada, (e-mail: minghong@cs.ualberta.ca). C. H. Li is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: chli@comp.hkbu.edu.hk).

H. Li is with the Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, AB T1K 3M4 Canada, (e-mail: huali@cs.uleth. Digital Object Identifier 10.1109/TMM.2006.870738

1520-9210/\$20.00 © 2006 IEEE

placed in the perceptually significant components to deter inof the Fourier magnitudes along the log-radius axis. Zhu et al. [6] inserted a watermark into all high-pass wavelet coefficients, pyramid structure. A spectrum of watermarking techniques has been surveyed in [7]. Over the past decade, fractal block coding has mainly been exploited for the purpose of image compression [13]-[17]. Re-

cently, fractal block coding is investigated for watermarking. Based on the fractal codes, several watermarking techniques have been proposed [8]-[11]. Most existing watermarking techniques using fractal codes

are developed based on the classification of the fractal codes and a watermark is hidden in a small part of fractal codes [8]-[11]. Davern and Scott [8] divided the domain block pool into two halves, the watermark is hidden in a set of selected range blocks according to which half the best-pair domain block belongs to. Similarly, Puate and Jordan [9] proposed hiding a 32-bits binary the principle that the original and decoded images possess the same fractal codes. Each range block is encoded by searching the local search region (LSR), which is defined as a square region around the range block. LSR is divided into two subregions: A and B, and all range blocks are classified into two categories according to whether the best-pair domain belongs to A or  ${f B}$ . Each bit is hidden with U randomly selected range blocks so that the embedded bit can survive even though the fractal codes of some range blocks are altered by the attacks. The analogous fractal watermarking technique was proposed by Li and Wang [10] and Wu and Chang [22], but each bit is hidden by the isometric transforms, instead of the geometric position of the best-pair domain block [8]. The eight isometric transforms are divided into two subgroups:  $I_0$  and  $I_1$ . If a bit is zero, the corresponding range block is encoded using  $I_0$ , otherwise, the corresponding range block is encoded using  $I_1$ . In addition, Bas et al. [11] proposed a binary watermarking technique based on classification of the fractal codes (or the maps). A part of original maps are replaced with modified ones. As a result, a binary watermark is placed into the image.

Because the contrast scaling and luminance offset depend on the domain block pool, the contrast scaling and luminance offset obtained from the attacked watermarked image are likely different from those obtained from the original image as illustrated in Fig. 1. As a result, the watermark is susceptible to tampering.

for users all over the world [1], [2], [3], [5].

1 BACKGROUND

the Internet and provide powerful computation capability

mechanism is topology-independent.

Since a grid connects numerous geographically distributed computers, and tasks are submitted to grid nodes in a changing grid environment. distributed fashion, an important issue is how to evenly distribute submitted tasks to nodes. This is a load balancing problem, one of the scheduling problems on the grid. By solving this problem, we can optimally utilize computatask can be divided into arbitrary chunks. Their scheduling tional resources of the grid. In this paper, we will propose mechanism assumed a master/worker architecture, i.e., a an agent-based load balancing mechanism

1.1 Scheduling on Grids The scheduling problem on grids has been widely studied [6], to task transfer time and network latency. Their algorithm,

[7], [8]. Many schedulers for grid computing have been called uniform multiround (UMR), allocates chunks using developed [9], such as AppLeS [10], [11], Nimrod-G [12], multiple rounds so as to overlap communication (i.e., J. Liu and X. Jin are with the Department of Computer Science, Hong Kong
i.e., makespan.

Zhongshan University, Guangzhou, China. E-mail: mcswys@zsu.edu.cn. Manuscript received 25 Feb. 2003; revised 14 June 2004; accepted 30 Sept. 2004; published online 20 May 2005. 2004; published online 20 May 2005.
For information on obtaining reprints of this article, please send e-mail to: type the computer org, and reference IEEECS Log Number TPDS-0018-0203.

will provide an agent-based, decentralized task allocation mechanism. Specifically, the paper will focus on load

Th order to meet the increasing demand of large-scale GrADs [13], [14], and Condor-G [15]. The scheduling issues  $oldsymbol{\mathsf{L}}$  scientific computation in the fields of life sciences, on grids lie in several aspects, among which resource allocation biology, physics, and astronomy, the notion of "computa-[6], i.e., how to allocate computational resources (e.g., CPUtional grid" was proposed in mid 1990s [1], [2], [3], [4]. It hours, storage, and network bandwidth) to submitted tasks, has been observed that computers (such as PCs, work- and task allocation [7], [8], i.e., how to allocate tasks to different stations, and clusters) in the Internet are often idle. Grid
nodes, attract most attention. In [6], Calstyan et al. proposed computing aims to integrate idle computational power over an agent-based resource allocation model for grid computing. Their model is based on the reinforcement learning technique, and consequently can be adaptive to a dynamically

IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 16, NO. 7, JULY 2005

In [7], [8], [16], [17], Casanova et al. have studied the task allocation problem in a grid environment, where the submitted task is arbitrarily divisible. In other words, a master, acting as a scheduler, is responsible for dividing the submitted tasks and allocating the obtained task chunks to different workers. In their work, they paid special attention transfer time and network latency) and computation and, consequently, decrease the total time for handling the task,

Baptist University, Kowloon Tong, Kowloon, Hong Kong.

E-mail: fjiming.jxl@comp.hkbu.edu.hk.

Y. Wang is with the School of Mathematics and Computational Science, makespan.

In Casanova et al.'s work, the main problem lies in the master/worker architecture. Because a grid environment usually involves a large number of nodes and tasks, such 1045-9219/05/520:00 © 2005 IEEE Published by the IEEE Computer Society

**(1)** 

## Agents, Machine-learning, and Data-mining

#### TOWARD NATURE-INSPIRED -Computing

NIC-based systems utilize autonomous entities that self-organize to achieve the goals of systems modeling and problem solving.

> By JIMING LIU and K.C. TSUI >>

examine NIC from two perspectives. First, as a way to help explain, model, complex real-world systems by formulating computing models and testing hypotheses through con-trolled experimentation. The end product is a potentially deep understanding or at least a better explanation of the working mechanism(s) of the autonomous (such as lifelike) behavior in solving computing problems. With detailed knowledge of the underlying mechanism(s), simplified abstracted | Figure 1) [5, 8]. autonomous lifelike behavior can be used as a model in practically any general-purpose problem-solving

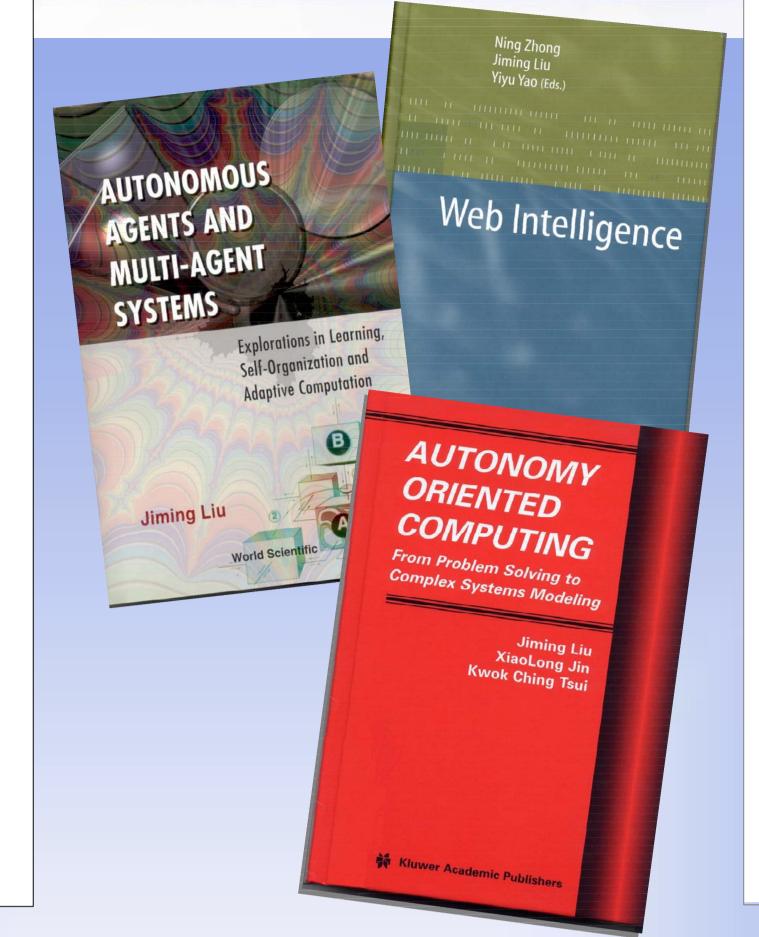
strategy or technique. Neither objective is achievable without formulating

ature-inspired computing (NIC) is an 1 modeling process can begin with a theoretical analysis emerging computing paradigm that | from either a macroscopic or microscopic view of the draws on the principles of self-organi- system. Alternatively, the application developer may zation and complex systems. Here, we adopt a blackbox or whitebox approach. Blackbox approaches (such as Markov models and artificial neural networks) normally do not reveal much about and characterize the underlying mechanism(s) of their working mechanism(s). On the other hand, whitebox approaches (such as agents with bounded rationality) are more useful for explaining behavior [7].

The essence of NIC formulation involves conceiving a computing system operated by population(s) of autonomous entities. The rest of the system is referred modeled system. And second, as a way to reproduce to as the environment. An autonomous entity consists of a detector (or set of detectors), an effector (or set of effectors), and a repository of local behavior rules (see A detector receives information related to its neigh

bots and to the environment. For example, in a simulation of a flock of birds, this information would include the speed and direction the birds are heading a model of the factors underlying the system. The | and the distance between the birds in question. The

COPHUNICATIONS OF THE ACM Dozobar 2006/Vol. 49, No. 10 55



now, there may have existed over one billion Web sites on model. Section 5 shows several experimental results on the Web at anytime, if projected based on the studies characterizing Web usage regularities. Section 6 discusses the reported in [1], [2]. Viewing the Web as a large directed effects on the emergent regularities under different condimodel to simulate certain regularities in user navigation behavior and suggested that the probability distribution of surfing depth (step) follows a two-parameter inverse Gaussian distribution. They conjectured that the probability

1.1 Organization of the Paper regularities on the Web. This is followed by Section 3 which personalize Web surfing operations.

states the problems as well as important issues to be dealt University, Kowloon Tong, Hong Kong.
E-mail: jiming@comp.hkw.edu.hk.
• S. Zhang and J. Yang are with the Department of Precision Machinery and Instrumentation, University of Science and Technology of China, 230026, Jinzai Rd. 96, Hefzi, Anhui, China.
E-mail: zhsw@mail.ustc.edu.cn, Jieyang@ustc.edu.cn.

Manuscript received 18 Dec. 2001; revised 21 Oct. 2002; accepted 17 Mar.

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 5, MAY 2004

#### Characterizing Web Usage Regularities with Information Foraging Agents

Jiming Liu, Senior Member, IEEE, Shiwu Zhang, and Jie Yang

Abstract-Researchers have recently discovered several interesting, self-organized regularities from the World Wide Web, ranging from the structure and growth of the Web to the access patterns in Web surfing. What remains to be a great challenge in Web log mining is how to explain user behavior underlying observed. Web usage regularities. In this paper, we will address the issue of how to characterize the strong regularities in Web surfing in terms of user navigation strategies, and present an information foraging agentbased approach to describing user behavior. By experimenting with the agent-based decision models of Web surfing, we aim to explain how some Web design factors as well as user cognitive factors may affect the overall behavioral patterns in Web usage

Index Terms—Web log, Web mining, power law, regularities, user behavior, decision models, information foraging, autonomous agents, agent-based simulation.

of finding a group surfing at a given level scales inversely in proportion to its depth, i.e.,  $P(L) \sim L^{-3/2}$ .

In order to further characterize user navigation regularities as well as to understand the effects of user interests, in this paper we will present an information foraging agent- and using the information on the Web. based model that takes into account the interest profiles, motivation aggregation, and content selection strategies of users and, thereafter, predicts the emerged regularities in user navigation behavior.

The remainder of this paper is organized as follows: In Web mining with a special focus on studies that deal with the

For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 115581.

THE contents and services on the World Wide Web (or the with in our present study. Section 4 presents the detailed I Web) have been growing at a very rapid rate. Until formulation of our proposed information foraging agent graph of nodes (i.e., Web pages) connected with links (i.e., tions in our model. Finally, Section 7 concludes the paper by hyperlinks), Huberman et al. [3] proposed a random-walk summarizing the key contributions and findings of this

This section provides an overview of research work related to Web mining. Generally speaking, Web mining is aimed to study the issues of 1) where and how information can be efficiently found on the Web and 2) how and why users motivation, and content organization on the user behavior, behave in various situations when dynamically accessing 2.1 Web Mining for Pattern-Oriented Adaptation

The first major task in Web mining may be called Web mining for pattern-oriented adaptation; that is, to identify the interrelationships among different Web sites, either based on the analysis of the contents in Web pages or based on the discovery of the access patterns from Web log files. Section 2, we will provide a survey of the existing work in

By understanding such interrelationships, we aim to develop adaptive Web search tools that help facilitate or

This task is certainly justified as studies have shown that 85 percent of users use search engines to locate information 4]. Even though good search engines normally index only J. Liu is with the Department of Computer Science, Hong Kong Beptist about 16 percent of the entire Web [2], an adaptive utility can still be useful to filter or rank thousands of Web pages that are often returned by search engines. For instance, some researchers have developed efficient search techniques that detect authorities, i.e., pages that offer the best resource of the information on a certain topic and hubs, i.e., pages that are collections of links to authorities [5], [6]. When it is difficult to directly find relevant information from search engines, navigating from one page to another

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, VOL. 35, NO. 4, AUGUST 2005

#### Equilibria, Prudent Compromises, and the "Waiting" Game Kwang Mong Sim

Abstract—While evaluation of many e-negotiation agents are as new opportunities/threats are constantly being introduced. carried out through empirical studies, this work supplements and complements existing literature by analyzing the problem of designing market-driven agents (MDAs) in terms of equilibrium points and stable strategies. MDAs are negotiation agents designed grids [22] must take into consideration the dynamics of a to make prudent compromises taking into account factors such as time preference, outside option, and rivalry. This work shows that 1) in a given market situation, an MDA negotiates optimally heaves it makes minimally account factors such as grid-computing environment because it is expected that resources and services are constantly being added/removed from a grid [23]. because it makes minimally sufficient concession, and 2) by modeling negotiation of MDAs as a game  $\Gamma$  of incomplete information, it is shown that the strategies adopted by MDAs are stable. In a bilateral negotiation, it is proven that the strategy pair of two MDAs forms a sequential equilibrium for  $\Gamma$ . In a multilateral negotiation, it is shown that the strategy profile of MDAs forms a

Index Terms-Automated negotiation, negotiation agent, sequential equilibrium.

#### I. INTRODUCTION

UTOMATED negotiation [1] among software agents is A becoming increasingly important because automated interactions between agents can occur in many different contexts. and research on engineering e-negotiation agents [2]-[9] has received a great deal of attention in recent years (see [2], [3] for a survey). Whereas a substantial portion of the existing work evaluates e-negotiation by experimentation and stochastic simulation, this work supplements and complements the literature by providing a game-theoretic analysis of market-driven agents (MDAs) [10]-[15]. MDAs are e-negotiation agents that make minimally sufficient concessions [1], taking into account time preference, deadline, outside option, differences in proposals, and market rivalry (see Section II-B). The impetus of this work is analyzing MDAs in terms of sequential equilibrium [16], [17], market equilibrium [17], [18], and stable (and dominant) strategies. While [19]-[21] have proven sequential equilibrium on the strategies of their agents in bilateral negotiations (see Section VI), this work shows that the strategies of MDAs tion II-B). A market-driven agent is an e-negotiation agent with are in sequential equilibrium (see Section V-A) and market the distinguishing feature that determines the amount of conequilibrium (see Section V-B) for bilateral and multilateral cession using the time, opportunity, and competition decision negotiations, respectively. Proving market equilibrium and functions. Section II-C discusses the collection of strategies for considering the influence of market condition are essential in a range of negotiation situations existing applications such as e-commerce because when traders enter/leave an e-market, the conditions for deliberation change A. Negotiation Protocol

Manuscript received August 1, 2004; revised December 1, 2004. This paper was supported by the National Science Council of Taiwan under Project Code NSC 93-2218-E-001-002 and the Institute of Information Science, Academia Sinica. This paper was recommended by Associate Editor Philippe De Wilde.
The author is with The Institute of Information Science, Academia Sinica,
Nankang, Taipei, Taiwan, R.O.C. (e-mail: kmsim@iis.sinica.edu.tw).
Digital Object Identifier 10.1109/TSMCB.2005.845395

that agents negotiating for resources in future computational Whereas empirical results derived from stochastic simula-

tions discussed in the author's previous work [12], [14], [15] verified some of the desirable properties (e.g., high average utility and relatively high success rate of reaching consensus) of (enhanced) MDAs, this work consolidates and further supports previous contributions by showing that i) for a given market situation, an MDA negotiates optimally by making minimally sufficient concession (see Section IV), and ii) in bilateral as well as multilateral negotiations, the strategies of MDAs are stable (see Sections V-A and B); this rests on the fact that there is a dominant strategy (see Section III) for an MDA in both bilateral and multilateral negotiations (see also Section IV). Results from this work show that i) MDAs are designed to optimize their utilities in a given market situation and trading constraints, and ii) since the strategies are stable, MDAs are motivated to behave in a desired manner, even though it is assumed that MDAs do not have complete information about their opponents (see Section II-A). In addition to being able to optimize utility, stability is also an essential criterion for evaluating the design of negotiation mechanisms [18]. Additionally, Section IV discusses how the "sit-and-wait" strategy [19] can be modeled in MDAs.

#### II. NEGOTIATION MECHANISM

This section presents the negotiation protocol (see Section II-A) and the market-driven negotiation strategy (Sec-

To set the stage for specifying the negotiation protocol and the market-driven strategy in Section II-B, some assumptions are given as follows:

> Agents do not have information about the deadline. reserve price, strategy, and time preference of other

1083-4419/\$20.00 @ 2005 IEEE

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, VOL. 36, NO. 6, DECEMBER 2006

#### Continuous-Time Negotiation Mechanism for Software Agents

Bo An, Kwang Mong Sim, Liang Gui Tang, Shuang Qing Li, and Dai Jie Cheng

Abstract—While there are several existing mechanisms and the buyer gets more desirable negotiation outcomes when it one-to-many negotiation, this paper develops a flexible one-to-many negotiation mechanism for software agents. Unlike the existing general one-to-many negotiation mechanism, in which an agent should wait until it has received proposals from all its trading partners before generating counterproposals, in the flexible one-to-many negotiation mechanism, an agent can make a proposal in a flexible way during negotiation, i.e., negotiation is conducted in continuous time. To decide when to make a proposal, the trading partner be the time from an agent's proposing to its receiving a counterproposal from the trading partner. In existing general one-to-many negotiation two strategies based on fixed waiting time and a fixed waiting ratio are proposed. Results from a series of experiments suggest that, guided by the two strategies for deciding when to make a proposal, the flexible negotiation mechanism achieved more favorable trading outcomes as compared with the general one-to-many negotiation mechanism. To determine the amount of concession, tiation agents are guided by four mathematical functions based on factors such as time, trading partners' strategies, negoti-ation situations of other threads, and competition. Experimental results show that agents guided by the four functions react to changing market situations by making prudent and appropriate rates of concession and achieve generally favorable negotiation

Index Terms-Automated negotiation, negotiation agents, oneto-many negotiation.

#### I. INTRODUCTION

A UTOMATED negotiation [19], [21] among software not flexible enough when negotiation agents are of different reaction times. mated interactions between agents [4], [8], [29] can occur in many different contexts (e.g., negotiation for resources [9]). negotiation mechanism, this research focuses on developing In terms of the number of agents participating in negotia- a more flexible mechanism than the general one, where are tions, agent-based automated negotiation can be divided into agent can decide when to make a proposal according to the three cases [6], namely: 1) one-to-one negotiation (bilateral synchronization situations of negotiation, which are determined negotiation); 2) many-to-many negotiation; and 3) one-to-many by the reaction time of each trading partner negotiation. Compared with auction mechanisms [18], one-tomany interactive negotiation is more flexible. For example, many negotiation mechanism. agents can adopt different negotiation strategies with different trading partners (alternatives), and negotiations can be taken under different negotiation environments and protocols [20]. In one-to-many negotiation (take the negotiation between a buyer and several sellers as an example), there are two alternatives: 1) buyer negotiates sequentially with all the sellers and 2) buyer negotiates concurrently with these sellers. Generally,

Manuscript received July 11, 2005; revised December 10, 2005 and January 7, 2006. This paper was recommended by Associate Editor G. Skarmeta. B. An, L. G. Tang, S. Q. Li, and D. J. Cheng are with the College of

systems addressing the crucial and difficult issues of automated negotiates concurrently with all the sellers in competitive situations in which there are information uncertainty and deadlines [16], [17]. In this paper, we assume that an agent negotiates concurrently with its trading partners.

Let a negotiation cycle be the time spent in a round of negotimechanisms and systems (e.g., [1], [7], [20], and [30]), taking the negotiation between a buyer and several sellers as an example, the buyer's negotiation with the set of sellers is divided into several rounds (indexed by  $\{0, 1, 2, \ldots\}$ ), i.e., negotiation is conducted in discrete time. A problem with the general one-tomany negotiation mechanism [7], [12], [16], [20] is that during negotiation, no matter how long an agent has to wait and how many proposals have been received, the agent cannot propose until it has received proposals from all its trading partners. In actual negotiation environments, as agents may have different negotiation strategies, reasoning mechanisms, communication time, constraints, and preferences, an agent generally receives its trading partners' proposals at different times in each round after it sent proposals to all its trading partners at the same time (i.e., different trading partners have different reaction times). Therefore, the general one-to-many negotiation mechanism is

There are three critical issues in designing a flexible one-to-

To overcome the limitation of the general one-to-many

1) How to coordinate all the subnegotiation threads (Section II): One-to-many negotiation can be treated as a series of subnegotiation threads, and different subnegotiation threads have different negotiation situations. A coordination strategy concerns issues such as whether all the subnegotiation threads are interactive or not and

<sup>1</sup>The reasons that bring about different trading partners with different reaction times vary. For example, agents located in different positions in the network may have different communication distances. The factors affecting communication quality of service (QoS), e.g., bandwidth, congestion, and Computer Science, Chongqing University, Chongqing 400044, China (e-mail: boancqu@gmail.com).

K. M. Sim is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: bsim@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TSMCB.2006.874686

1083-4419/\$20.00 @ 2006 IEEE

#### A Survey of Bargaining Models for Grid Resource Allocation

KWANG MONG SIM

Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, KLN, Hong Kong. Email: prof\_sim\_2002@yahoo.com

Whereas it is noted that various types of auctions, commodity market models, and contract-net (tendering) model are more widely used for managing Grid resources, this paper focuses on discussing bargaining (negotiation) models for Grid resource management. To this end, this survey supplements and complements existing surveys by reviewing, comparing, and highlighting the very few extant research initiatives on applying bargaining as a mechanism for managing Grid resources. The contributions of this paper are (i) discussing the motivations for considering bargaining models for Grid resource management, (ii) discussing the issues in building bargaining mechanisms for Grid resource management, (iii) comparing the strategies and protocols of state-of-the-art bargaining models for Grid resources, and (iv) discussing possible new directions.

Additional Key Words and Phrases: Automated negotiation, bargaining, Grid resource allocation, Grid resource management, computational economy.

#### 1. INTRODUCTION

Since a computational Grid focuses on large-scale resource sharing, a resource management system is central to its operations [1, p.135]. However, providing efficient resource allocation mechanisms in the Grid is a complex undertaking due to its scale and that resource owners and consumers may have different goals, preferences, and policies. In a position paper by Sim [2], it was argued that software agents (or automatic scheduling programs), in particular negotiation agents, can play an essential role in realizing the Grid vision. Numerous economic models for Grid resource management such as commodity market models, auction, contract-net/tendering models, bargaining models, posted price models, bid-based proportional resource sharing models. cooperative bartering models, and monopoly and oligopoly had been proposed in the literature and were summarized in [3-4]. While some of the more commonly-referenced work (e.g.,[5-8]) focused on commodity markets, auction and contract-net/tendering models for Grid resource management, this paper focuses on discussing bargaining (negotiation) models for Grid resource management. The intention of this paper is to supplement and complement the existing survey papers on Grid resource management [1,3-4] by reviewing and highlighting the very few extant research initiatives on applying automated negotiation as a mechanism for managing Grid resources. The contributions of this paper are listed as follows. Section 2 discusses the motivations for considering automated negotiation as a model for allocating Grid resources. Section 3 discusses the challenges of the bargaining problem in Grid resource management and identifies some issues for consideration when building negotiation mechanisms for Grid resource management. Section 4 reviews the state-of-the-art bargaining models for Grid resource management, of which there are not many. Section 5 concludes this paper by comparing the bargaining models in section 4, and discusses new directions and open problems.

Permission to make digital/hard copy of part of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication, and its date of appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific © 2005 ACM 1529-3785/2005/0700-0024 \$5.00

ACM SIGecom Exchanges, Vol. 5, No. 5, December 2005, Pages 22-32.

Computational Intelligence, Volume 20, Number 4, 2004

#### NEGOTIATION AGENTS THAT MAKE PRUDENT COMPROMISES AND ARE SLIGHTLY FLEXIBLE IN REACHING CONSENSUS KWANG MONG SIM

Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan

The contribution of this work is designing and developing enhanced market-driven agents with the flexibility have shown that market-driven agents (MDAs) make prudent compromises by reacting to changing market situations by taking into account factors such as competition, deadlines, and trading options. This work augments the design of an MDA with three fuzzy decision controllers that guide the agent in (i) relaxing trade aspiration in face of intense negotiation pressure, and (ii) raising trade aspiration in extremely favorable markets. Results from extensive simulations conducted using an implemented testbed suggest that when compared to MDAs, agents in this work achieved (1) higher success rates in reaching deals, (2) higher average utilities, and (3) higher expected utility. Key words: negotiation agent, automated negotiation, agent-based e-commerce, agent-based simulation, mul-

#### 1. INTRODUCTION

Even though there are many extant negotiation agents for e-commerce (e.g., Chavez et al. 1997; Faratin, Sierra, and Jennings 1998), the negotiation strategies of agents in these systems seldom (i) consider the dynamics of the market and/or (ii) examine the trade-off between the benefit of accepting a suboptimal proposal to reach a consensus more quickly (or with more certainty) and the benefit of postponing an agreement in search of an optimal outcome. While neglecting these issues in some conventional e-commerce systems (e.g., e-trading systems for used books) may not necessarily be crucial, it is envisioned that agents negotiating for resources in future computational grids must take into consideration both the dynamics of the grid computing environment as well as the speed at which resources are allocated. One would expect that resources and services are constantly being added or removed from a grid (Cao et al. 2001, 2002). Furthermore, because a grid reacts at machine speed, any delay incurred by waiting for resource assignment may be perceived as an overhead (Wolski, Plank, and Brevik 2001). Both of these considerations collectively provide the impetus for defining some of the design considerations of negotiation agents that are envisioned to participate in trading resources in future market-oriented grids (Wolski et al. 2003). Following these design considerations, the contribution of this work is designing and developing flexible negotiation agents with two novel features: (1) adopt a market-driven strategy when making a concession, and (2) have flexible zones of concession.

Market-driven strategy: In an e-market (respectively, a market-oriented grid), the conditions for deliberation change as new opportunities and threats (respectively, requests and services) are constantly being introduced as traders (respectively, consumers and resource/service providers) enter and leave a market. In addition, because a deadline puts negotiators under pressure (Kennedy 1994, p. 67), deliberation is also bounded by time. Even though there are negotiation agents that take into consideration the passage of time (e.g. Faratin et al. 1998; Kraus and Schechter 2003), they do not consider market dynamics. The strategic negotiation model of Kraus and Schechter (2003) enables self-motivated rational agents to share resources but the model only considers bilateral negotiation. Although Faratin et al. (1998) define a range of strategies and tactics for generating (counter-)proposals based

Address correspondence to Kwang Mong Sim, Institute of Information Science, Academia Sinica, 128 Academia Road.

© 2004 Blackwell Publishing, 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford OX4 2DO, UK.

#### Grid Commerce, Market-Driven G-Negotiation,

IBBE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 36, NO. 6, DECEMBER 2006

#### and Grid Resource Management Kwang Mong Sim

for realizing a computational grid, providing an efficient re-source allocation mechanism is a complex undertaking. Since Grid providers and consumers may be independent bodies, negotiation among them is necessary. The contribution of this paper is showing that market-driven agents (MDAs) are appropriate raised in [5] by designing and building negotiation agents that tools for Grid resource negotiation. MDAs are e-negotiation agents designed with the flexibility of: 1) making adjustable amounts of market-oriented Grid [9]-[11]. concession taking into account market rivalry, outside options, and time preferences and 2) relaxing bargaining terms in the face of intense pressure. A heterogeneous testhed consisting of several types of e-negotiation agents to simulate a Grid computational economies for controlling the resource allocation in computational Grid environments. The Grid can be viewed as a of MDAs against other e-negotiation agents (e.g., Kasbah) in a network of computations [12], and computations can be viewed

Index Terms-Grid commerce, Grid resource allocation, negotiation, resource management, software agent.

#### I. INTRODUCTION

RID COMPUTING is distinguished from conventional systems should provide the tools and mechanisms for both distributed computing because it focuses on large-scale providers and consumers to express their requirements and resource sharing [1, p. 200]. Hence, a resource management facilitate the realization of their goals. A Grid economy not system is central to the operation of a Grid [2, p. 135]. A only helps regulate the supply and demand for Grid resources, Grid is a very large-scale network computing system that can but also provides the incentives for providers to contribute potentially scale to Internet size, and the network computing resources and benefit from doing so and offers an efficient system can be viewed as a virtual computer consisting of a mechanism for managing resources [14, p. 699]. undertaking. Hence, automatic scheduling programs are needed different goals and knowledge. to (re)allocate computing resources because of both the complexity of the resource allocation problem and the dynami-systems and some of their assumptions seem to be familiar (e.g., cally changing performance of the Grid resources [3, p. 747]. many people can associate price to the supply-and-demand Agents (or autonomous problem solvers) that can act flexibly patterns of resources). Moreover, these economic principles in dynamic environments can provide supportive efforts for also extend to artificial decision-making agents in general

April 4, 2006. This work was supported in part by a research grant from the Hong Kong Baptist University, Project FRG/04-05/II-65. This paper was recommended by Associate Editor A. F. Gomez Skarmeta. by Associate Editor A. F. Gomez Skarmeta. by Communication of the Hong Kong Baptist University, Project FRG/04-05/II-65. This paper was recommended by Associate Editor A. F. Gomez Skarmeta. by Communication of the Hong Kong Baptist University, Project FRG/04-05/II-65. This paper was recommended by Associate Editor A. F. Gomez Skarmeta. The author is with the Hong Kong Baptist University, Kowloon, Hong Kong 2) Market-Driven G-Negotiation: Whereas [3], [7], and [8] (e-mail: prof\_sim\_2002@yahoo.com).
Digital Object Identifier 10.1109/TSMCB.2006.877791 focused on both commodity markets and auction formulations

Abstract—Although the management of resources is essential a computational Grid [4, p. 8]. Sim [5] argued that software 1) G-commerce and Market-Oriented Grid: In [3], [7], and

generally achieve: 1) higher budget efficiencies in many market situations than other e-negotiation agents in the testbed and 2) higher success rates in acquiring Grid resources under high Grid loadings. First, the utilization of Grid resources is not for free [3]. In

a market-oriented Grid, providers can receive royalties for the (computing and storage) resources and services they provide, whereas Grid users can attempt to mold the Grid systems to their needs by exercising their market powers as Grid consumers. In a Grid economy [14, p. 699], resource management

networked set of heterogeneous machines (owned by multiple Second, there is an enormous literature on economic theories organizations) that agrees to share their local resources with and principles for explicating and understanding the emergent each other [2, p. 135]. Due to its scale, and because resource behavior of the Grid and its constituents (participants). It was owners and consumers may have different goals, preferences, also noted in [13, p. 134] that using market models as an interests, and policies, providing an efficient resource manage- economic organization for computation is effective in promotment and coordination mechanism in the Grid is a complex ing efficient and cooperative interactions among entities with

[13, p. 134], including software agents and entities. Finally, it was noted in [14, p. 699] that some of the economic Manuscript received August 11, 2005; revised January 23, 2006 and models for resource allocation include: commodity market

#### A Pilot Study on the Impact of the Web-based Essay Critiquing System on Writing at the Tertiary Level

Kelvin C.K. Wong<sup>\*\*</sup>, Fion S.L. Lee\*, Cynthin Γ.K. Lee\*, William K.W. Cheung\*, Anders I. Morels and Jiming Liu\*

\*Department of Computer Science, Hong Kong Baptist University, Hong Kong. Language Centre, Hong Kong Baptist University, Hong Kong. InterMedia, University of Oslo, Norway. \*Corresponding Author: kckwong@comp.hkbu.edu.hk

#### ABSTRACT

As English is widely used worldwide, it is the preferred second language in Hong Kong. Many students find essay writing stressful because they do not have sufficient ideas to fully cover the topic of the essay. To alleviate learning berriers while writing, a web-based essay critiquing system was developed using Latent Semantic Analysis. (LSA). LSA is an automatic text analysis technique for providing just in time feedback to students. The feedback takes two forms: new sub-themes suggested to be included, and the visualization of the existing sub-thenes' organization. In this paper, we present our findings on students' performance and their perception of the usefulness of this system

Keyword: Essay Writing, Latent Semantic Analysis, Critiquing System.

#### 1. Inaudaction

Acquiring good essay writing skills is acknowledged as important, and yet challenging for students in Horg Kong, Many of them find the essay-writing task stressful because they are short of ideas to fully discuss the essay topic and complete the essay. If it is an in-class writing exercise, it will be difficult for the teacher to give immediate content-specific hints to each individual student. If it is a take-home writing exercise, getting immediate feedback from the teacher is simply impossible. Students usually make multiple drafts before finalizing their essays. Even if the teacher can afford to provide feedback to each draft, the turn-around time will often be in days at least. This makes the learning process quite meffective.

To alleviate the aforementioned learning barriers, a computer-supported critiquing system was developed using Latent Sementic Analysis (LSA). LSA (Landauer and Laham, 1998) is an automatic text analysis technique, for providing just-in-time feedback to students. The feedback takes two forms:

- new sub-themes suggested to be included, and
- the visualization of the existing sub-themes' organization.

# Networking and Multimedia

Dynamic Routing and Wavelength Assignment in the Presence of Wavelength Conversion for All-Optical Networks

Xiaowen Chu, Member, IEEE, and Bo Li, Senior Member, IEEE

effective Routing and Wavelength Assignment (RWA) algorithm wavelength conversion as they usually only take into account the wavelength conversion as they usually only take into account the current traffic, and do not explicitly consider the route lengths.

We then propose a weighted least-congestion routing and first-fit wavelength assignment (WLCR-FF) algorithm that considers both the current traffic load and the route lengths jointly. We further introduce an analytical model that can evaluate the blocking performance for WLCR algorithm. We carry out extensive numerical wavelength conversion see aliminate the wavelength along any route.

Wavelength conversion as they used in the current traffic least considers both the current traffic load and the route lengths jointly. We further introduce an analytical model that can evaluate the blocking performance for WLCR algorithm. We carry out extensive numerical studies over typical topologies including ring, mesh-torus, and the 14-node NSFNET; and compare the performance of WLCR-FF with a wide variety of existing routing algorithms including results conclusively demonstrate that the proposed WLCR-FF algorithm can achieve much better blocking performance in the

Index Terms—Routing and wavelength assignment (RWA), wavelength conversion, wavelength division multiplexing (WDM).

I. INTRODUCTION

WAVELENGTH-ROUTED all-optical WDM networks are considered to be candidates for the next generation of wide-area backbone networks [21]. A wavelength-routed all-optical WDM network comprises optical wavelength routing nodes (i.e., wavelength routers) interconnected by optical fiber links. A lightpath has to be established before the communication between any two wavelength routers [5]. To establish a lightpath, it is normally required that the same wavelength channel should be allocated on all the links along the route. This imitation is known as the wavelength continuity constraint,

Manuscript received June 21, 2002; revised April 17, 2004, approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor C. Qiao. This work was supported in part by grants from RGC under Contracts HKBU2159/04E, HKUST6204/03E and HKUST6104/04E, a grant from NSF China under Contract 60429202, a NSFC/RGC joint grant under Contract N\_HKUST605/02, and a grant from Microsoft Research under Contract MCCL02/03.EG01.

X.-W. Chu is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China (e-mail: chxw@comp.hkbu.edu.nk).

B. Li is with the Department of Computer Science, Hong Kong University and Technology, Hong Kong, China (e-mail: bli@cs.ust.hk).

Digital Object Identifier 10.1109/TNET.2005.850226

Abstract—Blocking probability has been one of the key performance indexes in the design of wavelength-routed all-optical WDM networks. Existing research has demonstrated that an effective Routing and Wavelength Assignment (RWA) algorithm as a random holding period. These lightpaths need to be set up dynamically by determining a route across the network and wavelength conversion are two primary vehicles for improving the blocking performance. However, these two issues have largely been investigated separately; in particular the existing RWA algorithms have seldom considered the presence of wavelength conversion. In this paper, we firstly demonstrate that the existing dynamic RWA algorithms do not work well in the presence of the destination, and allocating a free wavelength channel on each fiber link along the chosen route. In this paper, we assume that existing lightpaths cannot be re-routed to accommodate new lightpath requests until they are released. Some of the lightpath requests could be blocked if released. Some of the lightpath requests could be blocked if

Wavelength conversion can eliminate the wavelength con-

IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 13, NO. 3, JUNE 2005

tinuity constraint and thus improve the blocking performance [14], [17]. A wavelength router with wavelength conversion capability is called a wavelength convertible router (WCR). If all the wavelength routers can support wavelength conversion, this is referred to as full wavelength conversion. Because waveength converters are still very expensive at the current stage, much research work focuses on sparse wavelength conversion, in which only part of the wavelength routers have the capability of wavelength conversion. Subramaniam et al. have shown that, by using sparse wavelength conversion, a relatively small number of WCRs can achieve satisfactory performance [24]. The key issue in sparse wavelength conversion architecture is the wavelength converter placement problem, i.e., given a network topology, a certain number of WCRs, and the traffic statistics, how can the WCRs be placed so as to minimize the overall blocking probability? The algorithms for optimal converter placement for simple topologies, such as bus and ring, have been shown in [25]. However, optimal converter placement for arbitrary mesh is considered to be very hard. An efficient algorithm for optimal converter placement has been proposed in which makes the wavelength-routed networks different from [26], which is more efficient than exhausted search solution. A large number of heuristic algorithms also exist in the literature

> [1], [8], [10], [12], [27]. The existing research demonstrates that an effective routing and wavelength assignment (RWA) strategy and wavelength conversion are the two primary vehicles for improving the blocking performance [6], [13], [20]. However, these two issues have largely been investigated separately in that existing RWA algorithms have seldom considered the presence of wavelength conversion, whilst the wavelength converter placement algorithms have largely assumed that a static routing and

Paper approved by W. C. Kwong, the Editor for Optical Communications of the IEEE Communications Society. Manuscript received November 6, 2001; revised June 7, 2002; August 24, 2002; October 15, 2002. This work was supported in part by Research Grants Council (RGC) grants under Contract AoE/E-01/99, Contract HKUST 6195/02E. This paper was presented in part at the SPIE Optical Networking and Communications Conference (Doi:Compt.) 2002. Boston MA. July 29-August 2, 2002. Conference (OptiComm) 2002, Boston, MA, July 29-August 2, 2002. X. Chu and B. Li are with the Department of Computer Science, Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail:

ılamtac@utdallas.edu). Digital Object Identifier 10.1109/TCOMM.2003.810834

IEEE TRANSACTIONS ON COMMUNICATIONS, VOL. 51, NO. 4, APRIL 2003

Wavelength Converter Placement Under Different RWA Algorithms in Wavelength-Routed All-Optical Networks

Xiaowen Chu, Student Member, IEEE, Bo Li, Senior Member, IEEE, and Imrich Chlamtac, Fellow, IEEE

Abstract-Sparse wavelength conversion and appropriate routing and wavelength assignment (RWA) algorithms are the two key factors in improving the blocking performance in wavelength-routed all-optical networks. It has been shown that the optimal placement of a limited number of wavelength converters in an arbitrary mesh network is an NP-complete problem. There have been various heuristic algorithms proposed in the literature, in which most of them assume that a static routing and random-wavelength assignment RWA algorithm is employed. However, the existing work shows that fixed-alternate routing and dynamic routing RWA algorithms can achieve much better
blocking performance. Our study further demonstrates that the wavelength converter placement and RWA algorithms are closely related in the sense that a well-designed wavelength converter placement mechanism for a particular RWA algorithm might not work well with a different RWA algorithm. Therefore, the wavelength converter placement and the RWA have to be considered jointly. The objective of this paper is to investigate the wavelength converter placement problem under the fixed-alternate routing (FAR) algorithm and least-loaded routing (LLR) algorithm. Under the FAR algorithm, we propose a heuristic algorithm called minimum blocking probability first for wavelength converter placement. Under the LLR algorithm, we propose another heuristic traffic statistics, how can the wavelength converters be placed algorithm called weighted maximum segment length. The objective into the network in order to minimize the overall blocking probability? Usually, this is addressed as a separate issue that is blocking probability. Extensive simulation studies have been carer three typical mesh networks, including the 14-node NSFNET, 19-node EON, and 38-node CTNET. We observe that

Index Terms—Routing and wavelength assignment (RWA), assume that the static routing and random wavelength assignment (RWA) algorithm is employed.

Nevertheless, the literature results show that the blocking

I. Chlamtae is with the Erik Johnsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75083 USA (e-mail:

I. INTRODUCTION WAVELENGTH-ROUTED all-optical networks are considered to be candidates for the next generation wide-area backbone networks [4], [14]. In such networks, wavelength conversion (or translation) plays an important role on sparse wavelength conversion, which means that only conversion, while others have no conversion capability [15]. If all the network nodes are capable of wavelength conversion,

this is referred to as full wavelength conversion. It has been shown in [15] that, by using sparse wavelength conversion, a relatively small number of converters can achieve satisfactory performance. However, the problem of wavelength converter placement was not considered. That is, given a network topology, a certain number of wavelength converters, and solved by converter placement algorithms. The algorithms for optimal converter placement in simple topologies, such as bus the proposed algorithms not only outperform existing wavelength and ring, have been provided in [16]. However, optimal conconverter placement algorithms by a large margin, but they also verter placement for more realistic topologies such as arbitrary can achieve almost the same performance compared with full wavelength conversion under the same RWA algorithm. algorithms have been proposed [1], [9], [11], [18]. All of them

Nevertheless, the literature results show that the blocking probabilities of wavelength-routed networks are heavily de pendent on the RWA algorithms [5], [8], [12]. Our studies also demonstrate that a well-designed wavelength converter placement mechanism for the static RWA algorithm does not work well under a different RWA algorithm. Therefore, we argue that wavelength converter placement and RWA algorithms should

be considered jointly. In this paper, we investigate the problem of wavelength converter placement under two RWA algorithms, both of which have exhibited that better blocking performance can be obtained. The first one is the fixed-alternate routing and first-fit wavelength assignment (FAR-FF RWA) algorithm. The second

<sup>1</sup>In this paper, wavelength converter means a wavelength router with the capability of wavelength conversion.

0090-6778/03\$17.00 @ 2003 IEEE

IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, VOL. 21, NO. 7, SEPTEMBER 2003

#### Resource Management in An Integrated Optical Network

Kazem Sohraby, Senior Member, IEEE, Zhensheng Zhang, Senior Member, IEEE, Xiaowen Chu, Member, IEEE,

Abstract-We propose a novel integrated optical network nodes, and core network components. The customer premises switching architecture. The proposal offers an approach to signaling for the purpose of transport on an all-optical network of optical and nonoptical legacy network traffic. In order to provide effective end-to-end control and efficient transport services, new signaling and control techniques are required. Standard organizations such as Optical Interworking Forum (OIF) and Internet Engineering Task Force have developed interface methods be-tween client and transport networks, as well as signaling processes

ment along with the combination of electronic and optical
networking architectures results in a nonintegrated and often for resource allocation (Benzamin, 2001). We propose a network troller, which implements interfaces for such integration in the intermediate future, as well as provides a feasible path for the long-term objective of all optical networking. Performance and capacity issues for these systems introduce new dimensions to the existing set of networking problems, since optical paths can now be set up in real-time. There are two main contributions in this paper: 1) functional composition of a network controller, which translates legacy signaling to optical connection signaling and path establishment and 2) determining when to issue an optical nnection request based on the current network conditions such as link utilization, so that the integrated optical network can operate efficiently. Analytical approximations, as well as simulation results for call blocking performance are also presented.

therefore, critical to efficiently integrate legacy systems with the emerging optical networks.

Today's networks are an increasingly complex interconnec-

Manuscript received August 12, 2002; revised March 20, 2003. K. Sohnaby was with the Bell Labs, Lucent Technologies, Holmdel, NJ 07733 USA. He is now with the Department of Computer and Computer Engineering, College of Engineering, University of Arkansas, Fayetteville, AR 72701 USA Z. Zhang is with Waterridge Networks, San Diego, CA 92130 USA (e-mail: zhang@ieee.org).

X. Chu is with the Department of Computer Science, Hong Kong Baptist

Available at

www.ElsevierComputerScience.com

Information Sciences 157 (2003) 33-71

POWERED BY SCIENCE @ DIRECT®

The impact of load balancing

to object-oriented query execution

scheduling in parallel machine environment

David Taniar a,\*, Clement H.C. Leung b

<sup>a</sup> School of Business Systems, Monash University, P.O. Box 63B, Clayton, Vic. 3800, Australia

<sup>b</sup> School of Communication and Informatics, Victoria University of Technology,

P.O. Box 14428 MCMC, Melbourne 8001, Australia

University, Kowloon, Hong Kong (e-mail: chxw@comp.hkbu.edu.hk).

B. Li is with the Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: li@cs.ust.hk).
Digital Object Identifier 10.1109/JSAC.2003.815978

equipment interface, multiplexing, switching, optical routing, and transport are performed separately under the control of hubs and routers, multiplexers, local switches, optical add/drop multiplexers (OADMs), optical cross-connects (OXCs), and high-capacity DWDM. Currently, this complex set of equipnetworking architectures results in a nonintegrated and often separate control for the user traffic over the electronic and optical networks. In principle, the interaction among electronics and optical devices for control and routing makes it possible to deliver traffic from premises equipment to the network for end-to-end transport. However, there does not exist an integrated, single point of control between electronic and optical domains. Another requirement for integrated optical network in-

frastructure is as a direct result of increase of volume of

multimedia services on the Internet. With the high growth of the data traffic and multimedia services, utilization of optical transport has grown at an unprecedented rate. Along Index Terms-Control, networking, optical networks, perfor-(possibly with optical interfaces) connect to the all-optical transport networks. These developments have fundamentally changed the way in which transport networks are designed and operated. The emerging new applications and services on VER THE LAST decade, technologies related to optical networking, especially the dense wavelength division the optical transport network, such as "bandwidth on demand" and "bandwidth trading" require that the transport network multiplexing (DWDM), have made significant progress. Many respond dynamically to mostly unpredictable demands. An DWDM based systems have been deployed to satisfy the ever important conclusion drawn from the new set of applications increasing Internet traffic demand. However, a large number and services is that bandwidth of the transport network needs of legacy networks such as ATM, frame relay, SONET/SDH to be dynamically controlled and allocated on a real-time basis. tegrated optical networks. In this paper, we study several key components in an integrated optical network; a network controller which for the purpose of traffic transport, protion of customer premises equipment, access and switching vides signaling and interface to the legacy systems, optical cross-connects, and switching equipment, see Fig. 1. The controller allocates resources for the incoming requests either for the optical path setup and wavelength establishment, or for establishing connections for the requests on the current optical paths; and performs the functions necessary for provisioning of quality-of-service (QoS) and performance guarantees. A major role of the controller, in addition to signaling and control is to monitor the resources in the network in a way that request for a connection over an existing optical path, or a wavelength setup can be optimally allocated while maintaining the performance expectations. The controllers can be managed and operated by the network operators, or the service providers.

INFORMATION

**SCIENCES** 

www.elsevier.com/locate/ins

0733-8716/03\$17.00 @ 2003 IEEE

### Provision of Continuous VCR Functions in Interactive

IEEE TRANSACTIONS ON BROADCASTING, VOL. 51, NO. 4, DECEMBER 2005

Broadcast VoD Systems W.-F. Poon, K.-T. Lo, and J. Feng

1063-6692/\$20.00 @ 2005 IEEE

Abstract-Video broadcasting is one of the feasible solutions

Index Terms—Interactive TV, VCR functionality, video broadcast, video-on-demand (VoD).

I. INTRODUCTION

W.-F. Poon and K.-T. Lo are with the Centre for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: enktlo@polyu. J. Feng is with the Department of Computer Science, Hong Kong Baptist iversity, Kowloon Tong, Hong Kong. Digital Object Identifier 10.1109/TBC.2005.857607

Abstract—Video broadcasting is one of the feasible solutions to implement a large-scale video-on-demand (VoD) system. Nevertheless, it is still an open issue for the provision of continuous VCR functions in a delay insensitive broadcast VoD system. In this paper, we propose to jointly optimize an active buffer management scheme with contingency channels to support the VCR functions in an efficient protocol called partitioned broadcasting. We develop a greedy channel management scheme by exploiting the property of the broadcasting protocol such that the system bandwidth capacity can be fully utilized. Incorporating the channel management scheme with the partitioned video broadcast, the VoD system not only provides delay insensitive services but also handles all the interactive requests. Extensive simulation results demonstrate that the partitioned broadcasting system outperforms the traditional system based on the staggered broadcasting protocols. It is found that 20 broadcasting channels and 10 contingency channels are sufficient to support on average 720 customers for a single video with staggered start time. The start-up delay is equal to the length of the video divided by the number of the channels allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the delay of broadcasting allocated to this video. For instance, the de a video with staggered start time. The start-up delay is equal to ferent data broadcasting techniques for VoD services can be

However, in such efficient broadcasting systems, it is still an open issue for implementation of continuous VCR functions. In [13], Almeroth et al. first proposed to use the Set-Top-Box TIDEO-ON-DEMAND (VoD) [1], [2] services allow geo- (STB) buffer to support the VCR functions in the staggered graphically distributed subscribers to interactively access broadcasting system. Later on, the Split-And-Merge (SAM) video files from remote VoD servers. Although the unicast VoD protocol [14] was proposed to cover the details for all the insystem, which serves each user by an individually allocated teraction types that include Jump Forward/Backward (JF/JB), channel, can respond to requests immediately, it is neither ef- Pause (PAU) and Fast Forward/Reverse (FF/FR). The basic idea ficient nor cost effective to support a large number of users of SAM is to use separated channels called interactive channels mainly due to limited capacity in a central server. Many studies (called contingency channels in this paper) to serve for the then explored multicast/broadcast delivery strategies to imple- VCR functions. When the playout resumes, the server attempts ment a Near-VoD (NVoD) system. In general, scheduled multicast [3]-[6] and periodic broadcast are the two major tech- Each interactive user is allocated a buffer of maximum size SB niques for NVoD provision. The former approach allocates a located at the access nodes for the synchronization purpose. The video channel to a group of users waiting for the same video similar idea was proposed in [15] but the authors suggested that according to a scheduling policy such as FCFS [3] and MFQL a decentralized nonshared buffer should be more suitable for a [4]. In the periodic broadcast, the system dedicates several channels to a single video and each channel broadcasts the whole VCR functions were also analyzed in [16] but the system is also video or a part of it in cycles. Different from the scheduled mul- based on the staggered broadcasting scheme. In addition to using ticast, a user does not explicitly request for a video but rather the contingency channels, a novel management scheme called waits till the video is broadcast. One of the traditional schemes is Active Buffer Management (ABM) [17], [18] was developed to called staggered broadcasting [7] which periodically broadcasts support the VCR functions. With ABM, a user will receive data from several broadcasting channels at the same time in order to Manuscript received August 20, 2003; revised August 15, 2005. The work of maintain the play point in the middle of the receiver buffer. Nev-K.-T. Lo was supported by the Hong Kong Polytechnic University under Grant ertheless, if the user has limited storage capacity, a continuous VCR function still cannot be guaranteed in such a system. Although many works have been done to implement interactive broadcasting VoD systems, they are all based on a simple staggered VoD system that introduces significant start-up latency. To the best of our knowledge, no solution has been provided in other efficient broadcasting systems. This motivates

**Standards** 

John R. Smith

#### **Multimedia Retrieval Benchmarks**

National Institute

However, two recent benchmarking efforts, TREC information to the participants. Video Retrieval Evaluation (TRECVID) and the Then in November, the research groups who Clement Leung | IAPR TC-12 Benchmark for Visual Information completed one or more of the tasks present their Victoria | Search, show great promise for letting the multi-results and exchange ideas with each other at a University, media retrieval community collaborate on short workshop at NIST. Because multiple groups Australia | advancing the state of the art. This Standards | have worked on the same tasks with the same department features two articles that describe data, it makes for productive comparisons and these respective benchmarking efforts and disdiscussions. Horace Ip | cuss the benefits to the multimedia community.

City University o Hong Kong

Grubinger

**TREC Video Retrieval Evaluation** tion. Open, metrics-based evaluations organized have demonstrated their value in accelerating nition and text information retrieval systems. With this in mind, IAD's Retrieval Group launched a benchmarking effort with interested | high-level feature (concept) extraction, and researchers in 2001 aimed at establishing a com-

mon evaluation framework for the scientific comparison of digital video retrieval technologies and systems: the TREC Video Retrieval Evalua-Advanced Research and Development Activity tralia, Europe, and the US. (ARDA) and NIST. Annual TRECVID cycle

proposed set of guidelines for that year's evaluathe test data, the tasks the systems will be tested Each topic contains a short textual description of

ultimedia retrieval systems' development on, how systems will be evaluated, and the schedule. During the summer and early autumn, standard benchmarks for evaluating and com- researchers run their systems on the test data, paring the performance of emerging techniques. and NIST evaluates the results and returns that

**Evolving evaluation framework** 

Over the last 3 years, TRECVID has evolved in Within the US National Institute of Standards many ways. The data used, while always in and Technology's (NIST) Information Technolo- MPEG-1, has grown from 11 hours to 133 and gy Laboratory, the Information Access Division changed from ephemeral, informational materi-(IAD) is charged with providing measurements al from 1940 to the 1970s (Internet/Prelinger and standards to advance technologies dealing Archive) to US news programming (ABC/CNN with multimedia and other complex informa- and C-SPAN) from 1998. Four tasks now exist:

by IAD using standard data, tasks, and measures | shot boundary detection (gradual versus cut),

progress in developing automatic speech recog-

search (fully interactive and not).

The project started with 12 groups of particition (TRECVID). TRECVID is funded by the US pants and has increased to 24 from Asia, Aus-The 17 features tested in 2003 included various

mid- to high-level concepts such as outdoors, peo-

ple, female speech, news subject monologue, TRECVID works like this. Early in the calendar sporting event, zoom in, and physical violence. If year a call for participation goes out. Interested a feature's target is present in at least one frame of parties respond and work with NIST to refine a a shot, the feature is considered true for the shot. The search task is based on 25 statements of tion. The guidelines spell out the training data, information needs (or topics) NIST developed.

1070-986X/04/\$20.00 © 2004 IEEE

Published by the IEEE Computer Society

### Received 19 February 2003; received in revised form 16 March 2003; accepted 2 May 2003

Abstract

Skewness has been one of the major problems not only in parallel relational database systems, but also in parallel object-oriented database systems. To improve performance of object-oriented query processing, a careful and intelligent skew handling for load balancing must be established. Depending on the parallel machine environment, whether it is a shared-memory or a shared-nothing architecture, load balancing can be achieved through "physical" or "logical" data re-distribution. It is not the aim of this paper to propose or to investigate skew handling methods, but rather to analyze the impact of load balancing to query execution scheduling strategies. Our analysis shows that when load balancing is achieved, "serial" execution scheduling is preferable to "parallel" execution scheduling strategy. In other words, allocating full resources to a sub-query seems to be better than dividing resources to multiple sub-queries. © 2003 Elsevier Inc. All rights reserved.

Keywords: Query access plans; Execution scheduling; Serial and parallel execution scheduling; Inter/intra-operation parallelization; Parallel query processing; Parallel object-oriented databases

\*Corresponding author. Tel.: +61-3-99059693; fax: +61-3-99055159. E-mail addresses: david.taniar@infotech.monash.edu.au (D. Taniar), clement@matilda.vu. edu.au (C.H.C. Leung).

0020-0255/\$ - see front matter © 2003 Elsevier Inc. All rights reserved. doi:10.1016/S0020-0255(03)00152-X

#### COMPARATIVE EVALUATION OF WEB IMAGE SEARCH ENGINES FOR MULTIMEDIA APPLICATIONS

0018-9316/\$20.00 @ 2005 IEEE

Keon Stevenson Clement Leung School of Computer Science and Mathematics Victoria University, Footscray, Vic 3011, Aus keon182@yahoo.com clement.leung@vu.edu.au

ABSTRACT While text-oriented document searching are relatively

mature on the Internet, image searching, which requires much more than text matching, significantly lags behind. The use of image search engines significantly enlarges the scope of images to users accessibility. This paper provides an understanding of current technologies in image searching on the Internet, and points to future areas of improvement for multimedia applications. We develop a systematic set of image queries to assess the competence and performance of the major image search engines. We find that current technology is only able to deliver an average precision of around 42% and an average recall of around 12%, while the best performers are capable of producing over 70% for precision and around 27% for recall. The reasons for such differences, and mechanisms for search improvement, are also

#### 1. INTRODUCTION

There are a number of different types of Image Search 2, 6-8, 10, 11]. Currently most major search engines fall ISE such as Google and Yahoo are of basic design having large databases and relying on mixed retrieval for accuracy. The technology is very basic; it is indexed in erms of text label/words coupled with an image. Specialized ISEs that are dedicated to indexing mages or multimedia such as Corbis & Getty Images. These sites are often experimental and have limited atabases restricted by size when compared with enormous sites such as Google. Finally there are image search engines that are called Meta-Search engines, which send users requests to multiple search engines and then display the 'multiple'

Content-based systems use technologies that refine searches by focusing on content and objects within the image rather then text. Visual-meta tagging is another method like content-based retrieval that organizes images into groups of relevance using their visual content. Finally, whilst most current images search technologies on the Internet are based on 2D images, it is worth mentioning that a method that is being developed [4,5,9] for 3D image search that uses algorithms and matical models of each shape to determine image

0-7803-9332-5/05/\$20.00 ©2005 IEEE

Whilst current major image search engines in use are limited in terms of high functionality (precision and recall), the need for a performance evaluation of current engines will be of great benefit to users. It will provide them with an understanding of current technologies, advanced searches and ways of making use of the results for specific multimedia applications.

#### 2. THE SEARCH MECHANISMS With most ISEs, the options available require that an

online user will type in a keyword search and it will match the keywords and present them in a thumbnail manner with the appropriate links. By specifying the file type the ISE can detect the tag associated with the file e.g. JPEG or GIF. In matching a keyword search with a file type an ISE can approximate the content of the file. Engines look for and index online Web sites where the titles of the files accurately describe the content of the image. In using such a method it can lead to error but it also means it can support a larger database as they are catalogued more easily rather than looking at its actual content. As an example, in a search for "cat", the first few result pages will be relevant but it will also match other images that include cat i.e. "cat scan" "different felines" and so on. Other strategies that can assist an ISE can be human involvement where they search the World Wide Web manually and look for images and catalogue them according to their relevant content. In cataloguing images this way more accurate results are achieved but as there are millions of images on the World Wide Web it requires very intense labour to do so. A positive aspect of current image search engines is the way in which they display results from a search. The results are displayed in thumbnail form with an associated link of where the image is originally located and some file information i.e.

When text accompanying an image has little relevance to the content they still can be indexed with the use of ALT tagging which is now a requirement in the US, with the Americans with Disabilities Act [6]. ALT tagging is known as alternative text, which is method used to actively describe an image through text. This allows users that are visually impaired to use voice activation software that informs them with the use of ALT tagging within an image.

If a keyword search within an image search engines is broad then current systems work quite well. Problems begin to surface as searches become more



Available at www.ElsevierComputerScience.com POWERED BY SCIENCE @DIRECT® Information Sciences 157 (2003) 33-71

www.elsevier.com/locate/ins

#### The impact of load balancing to object-oriented query execution scheduling in parallel machine environment

David Taniar a,\*, Clement H.C. Leung b <sup>a</sup> School of Business Systems, Monash University, P.O. Box 63B, Clayton, Vic. 3800, Australia

School of Communication and Informatics, Victoria University of Technology, P.O. Box 14428 MCMC, Melbourne 8001, Australia

Received 19 February 2003; received in revised form 16 March 2003; accepted 2 May 2003

#### Abstract

Skewness has been one of the major problems not only in parallel relational database systems, but also in parallel object-oriented database systems. To improve performance of object-oriented query processing, a careful and intelligent skew handling for load balancing must be established. Depending on the parallel machine environment, whether it is a shared-memory or a shared-nothing architecture, load balancing can be achieved through "physical" or "logical" data re-distribution. It is not the aim of this paper to propose or to investigate skew handling methods, but rather to analyze the impact of load balancing to query execution scheduling strategies. Our analysis shows that when load balancing is achieved, "serial" execution scheduling is preferable to "parallel" execution scheduling strategy. In other words, allocating full resources to a sub-query seems to be better than dividing resources to multiple sub-queries. © 2003 Elsevier Inc. All rights reserved.

Keywords: Query access plans; Execution scheduling; Serial and parallel execution scheduling; Inter/intra-operation parallelization; Parallel query processing; Parallel object-oriented databases

\*Corresponding author. Tel.: +61-3-99059693; fax: +61-3-99055159. E-mail addresses: david.taniar@infotech.monash.edu.au (D. Taniar), clement@matilda.vu.

0020-0255/\$ - see front matter © 2003 Elsevier Inc. All rights reserved. doi:10.1016/S0020-0255(03)00152-X

JOURNAL OF LIGHTWAVE TECHNOLOGY, VOL. 24, NO. 9, SEPTEMBER 2006

#### Lightpath Concentrators for All-Optical Networks Yiu-Wing Leung, Senior Member, IEEE

Abstract—In some deployments of all-optical networks, it is necessary to concentrate the lightpaths from some fibers to fewer fibers. An  $N \times M$  lightpath concentrator is an optical composition nent for this purpose, and it concentrates the lightpaths from N incoming fibers to M outgoing fibers. In this paper, three designs of  $N\times M$  lightpath concentrators are proposed. The first design is a generalization of optical crossconnects, and it requires 2) Example 2: In metro networks, a metro hub collects the concept of partial concentration so that it requires only m imes moptical switches (where m < M) at the expense of a larger

blocking probability. The switch size m can be selected to make

a tradeoff between cost and performance. In the third design, it

is assumed that small lightpath concentrators are available, and these concentrators are used as building blocks to construct large Index Terms-All-optical networks, concentration, concentra-

I. INTRODUCTION

N OPTICAL fiber can provide hundreds of high-speed A channels. If each network node converts the incoming optical signals into electronic signals for routing and switching, it would require hundreds of electronic line cards for each optical fiber. If the node has multiple incoming links with multiple fibers per link, it would have a very large complexity. To avoid this problem, an all-optical network transfers data in the optical domain within the network [1], [2]. Each sourceto-destination path is known as a lightpath [1], [3] because it transfers data optically.

All-optical networks involve a number of research problems.

In the literature, there are studies on network design (e.g., design of logical topology [4], physical topology [5], and node configuration [6]), wavelength conversion [7]-[11], and network control and analysis (e.g., routing and wavelength assignment [3], [8], [12]-[14], distributed network control [15], and analysis of blocking probability [9]-[11]). In some deployments of all-optical networks, it is necessary to concentrate the lightpaths from N fibers to M fibers, where

M < N. Two examples are given as follows. 1) Example 1: In fiber-to-the-curb, an optical network unit

and sends the multiplexed traffic to a regional hub through channels can probably provide well enough bandwidth to handle the traffic from one or a few buildings. The

Manuscript received March 14, 2005; revised May 25, 2006. This work was supported by the Hong Kong Research Grant Council and a HKBU Faculty reported by the Hong Kong Research Grant Council and a HKBU Faculty
essearch Grant.

The author is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: ywleung@comp.hkbu.edu.hk).
Digital Object Identifier 10.1109/JLT.2006.878496

lightpaths from multiple regional hubs through N fiber and transfers these lightpaths to the backbone network through M fibers, where M < N. In this case, the metro hub concentrates the lightpaths from N fibers to M fibers. To the best of our knowledge, there is no existing result on

outgoing fibers for outgoing transmission.

regional hub has multiple incoming fibers from multiple

ONUs, where each fiber only carries a few wavelength

channels. In this case, the regional hub concentrates the

wavelength channels from the incoming fibers to fewer

lightpath concentration. In March 2005, we used some relevant keywords (namely concentrator, concentration, all-optical networks, wavelength-routed networks, etc.) to search the IEEE Xplore Digital Library but did not find a relevant paper. We define an  $N \times M$  lightpath concentrator to be an optical component that concentrates the lightpaths from N incoming fibers to M outgoing fibers. Let there be C wavelength channels at  $\lambda_1, \lambda_2, \dots, \lambda_C$  in each fiber. The concentrator provides the following lightpath concentration function: If there are M or fewer incoming lightpaths at the same wavelength, these lightpaths can be concentrated to the outgoing fibers

[e.g., see Fig. 1(b)]. In this paper, we introduce the problem of designing lightpath concentrators for all-optical networks. In Section II, we identify and discuss the design considerations involved. In Sections III-V, we propose three designs of  $N \times M$  lightpath concentrators. The first design has the smallest blocking probability, the second design is less costly at the expense of a larger blocking probability, and the third design uses small and given lightpath concentrators to construct large lightpath concentration networks.

[e.g., see Fig. 1(a)]; otherwise, some lightpaths are blocked

#### II. DESIGN CONSIDERATIONS

Both electronic and lightpath concentrators are used to concentrate network traffic. However, an electronic concentrator is composed of logic gates, whereas a lightpath concentrator is (ONU) collects the traffic from one or a few buildings composed of optical devices. The problem of designing lightpath concentrators is similar to but different from the problem an optical fiber. For the latter purpose, a few wavelength of designing electronic concentrators. We identify three design considerations for lightpath concentrators as follows.

1) Cost of optical devices: A lightpath concentrator is composed of optical devices. Different optical devices may have very different costs. In general, active devices are more expensive than passive ones, and large devices are more expensive than small ones. In particular, optical switches are active devices, and large optical switches are much more expensive than small ones because of

0733-8724/\$20.00 @ 2006 IEEE

# 0 0

# Networking and Multimedia

IEEE TRANSACTIONS ON COMMUNICATIONS, VOL. 50, NO. 1, JANUARY 2002

#### Design of Node Configuration for All-Optical Multi-Fiber Networks

Yiu-Wing Leung, Senior Member, IEEE, Gaoxi Xiao, Member, IEEE, and Kwok-Wah Hung

future growth, and fault tolerance [15], [16].

When each link has more fibers, the network can provide a

example, if L = 5 and F = 100, a node requires  $500 \times 500$ 

5000×5000 optical switches. This results inthe following two

1) It is difficult to realize large optical switches because of

various technological constraints (e.g., the constraints im-

posed by insertion loss and control complexity). At the

time of revising this paper, the largest all-optical switch

in shipment is 1024×1024. In this case, if a node has

more than 1024 incoming/outgoing fibers, the existing

small optical switches. For example, an optical switch is

composed of a cross-connect and other parts for various

functions such as I/O and control, and the prices of a 8×8

cross-connect and a 1024×1024 cross-connect are about

In this paper, we design a node configuration that requires

· Compared with the existing node configuration, the

proposed design requires significantly smaller optical

switches while it can result in nearly the same blocking

probability. To design a new all-optical network or up-

grade an existing one with spare fibers, we can adopt

the proposed node configuration and slightly more fibers

in each link so that the network requires small optical

Using the existing node configuration, the nodes with dif-

ferent number of incoming/outgoing fibers require optical

switches of different sizes. Using the proposed node con-

figuration, all the nodes can use optical switches of the

same size. This can simplify network implementation and

switches while it has a small blocking probability.

two thousand and four million, respectively.

2) Large optical switches are much more expensive than

node configuration is not applicable.

Abstract—It is cost-effective to install multiple fibers in each derground or undersea. With multiple fibers per link, the netlink of an all-optical network, because the cost of fibers is relatively work can provide a large capacity for good quality of service, low compared with the installation cost. The resulting network can provide a large capacity for good quality of service, future growth, and fault tolerance. If a node has more incoming/outgoing fibers, it requires larger optical switches. Using the current photonic technology, it is difficult to realize large optical switches. Even if they node has L incoming/outgoing links, each link has F fibers and can be realized, they are expensive. To overcome this problem, we each fiber has C channels at wavelengths  $\lambda_1, \lambda_2, \ldots, \lambda_C$ , then can be realized, they are expensive. It is not be realized, they are expensive to this node configuration for all-optical networks. We exploit the this node has N=LF incoming/outgoing fibers. Using the exflexibility that, to establish a lightpath across a node, we can select any one of the available channels in the incoming link and any one of the available channels in the outgoing link. As a result, the switch for each wavelength. Fig. 1 shows an example. When proposed node configuration requires significantly smaller optical switches while it can result in nearly the same blocking probability going fibers and hence it requires larger optical switches. For as the existing one. We demonstrate that a good network design is to adopt the proposed node configuration and slightly more fibers in each link, so that the network requires small optical switches optical switches; if F is increased to 1000, the node requires while it has a small blocking probability.

Index Terms—All-optical networks, blocking probability, node drawbacks. configuration, optical switches.

I. INTRODUCTION

WAVELENGTH-DIVISION multiplexing (WDM) can effectively exploit the enormous bandwidth of an optical fiber [1]. Consequently, WDM networks can support high data rates and provide large network capacity. If a WDM network delivers information in the optical domain within the network, it is known as an all-optical network [2]-[16]. It can avoid many overheads within the network, such as O/E and E/O conversion, processing and buffering.

Before all-optical networks can be used for real-world applications, many issues have to be solved. In the literature, there are studies on routing and wavelength assignment [2]-[7], small optical switches even when the node has many inanalysis of blocking probability [8]-[10], wavelength convercoming/outgoing fibers. We demonstrate that the proposed sion [3], [8]-[10], allocation of wavelength converters [11], node configuration has several advantages: distributed network control [12], design of logical topology [13], and design of physical topology [14]. Recently, there is a growing interest to study all-optical networks with multiple fibers per link [15], [16]. It is cost-effective

to install multiple fibers in every link, because the cost of fibers is relatively low compared with the cost of installing fibers un-Paper approved by W. C. Kwong, the Editor for Optical Communications the IEEE Communications Society. Manuscript activities ications Society. Manuscript received September 12, 2000;

Y.-W. Leung is with the Division of Communication Engineering, School of Electrical and Electronic Engineering, Nanyung Technological University, Singapore.

G. Xiao and K.-W. Hung are with the School of Professional Education and Kong.
Publisher Item Identifier S 0090-6778(02)00516-0.

0090-6778/02\$17.00 © 2002 IEEE

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 35, NO. 5, SEPTEMBER 2005

#### Assignment of Movies to Heterogeneous Video Servers

Yiu-Wing Leung, Senior Member, IEEE, and Ricky Yuen-Tan Hou, Member, IEEE

Server system

Fig. 1. Typical video-on-demand (VOD) system

Centralized server configuration [Fig. 2(a)]—The

server system uses a high-end computer (e.g., a multi-

Distributed server configuration [Fig. 2(b)]—The

server system uses multiple low-end computers (e.g.,

personal computers) as video servers. Each server

stores some of the movies, and it can serve multiple

users concurrently. Overall, the server system can

Good scalability-The system can easily be scaled up

High availability-The system can still provide ser-

vice when some servers fail or are under preventive

Competitive performance-to-price ratio-Personal

Using the distributed server configuration, it is necessary to

· Little and Venkatesh [6] formulated the assignment

problem for a server system with identical servers. Their

objective was to minimize the blocking probability,

which is the probability that a request for VOD service is blocked (rejected) because the server system does

assign each movie to one or more video servers. This assignment

problem has been investigated in the literature [6]-[8].

computers are currently fast and cheap.

Serpanous and Bouloutas [4] made a comprehensive com-

processor computer) as a video server.

serve many users concurrently.

by adding more video servers.

VOD system called iTV system [5].

Abstract-A video-on-demand (VOD) system provides an electronic video rental service to geographically distributed users. It can adopt multiple servers to serve many users concurrently. As a VOD system is being used and evolved, its servers probably become heterogeneous. For example, if a new server is added to expand the VOD system or replace a failed server, the new server may be faster with a larger storage size. This paper investigates how to assign movies to heterogeneous servers in order to minimize the blocking probability. It is proven that this assignment problem is NP-hard, and a lower bound is derived on the minimal blocking probability. The following approach is proposed for assignment:

1) problem relaxation—a relaxed assignment problem is formulated and solved to determine the ideal load that each server should headlest a solve the relaxation. handle, and 2) goal programming—an assignment and reassign-ment are performed iteratively while fulfilling all the constraints so that the load handled by each server is close to the ideal one. This approach is generic and applicable to many assignment prob-lems. This approach is adopted to design two specific algorithms for movie assignment with and without replication. It is demonstrated that these algorithms can find optimal or close-to-optimal

Index Terms-Assignment, server system, video-on-demand

I. INTRODUCTION

A. Background video-on-demand (VOD) system provides an electronic A video rental service to geographically distributed users [1], [2]. Using this service, users can select and watch movies parison between these two configurations. In particular, the disat their convenient time and places, and they may interact with tributed server configuration is attractive in three aspects. the movies using interactive operations such as fast forward, rewind, and pause. VOD is considered to be a promising Internet service [3]. Fig. 1 shows a typical VOD system. A server system stores a collection of movies. When a user requests to watch a movie, the server system retrieves this movie and delivers it to

the user through a communication network. VOD service is usually open to many users and so the server system should be able to serve multiple users concurrently. To The distributed server configuration is adopted by a commercial serve each user, the server system retrieves and delivers video at the video playback rate (e.g., 1.5 Mb/s for MPEG video). Consequently, the server system should have a large capacity B. Relevant Results in the Literature to handle a large volume of bits. To satisfy this requirement, the server system can adopt one of two configurations [4].

Manuscript received February 22, 2002; revised April 23, 2004. This project was supported by the RGC Research Grant HKBU 2092/01E. This paper was recommended by Associate Editor C. Hsu. recommended by Associate Editor C. Hsu.

The authors are with the Department of Computer Science,
Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail:
ywleung@comp.hkbu.edu.hk; rickyhou@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TSMCA.2005.851136

1083-4427/\$20.00 @ 2005 IEEE

#### Design of an Interactive Video-on-Demand System Yiu-Wing Leung, Senior Member, IEEE, and Tony K. C. Chan

Abstract—We design an interactive video-on-demand (VOD) sign can provide an ideal interactive service to the customers, system using both the client-server paradigm and the broad-but it needs dedicated resources (such as I/O bandwidth) to cast delivery paradigm. Between the VOD warehouse and the maintain a video stream for each ongoing customer. For large customers, we adopt a client-server paradigm to provide an interactive service. Within the VOD warehouse, we adopt a broadcast delivery paradigm to support many concurrent customers. In r, we exploit the enormous bandwidth of optical fibers for broadcast delivery, so that the system can provide many video serve more concurrent customers. The main idea is that the program and maintain a small access delay. In addition, we design system waits for a time interval (called batch window) to collect and adopt an interleaved broadcast delivery scheme, so that every video stream only requires a small buffer size for temporary storage. A simple proxy is allocated to each ongoing customer, and eves video from the optical channels and delivers the video to the customer through an information network. The proposed customers simultaneously. However, the customers have to wait VOD system is suitable for large scale applications with many customers, and it has several desirable features: 1) it can be scaled up to serve more concurrent customers and provide more video programs, 2) it provides interactive operations, 3) it only requires inication between the VOD warehou the customer and it does not involve any network control, 4) it has a small access delay, and 5) it requires a small buffer size for each

Index Terms-Broadcast delivery paradigm, client-server program, video-on-demand.

I. INTRODUCTION

N INTERACTIVE video-on-demand (VOD) system pro-A vides an electronic video rental service to geographically distributed customers [1]. It retrieves video programs from its storage and delivers them to the customers through an information network. The customers can select and watch video programs at their convenient time and place, and they can interact with the programs via interactive operations such as pause, fastforward, and rewind.

A VOD system has to serve multiple customers concurrently, and therefore it must have a large enough capacity to provide multiple video streams. Many designs have been proposed for this purpose and they can be classified into two categories: client-server design and broadcasting design.

A. Client-Server Design

The client-server design adopts the client-server paradigm. The system is composed of one or more servers [2], [3]. It maintains a dedicated video stream for each ongoing customer. When adigm [9], [10] to serve many concurrent customers. There are the customer performs an interactive operation, the system retrieves and delivers the corresponding video for him. This de-

Manuscript received February 29, 2000; revised May 30, 2001. The associate editor coordinating the review of this manuscript and approving it for publica-tion was Dr. Thomas R. Gardos. The authors are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: ywleung@comp.hkbu.edu.hk; tkcchan@hkbu.edu.hk).

Digital Object Identifier 10.1109/TMM.2003.808818

The first design is called periodic broadcasting [11]. It broadcasts multiple streams of the same video program at staggered times periodically. To watch a video program, a customer waits until a new video stream for this program is broadcast and then he receives this stream. The system can serve many concurrent customers because many customers can receive the same video stream from a broadcast channel simultaneously. However, it has a long mean access delay or

IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 5, NO. 1, MARCH 2003

scale applications with many customers, this design requires

A client-server design can use a batching policy [4]-[8] to

a batch of requests for a video program. Then the system creates

one video stream for this program and multicasts it to a batch of

customers. In this manner, one video stream can serve multiple

before starting a VOD session (the waiting time is called access

delay) and they cannot perform (or can only perform some con-

strained) interactive operations. Several batching policies have

· Dan et al. [5] proposed that when the system can establish

a new video stream, it selects the batch with the largest

number of waiting customers and creates a video stream to

serve all the customers in this batch. This batching policy

can minimize the mean access delay, but some customers

· Dan et al. [6] proposed to choose a shorter batch window

for the more popular video programs. They developed an

analytical model and determined the window size for each

· Almeroth et al. [7] proposed a batching policy that sup-

ports some constrained interactive operations by buffering

a certain portion of the video program or joining the cus-

· Liao and Li [8] proposed a batching policy called

split-and-merge. When a customer performs an interac-

tive operation, the system splits him from his original

video stream and attempts to create a new video stream

for him. If this is not possible, the customers waits. Once

the interaction is done, the system attempts to merge this

customer back to an existing video stream via buffering.

The broadcasting design adopts the broadcast delivery par-

been proposed in the literature and they are as follows.

may experience long access delay.

tomer to another existing video stream.

If this is not possible, the customer waits.

video program.

B. Broadcasting Design

large amount of resources.

1520-9210/03\$17.00 © 2003 IEEE

Server

and whether a server mechanism will change the self-similarity

In [13] and [14], Tsybakov and Georganas point out that

the superposition of two uncorrelated self-similar processes

retain some asymptotic self-similarity property. Vamvakos and Anantharam [15] consider a special case of a leaky bucket

system with long-range-dependent input traffic, and prove that the output (departure) process is also long-range dependent.

In our previous work [12], we take a first look at the traffic

characteristics of the output process. In this letter, we carry on

with our previous work and focus on the superposition of self-

similar processes, and further explore the property of the output

is organized as follows. Section II discusses the concepts of self-

similar processes and presents the self-similar definitions and

superposition of two or more uncorrelated strong asymptotically

self-similar processes (or long-range-dependent processes) is

strong asymptotically self-similar (or long-range dependent).

Since traffic arrival to a switch is multiplexed from many

connections, this superposition property is very important for

the analysis of queueing systems. In Section III, we also discuss

Section IV considers a model of a single server with infi-

queue-length process is finite, the strong asymptotically second-

order self-similar (sas-s) properties of the input process and

that of the output process are equivalent, which means that the

Is Xout self-similar?

#### Some Results on the Self-Similarity Property in Communication Networks Shibin Song, Joseph Kee-Yin Ng, Senior Member, IEEE, and Bihai Tang

Is X' + X" self-similar?

Abstract—Due to the strong experimental evidence that packet network traffic is self-similar in nature, it is important to study the problems to see whether the superposition of self-similar processes retains the property of self-similarity, and whether the service of a server changes the self-similarity property of the input traffic. In this letter, we first discuss some definitions and superposition properties of self-similar processes. We obtain some good results about the property of merging self-similar data streams. Then we present a model of a single server with infinite buffer and prove that when the queue length has finite second-order moment, the input process, being strong asymptotically second-order self-similar (sas-s), is equivalent to the output process which also bears the

Index Terms-Long-range dependent, packet networks, self-similar, short-range dependent.

#### I. INTRODUCTION

CEVERAL empirical studies on the local-area network (LAN), the variable bit rate (VBR) video traffic, the Integrated Services Digital Network (ISDN), and other communication systems indicate that this traffic is self-similar in nature. For instance, Leland et al. [7] have demonstrated the self-similar nature of Ethernet traffic by a statistical analysis process from a server with self-similar input. The rest of the letter of the Ethernet traffic measurements at Bell-Core; Beran et al. [2], [3], [5] have demonstrated long-range dependence their relationships. In Section III, we present the superposition in samples of VBR video traffic generated by a number of different codecs; and Paxson and Floyd [11] have concluded property of self-similar processes. We obtained the result that the the presence of long-range dependence in TELNET and other wide-area network traffic.

In the light of this strong experimental evidence, it is important to examine in more detail the possible implications that selfsimilar traffic may have on the design and performance of network systems. For example, real-time communications require the network to provide an end-to-end delay guarantee. In order the superposition of two correlated self-similar processes, and to analyze the delay of networks with self-similar traffic, we need to know the property of queueing systems with self-similar input traffic. In particular, there are two important questions we need to study (as shown in Fig. 1): whether the superposition of self-similar processes retains the self-similarity properties,

Paper approved by M. Hamdi, the Editor for Network Architecture of the IEEE Communications Society. Manuscript received May 9, 2002; revised December 3, 2003. This work was supported in part by the Research Grants Council under Earmarked Research Grant RGC/97-98/54, and by the Faculty Research Grant program under FRG/97-98/II-76 and FRG/98-99/II-68. S. Song is with the Department of Risk Management and Insurance, Lingnan College, Zhongshan University, GuangZhou 510275, China (e-mail:

J. K.-Y. Ng is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e.mail: jng@comp.hkbu.edu.hk).

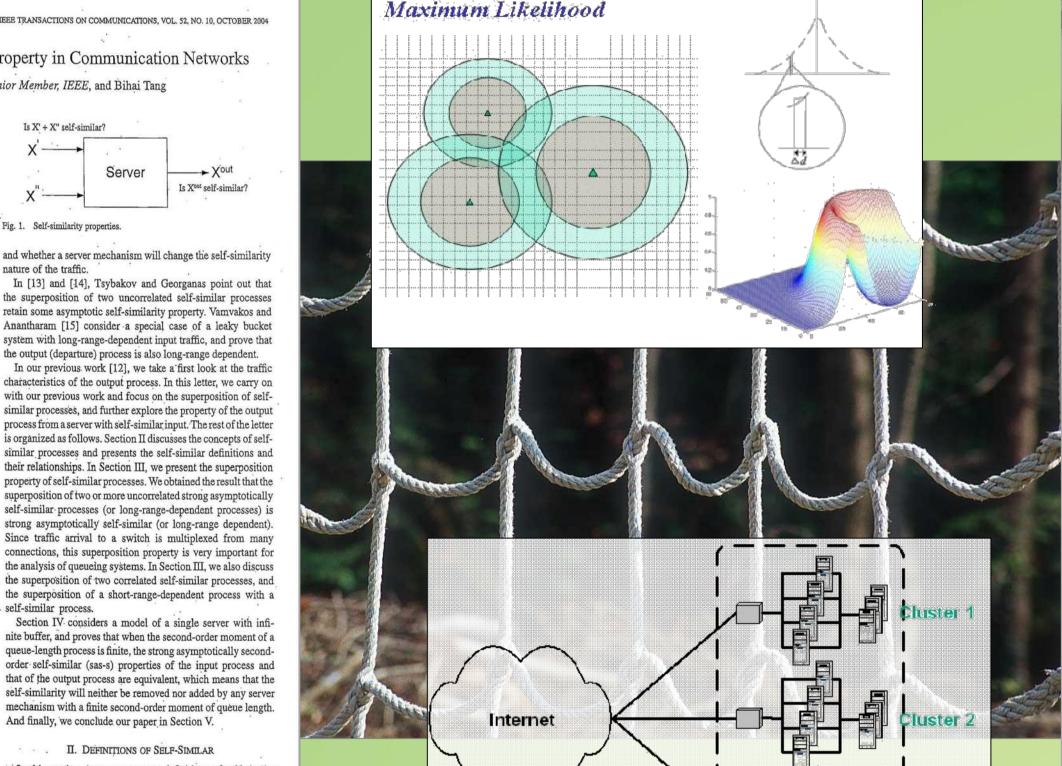
B. Tang is with the Department of Quantitative Economics, Guangzhou Academy of Economics, GuangZhou 510600, China (e-mail: tb-

Digital Object Identifier 10.1109/TCOMM.2004.833136

self-similarity will neither be removed nor added by any server And finally, we conclude our paper in Section V.

II. DEFINITIONS OF SELF-SIMILAR In this section, we present some definitions of self-similar processes which are based on a second-order-stationary realnumber stochastic process. We begin with the introduction of  $X = (X_1, X_2, ...)$ , a semiinfinite segment of a second-order-stationary real-number sto-

chastic process of discrete argument (time)  $t \in I_1 \triangleq \{1, 2, \ldots\}$ . 0090-6778/04\$20.00 © 2004 IEEE



Real-Time Systems, 30, 217-256, 2005 © 2005 Springer Science + Business Media, Inc.. Manufactured in The Netherlands. DOI: 10.1007/S11241-005-1401-1

#### A OoS-Enabled Transmission Scheme for MPEG Video Streaming\*

jng@comp.hkbu.edu.hk Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong

kleung@computer.org Department of Information & Communications Technology, Hong Kong Institute of Vocational Education (Tsing Yi), Tsing Yi Island, Hong Kong

CALVIN KIN-CHEUNG HUI kchui@comp.hkbu.edu.hk Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong

#### Published online: 9 June 2005

Abstract. While MPEG is the de facto encoding standard for video services, online video streaming service is becoming popular over the open network such as the Internet. As the performance of open network is nonpredictable and uncontrollable, the tuning of the quality of service (QoS) for on-line video streaming services is difficult. In order to provide better QoS for the delivery of videos, there are proposals of new encoding formats or new transmission protocols for on-line video streaming. However, these results are not compatible with popular video players or network protocols and hence these approaches are so far not very successful. We use another approach which tries to by-pass these problems. We designed a QoS Tuning Scheme and a QoS-Enabled Transmission Scheme for transmitting MPEG videos from video servers to clients. According to the traffic characteristics between the video server and each individual client, the QoS Tuning Scheme tunes the QoS to be delivered to each individual client on the fly, Furthermore, our QoS-Enabled Transmission Scheme can be applied over any protocol, such as HTTP which is the most popular protocol over the open network. With our transmission scheme, bandwidth can be better utilized by reducing transmitted frames which would have missed their deadlines and would eventually be discarded by the clients. This is achieved by sending frames according to their impact on the QoS in the playback under the allowed throughput. With these schemes, users can enjoy video streaming through their favorite video players and with the best possible QoS. In order to facilitate the real time QoS tuning, a metric, QoS-GFS, is developed. This QoS-GFS is extended from the QoS-Index, another metric which has taken human perspective in the measurement of video quality. Hence QoS-GFS is better than the common metrics which measures QoS by means of rate of transmission of bytes or MPEG frames. We designed and implemented a middleware to perform empirical tests of the proposed transmission scheme and QoS tuning scheme. Experiment results show that our schemes can effectively enhance the QoS for online MPEG video streaming services.

Keywords.: MPEG video streaming, transmission scheme, quality of services, QoS control, video on demand

#### 1. Introduction

Online video streaming services has become a popular services in the open network. Video streaming is the concurrent processes of sending video images from servers to clients over

\*The work reported in this paper was supported in part by the RGC Earmarked Research Grant under RGC HKBU 2074/01E, and by the FRG under FRG 00-01/I.

Real-Time Syst (2006) 34:83-99 DOI 10.1007/s11241-006-7982-5

#### Scheduling real-time requests in on-demand data broadcast environments\*,†

Victor C. S. Lee · Xiao Wu · Joseph Kee-Yin Ng

Published online: 15 May 2006 © Springer Science + Business Media, LLC 2006

Abstract On-demand broadcast is an attractive data dissemination method for mobile and wireless computing. In this paper, we propose a new online preemptive scheduling algorithm, called PRDS that incorporates urgency, data size and number of pending requests for real-time on-demand broadcast system. Furthermore, we use pyramid preemption to optimize performance and reduce overhead. A series of simulation experiments have been performed to evaluate the real-time performance of our algorithm as compared with other previously proposed methods. The experimental results show that our algorithm substantially outperforms other algorithms over a wide range of workloads and parameter settings.

Keywords Real-time scheduling algorithm · On-demand broadcast · Preemption · Mobile

#### 1. Introduction

With the increasing availability of high bandwidth links as well as the popularity of portable wireless devices, such as notebook computers and PDA, mobile and wireless computing

\*The work described in this paper was partially supported by grants from CityU (Project No. 7001841) and RGC CERG Grant No. HKBU 2174/03E.

<sup>†</sup>This paper is an extended version of the paper "A preemptive scheduling algorithm for wireless real-time on-demand data broadcast" that appeared in the 11th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications.

Springer 2

Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong e-mail: csvlee@cityu.edu.hk

e-mail: wuxiao@cs.citvu.edu.hk

Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong e-mail: jng@comp.hkbu.edu.hk

#### Location Estimation via Support Vector Regression

Zhi-li Wu, Chun-hung Li, Member, IEEE, Joseph Kee-Yin Ng, Senior Member, IEEE, and Karl R.P.H. Leung, Senior Member, IEEE

Abstract—Location estimation using the Global System for Mobile communication (GSM) is an emerging application that infers the location of the mobile receiver from multiple signals measurements. While geometrical and signal propagation models have been deployed to tackle this estimation problem, the terrain factors and power fluctuations have confined the accuracy of such estimation. Using support vector regression, we investigate the missing value location estimation problem by providing theoretical and empirical analysis on existing and novel kernels. A novel synthetic experiment is designed to compare the performances of different location estimation approaches. The proposed support vector regression approach shows promising performances, especially in terrains with local variations in environmental factors.

Index Terms-Location estimation, support vector regression, statistical estimation, Global System for Mobile communication.

#### INTRODUCTION TO GSM POSITIONING

global positioning system (GPS) has been in service for many years, its wide-spread application has been limited by its reliance on special hardware. Furthermore, in metropolitan areas, the access to GPS signals is often limited.

On the other hand, location estimation technology based on GSM technology has advanced rapidly in recent years. In 1996, the Federal Communications Commission (FCC) directed the rules that wireless service companies should provide location identification for wireless emergency calls, and the accuracy of location estimation should be within 125 meters [2]. Since 2000, the FCC has stipulated separate requirements for both terminal-based and network-based positioning methods, and the FCC rules revised that 67 percent of estimations should be less than 50 meters and 100 meters in terminal-based and network-based solutions, respectively [6]. A terminal-based positioning solution based on modifications of handsets, SIM cards, or both, means that subscribers have to upgrade their mobile phones or renew SIM cards to enable this location service. By contrast, a network-based solution estimates the location by upgrading the operator network instead of the handset; therefore, it can benefit all mobile phone subscribers. In July of 2002, these wireless Enhanced 911 rules were revised again to improve the effectiveness and reliability of 911 service by reporting the telephone number

Z.-l. Wu, C.-h. Li, and J.K.-Y. Ng are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong. E-mail: (vincent, chli, jng)@comp.hkbu.edu.hk. K.R.P.H. Leung is with the Department of Information and Communica-tions Technology, Hong Kong Institute of Vocational Education, 20 Tsing

Yi Road, Tsing Yi Island, Hong Kong. É-mail: kleung@computer.org. Manuscript received 8 June 2004; revised 6 June 2005; accepted 12 Sept. 2005; published online 16 Jan. 2007. For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-0185-0604.

OCATION estimation service is very important for mobile of a wireless 911 caller and the location of the cellular Land ubiquitous computing. Its range of applications phone within 50 to 100 meters in most cases [4]. Hence, includes logistics, field worker deployment, resource there have been a lot of investigations on location management, personal safety, and tourism. Although the estimations for satisfying the rules of Enhanced 911 from

In general, the terminal-based method can provide a more accurate estimate than the network-based method, which is also reflected in earlier FCC rules. However, it requires the modification of the handsets. From an engineering point of view, it is not easy to enable the functionality of location estimates by modifying handsets; even GPS technology, which is the most mature and desirable system to be introduced to assist mobile positioning, is also inaccurate for populated or indoor areas. Moreover, it is an engineering concern to upgrade millions of handsets already in operation. Thus, there is a need to develop location estimation algorithms for unmodified handsets. There are several network-based technologies available

for cellular phone positioning. These methods utilize the signal attenuation, angle of arrival (AOA), time of arrival (TOA), time difference of arrival (TDOA), enhanced observed time difference (E-OTD), and time advance (TA) to locate unmodified cellular phones. Generally, all of these methods are based on knowing some location-related information of reference points and then relating them to the location of the mobile station. For example, the time for the signals to be returned from a mobile device to a base station is known in the TDMA system. This information can be the distance measurement of the mobile device to the base station, which can be further exploited for location estimation.

But, for the GSM network, to implement time-involved estimate algorithms such as TOA, TDOA, and E-OTD, expensive clocks for precise time synchronization should be installed. Cellocate is an example of a TDOA-driven system that uses GPS time synchronization. Their field tests, based on four base stations, resulted in an accuracy of between 187 and 287 m [7]. Even though a GPS receiver can be used for synchronization purposes, extra hardware would add a heavy burden in terms of installation costs for the network operator.

1536-1233/07/\$25.00 © 2007 IEEE Published by the IEEE CS, CASS, ComSoc, IES, & SPS

IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 8, NO. 5, OCTOBER 2006

#### Fine-Grained Scalable Video Caching for Heterogeneous Clients

Jiangchuan Liu, Member, IEEE, Jianliang Xu, Member, IEEE, and Xiaowen Chu, Member, IEEE

Abstract—Much research has focused on caching adaptive To address this problem, many partial caching algorithms

are not adjustable.

videos to improve system performance for heterogeneous clients with diverse access bandwidths. However, existing rate-adaptive caching systems, which are based on layered coding or transcoding, often suffer from a coarse adaptation and/or a high computation is stored at the proxy, the network resource requirement can be significantly reduced. Most of these proposals however. overhead. In this paper, we propose an innovative rate-adap-tive caching framework that enables low-cost and fine-grained be significantly reduced. Most of these proposals, however, assumed that the streaming rate (either constant or variable) adaptation by using MPEG-4 fine-grained scalable videos. The of the video is predetermined. Because of this nonadaptability, proposed framework is both network-aware and media-adaptive; they suffer from two limitations: first, it is difficult to meet the i.e., the clients can be of heterogeneous streaming rates, and the backbone bandwidth consumption can be adaptively controlled. We develop efficient cache management schemes to determine the best contents to cache and the optimal streaming rate to each the best contents to cache and the optimal streaming rate to each client under the framework. We demonstrate via simulations that, control the backbone (server-to-proxy) bandwidth consumpcompared to nonadaptive caching, the proposed framework with the optimal cache management not only achieves a significant are not adjustable. reduction in the data transmission cost, but also enables a flexible utility assignment for the heterogeneous clients. Our results also show that the framework maintains a low computational overhead, which implies that it is practically deployable. Index Terms-Fine-grained scalable video, proxy caching, re-

source allocation, streaming media.

I. INTRODUCTION

WING to the increasing demand for video distribution over the Internet, caching video objects at proxies close to clients has attracted a lot of attention in recent years [9]. However, video objects have several distinct features which make conventional Web caching techniques inefficient, if not entirely inapplicable. In particular, a video object usually has a high data rate and a long playback duration, which together yield a very trol by partitioning the video stream at specific rates [7]. The high data volume. For example, a 1-h MPEG-1 video has a data volume of about 675 MB; caching the video in its entirety is clearly impractical, as several objects of such a size would exhaust the capacity of a typical cache.

version of this paper was published in IEEE INFOCOM'04. The work of J. Liu was supported in part by a Canadian NSERC Discovery Grant 288325, an NSERC Research Tools and Instruments Grant, a Canada Foundation for Innovation (CFI) New Opportunities Grant, and an SFU President's Research Grant. The work of J. Xu work was supported in part by grants from the Research Grants Council of the Hong Kong SAR, China (Projects HKBU 2115/05e and HKBU FRC/03-04/II-19). The work of X. Chu was supported by RGC HKBU2159/04E and HKBU210605. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Pascal

Manuscript received January 12, 2005; revised November 4, 2005. An earlier

J. Liu is with School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6 Canada (e-mail: jcliu@cs.sfu.ca). J. Xu and X. Chu are with Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: xujl@comp.hkbu.edu.hk; chxw@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TMM.2006.879859

bandwidth consumption

proposed framework. We advocate a semistatic caching paradigm [9], [12], and there are two dimensions to explore of how to cache an FGS video: the length and the rate of the portion to be cached. We stress that caching decisions must take into account the interactivities in video playback; i.e., nonuniform accesses of different portions. Moreover, when a cached video is delivered to a client, different streaming rates can be selected as long as the rate is no higher than the client's bandwidth. Consequently, the proxy cache management becomes considerably more complex than that for a nonadaptive video based system. Efficient solutions are developed to optimize the resource uti-

diverse bandwidth conditions of heterogeneous clients, as a single streaming rate would either overuse or underuse some

One method of overcoming these two limitations is to use

rate-adaptive videos. However, this poses significant challenges

to caching. Most conventional rate adaptation mechanisms are

executed during the encoding process (e.g., adjusting quantizers

[8], [22]) and, hence, are difficult to apply to cached videos.

There have been research efforts to combine proxy caching

with video layering or transcoding [10], [14], [18], [20], but

these adaptive systems suffer from either coarse adaptation

granularity (due to the inflexible structures of existing layered

coders) or a high computation overhead (due to the transcoding

In this paper, we propose a novel video caching framework

to achieve low-cost and fine-grained rate adaptation. The frame-

work employs the MPEG-4 fine-grained scalable (FGS) video

with bit-plane coding, which enables post-encoding rate con-

post-encoding rate control operations can be efficiently imple-

mented at the server or the proxy, resulting in a low computa-

network aware and media adaptive: clients can be of heteroge-

neous access bandwidths, and adaptive FGS videos are used to

meet the clients' bandwidth conditions and control the backbone

We examine the critical design and management issues in the

1520-9210/\$20.00 © 2006 IEEE

# Networking and Multimedia

IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. X, NO. X, MONTH 200X

#### Optimizing Lifetime for Continuous Data Aggregation with Precision Guarantees

in Wireless Sensor Networks

Xueyan Tang, Member, IEEE, and Jianliang Xu, Member, IEEE

Abstract-This paper exploits the tradeoff between data quality and energy consumption to extend the lifetime of wireless sensor networks. To obtain an aggregate form of sensor data with precision guarantees, the precision constraint is partitioned and allocated to individual sensor nodes in a coordinated fashion. Our key idea is to differentiate the precisions of data collected from different sensor nodes to balance their energy consumption. Three factors affecting the lifetime of sensor nodes are identified: (1) the changing pattern of sensor readings; (2) the residual energy of sensor nodes; and (3) the communication cost between precision allocation in terms of network lifetime and propose an adaptive scheme that dynamically adjusts the precision deration the topological relations among sensor nodes and the effect of in-network aggregation. Experimental results improves network lifetime compared to existing methods. Index Terms-data aggregation, data accuracy, energy effi-

I. INTRODUCTION

exchange data with applications to accomplish their missions. of 1°C." In this way, the sensor nodes do not have to report While the base station can have continuous power supply, all readings to the base station. Only the updates necessary to the sensor nodes are usually battery-powered. The batteries are guarantee the desired level of precision need to be sent.



value of a sensor network is determined by the time duration constraints at the sensor nodes. The adaptive scheme also takes before it fails to carry out the mission due to insufficient number of "alive" sensor nodes. This duration is referred to as the network lifetime [1]. It is both mission-critical and economically desirable to manage sensor data in an energyefficient way to extend the lifetime of sensor networks.

The data captured by the sensor nodes are often converted

into an aggregate form requested by the applications (e.g.,

average temperature reading). Primarily designed for monitoring purposes, many sensor applications require continuous WIRELESS sensor networks are used in a wide range of aggregation of sensor data [3]. Exact data aggregation requires substantial energy consumption because each sensor node has ronmental data [1], [2]. A wireless sensor network typically to report every reading to the base station. In wireless sensor consists of a base station and a group of sensor nodes (see networks, communication is a dominant source of energy Figure 1). The sensor nodes are responsible for continuously consumption [4], [5]. To save energy, data semantics can be sampling physical phenomena such as temperature and hu-relaxed to allow approximate data aggregation with precision midity. They are also capable of communicating with each guarantees [6], [7], [8], [9]. The precision can, for example, other and the base station through radios. The base station, on be specified in the form of quantitative error bounds: "average the other hand, serves as a gateway for the sensor network to temperature reading of all sensor nodes within an error bound

inconvenient and sometimes even impossible to replace. When It is, however, a challenging task to optimize network lifea sensor node runs out of energy, its coverage is lost. The time under approximate data aggregation because the sensor mission of a sensor application would not be able to continue nodes are inherently heterogeneous in energy consumption. if the coverage loss is remarkable. Therefore, the practical First, when the data captured by different sensor nodes change at different magnitudes and frequencies, the sensor nodes may Manuscript received October 24, 2006; revised February 23, 2007; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor N. Shroff. This work was supported in part by a grant from Nanyang Technological University (Project No. RC47/06). Jianliang Xu's work was supported in part by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. HKBU 2115/05E). A preliminary report of this work was presented at IEEE INFOCOM'2006. X. Tang is with the School of Computer Engineering, Nanyang Technodes report data at the same rate, their energy consumption nological University, Nanyang Avenue, Singapore 639798 (e-mail: asxy-can be highly unbalanced, thereby reducing network lifetime. In addition to reporting local sensor readings, the intermediate University, Kowloon Tong, Hong Kong (e-mail: xujl@comp.hkbu.edu.hk).

In addition to reporting local sensor readings, the intermediate nodes in a multi-hop network are also responsible for relaying

IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 3, MARCH 2006

#### An Error-Resilient and Tunable Distributed Indexing Scheme for Wireless Data Broadcast

Jianliang Xu, Member, IEEE, Wang-Chien Lee, Member, IEEE, Xueyan Tang, Member, IEEE, Qing Gao, and Shanping Li

Abstract—Access efficiency and energy conservation are two critical performance concerns in a wireless data broadcast system. We propose in this paper a novel parameterized index called the exponential index that has a linear yet distributed structure for wireless data broadcast. Based on two tuning knobs, index base and chunk size, the exponential index can be tuned to optimize the access latency with the tuning time bounded by a given limit, and vice versa. The client access algorithm for the exponential index under unreliable broadcast is described. A performance analysis of the exponential index is provided. Extensive ns-2-based simulation experiments are conducted to evaluate the performance under various link error probabilities. Simulation results show that the exponential index substantially outperforms the state-of-the-art indexes. In particular, it is more resilient to link errors and achieves more performance advantages from index caching. The results also demonstrate its great flexibility in trading access latency with

1041-4347/06/\$20.00 © 2006 IEEE Published by the IEEE Computer Society

Index Terms-Index structure, data broadcast, energy conservation, mobile computing.

Wireless data broadcast has received a lot of attention limited battery capacity on mobile clients [25]. To facilitate from industries and academia in recent years. It has been available as commercial products for many years (e.g., two operation modes: active mode and doze mode. The device StarBand [20] and Hughes Network [21]). In particular, the normally operates in the active mode; it can switch to the recent announcement of the smart personal objects technology (SPOT) by Microsoft [16] further highlights the industrial interest in and feasibility of utilizing broadcast for wireless data services. With a continuous broadcast network (called DirectBand Network) using FM radio subcarrier frequen-been used to measure access efficiency and energy concies, SPOT-based devices (e.g., PDAs and watches) can servation, respectively [6], [9], [10]: continuously receive timely information such as stock quotes, airline schedules, local news, weather, and traffic

Access efficiency and energy conservation are two main performance issues for the clients in a wireless data broadcast system. Access efficiency concerns how fast a request is satisfied, and energy conservation concerns how To retrieve a data item from wireless data broadcast, to reduce a mobile client's energy consumption when it the mobile client has to continuously monitor the broadaccesses the data of interest. While access efficiency is a cast until the data arrives. This will consume a lot of constantly tackled issue in most system and database energy since the client has to remain active during its

- J. Xu and Q. Gao are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, KLN, Hong Kong.
   E-mail: [xull, agao]@comp.hkbu.edu.hk.
   W.-C. Lee is with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802.
- E-mai: Weevest-psu-teut.
  X. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.

Manuscript received 6 Jan. 2005; revised 6 June 2005; accepted 8 Sept. 2005; fically, to start an index search, the client needs to wait For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0007-0105.

Access latency. The time elapsed between the moment when a query is issued and the moment when it is responded. Tuning time: The amount of time a mobile client

doze mode to save energy when the system becomes idle.

For example, a typical wireless PC card, ORINOCO,

consumes 60 mW during the doze mode and 805-1,400 mW

during the active mode [25]. In the literature, two perfor-

mance metrics, namely, access latency and tuning time, have

stays active to receive the requested data. research, energy conservation is very critical due to the waiting time. A solution to this problem is air indexing. The basic idea is to include some index information about the arrival times of data items on the broadcast channel. By accessing the index, the mobile client is able to predict the arrival of its desired data. Thus, it can stay in the doze mode during waiting time and tune into the broadcast channel only when the data item of interest arrives. Several traditional disk-based indexing techniques Technological University, Nanyang Avenue, Singapore 605755.

E-mail: assyrtang@ntu.edu.sg.

S. Li is with the College of Computer Science, Zhejiang University,

centralized tree structures, which are not performance efficient for the sequential-access broadcast media. SpeciIEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 6, JUNE 2007 Top-k Monitoring in Wireless Sensor Networks

Minji Wu, Student Member, IEEE, Jianliang Xu, Member, IEEE, Xueyan Tang, Member, IEEE, and Wang-Chien Lee, Member, IEEE

Abstract—Top-k monitoring is important to many wireless sensor applications. This paper exploits the semantics of top-k query and proposes an energy-efficient monitoring approach called FILA. The basic idea is to install a filter at each sensor node to suppress unnecessary sensor updates. Filter setting and query reevaluation upon updates are two fundamental issues to the correctness and efficiency of the FILA approach. We develop a query reevaluation algorithm that is capable of handling concurrent sensor updates. In particular, we present optimization techniques to reduce the probing cost. We design a skewed filter setting scheme, which aims to balance energy consumption and prolong network lifetime. Moreover, two filter update strategies, namely, eager and lazy, are proposed to favor different application scenarios. We also extend the algorithms to several variants of top-k query, that is, orderinsensitive, approximate, and value monitoring. The performance of the proposed FILA approach is extensively evaluated using real data traces. The results show that FILA substantially outperforms the existing TAG-based approach and range caching approach in terms of both network lifetime and energy consumption under various network configurations.

Index Terms—Sensor network, data management, energy efficiency, top-k, continuous query.

#### 1 INTRODUCTION

 $\mathbf{R}$ ECENT advances in signal processing, microelectronics, and wireless communications have enabled the deployment of large-scale sensor networks for many applications such as habitat and environment monitoring [28]. A wireless sensor network typically consists of a base station and a group of sensor nodes (see Fig. 1). The base station serves as a gateway for the sensor network to exchange data with external users. The sensor nodes, on the other hand, are responsible for sensing and collecting data from their local environments. They are also capable of processing sensed data and communicating with their neighbors and

Monitoring aggregate forms of sensed data is important to many sensor applications and has drawn a lot of research attention [6], [7], [16], [19], [29], [30]. Among those aggregates, a top-k query requests the list of k sensor nodes

- with the highest (or lowest) readings. For example: • Environmental Monitoring. Consider an environment-monitoring sensor network. A top-k query is issued to find out the nodes and their corresponding areas with the highest pollution indexes for the purpose of pollution control or research study.
- Network Management. Power supply is critical to
- M. Wu and J. Xu are with the Department of Computer Science, Hong Kong Buptist University, Kowloon Tong, KLN, Hong Kong.

  E-mail: talexwu, xujl/@comp.hkbu.edu.hk.

  X. Tang is with the School of Computer Engineering, Nanyang Technological University, Narvang Assault Science 630708. X. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.
   E-mail: asxylang@ntu.edu.eg.
   W.-C. Lee is with the Department of Computer Science and Engineering, The Penn State University, University Park, PA 16802.

E-mail: wlee@cse.psu.edu. Manuscript received 3 Mar. 2006; revised 13 Oct. 2006; accepted 11 Jan. 2007; published online 7 Mar. 2007. For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0132-0306, Digital Object Identifier no. 10.1109/TKDE.2007.1038.

use a centralized approach in which all sensor readings are periodically collected by the base station, which then computes the top-k result set. To reduce network traffic in data collection, an in-network aggregation technique, known

A naive implementation of monitoring top-k query is to

extend network lifetime.

top-k query may be issued to continuously monitor

the sensor nodes with the least residual energy so

that these sensor nodes can be instructed to adapt

themselves (for example, reducing sampling rates) to

as TAG, has been proposed [16]. In this approach, a routing tree rooted at the base station is first established, and the data are then aggregated and collected along the routing tree to the base station. Consider an example shown in Fig. 2a, where sensor nodes A, B, and C form a routing tree. The readings of these sensor nodes at three successive sampling instances  $t_1$ ,  $t_2$ , and  $t_3$  are shown in the tables in Fig. 2a. Suppose we are monitoring a top-1 query. Employing TAG, at each sampling instance, nodes B and C send their current readings to the parent (that is, node A), which compares the data received with its own reading and sends the highest reading to the base station. In this example, a total of nine messages are sent. Node B is involved in the the operation of a wireless sensor network. Thus, a message exchange at each sampling instance though the top-1 result is always node C. Therefore, this approach incurs unnecessary updates in the network and is not

In this paper, we exploit the semantics of top-k query and propose a novel filter-based monitoring approach called FILA. The basic idea is to install a filter at each sensor node to suppress unnecessary sensor updates. The base station also keeps a copy of the filter setting to maintain an (error bounded) approximate view of each node's reading. A sensor node updates its reading with the base station only when the reading passes the filter. The correctness of the top-k result is ensured if all sensor nodes perform updates according to their filters. Fig. 2b shows an 1041-4347/07/\$25.00 @ 2007 EEE Published by the IEEE Computer Society

IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 17, NO. 1, JANUARY 2006

#### Time-Critical On-Demand Data Broadcast: Algorithms, Analysis, and Performance Evaluation

Jianliang Xu, Member, IEEE, Xueyan Tang, Member, IEEE, and Wang-Chien Lee, Member, IEEE

Abstract—On-demand broadcast is an effective wireless data dissemination technique to enhance system scalability and deal with dynamic user access patterns. With the rapid growth of time-critical information services in emerging applications, there is an increasing need for the system to support timely data dissemination. This paper investigates online scheduling algorithms for time critical on-demand data broadcast. We propose a novel scheduling algorithm called SIN- $\alpha$  that takes the urgency and number of outstanding requests into consideration. An efficient implementation of SIN- $\alpha$  is presented. We also analyze the theoretical bound of request drop rate when the request arrival rate rises toward infinity. Trace-driven experiments show that  $SIN-\alpha$  significantly outperforms existing algorithms over a wide range of workloads and approaches the analytical bound at high request rates.

Index Terms-Mobile computing, on-demand data broadcast, scheduling, content delivery, time constraint.

#### 1 INTRODUCTION

rise to vast research efforts on improving the performance for certain applications (e.g., a small set of data items with of Web accesses. As the system scale and user base continue stable access patterns), on-demand broadcast is more to grow, there is an increasing demand for information providers to be capable of concurrently delivering a large like that in the Internet. amount of information to a huge number of users, especially in popular events such as elections and Olympics services and business-oriented applications, there is an games. As a result, innovative delivery technologies, ncluding satellite communications (e.g., StarBand [26] and DIRECWAY [27]), cable networks, and wireless networks (e.g., 2.5G and 3G), have been developed and deployed to provide shared broadband Internet accesses. Different from traditional networks, a distinguished feature of these new technologies is that they naturally support broadcast. In contrast to unicast, where a data item of interest to multiple clients must be sent individually to each wireless location-based services [18], the queried informaclient, broadcast satisfies all outstanding requests for the tion (e.g., the local theaters) is valid only within a local area. same item by a single transmission. This leads to a more When the mobile user moves away from the area, the efficient use of shared bandwidth, hence improving the information becomes invalid. In addition, a service level system throughput and user-perceived response time [19], [30]. In general, there are two data broadcast approaches [8], [20]:1 Push-based broadcast computes the broadcast program based on historical access statistics; on-demand broadcast

 $1.\ A$  third approach is hybrid broadcast that combines on-demand broadcast with push-based broadcast.

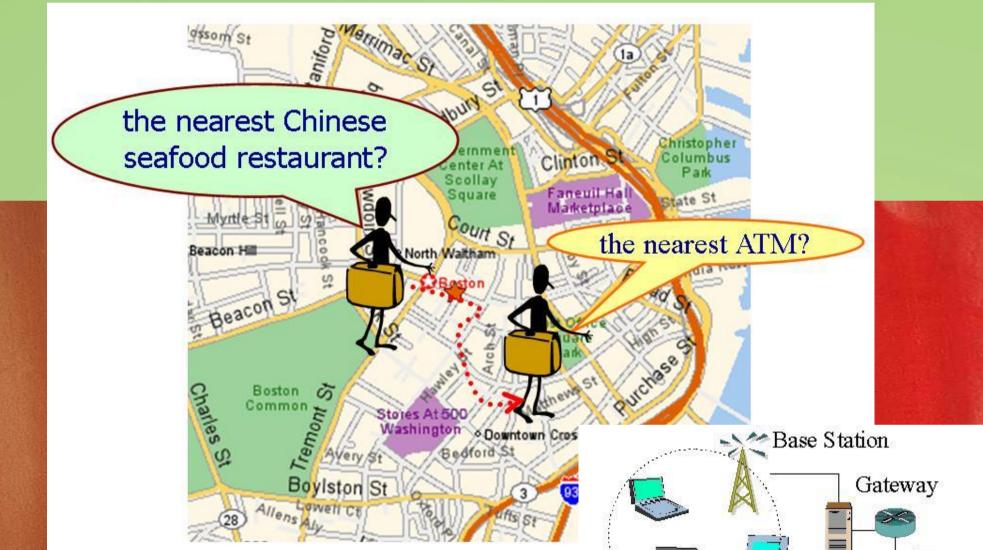
- J. Xu is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, KLN, Hong Kong. E-mail: xuil@comp.hkbu.edu.hk. 3-тан: хиривсотр.пкон.си.пк. X. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798.
- E-mail: asxytang@nti.edu.sg.
  W.-C. Lee is with the Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802.

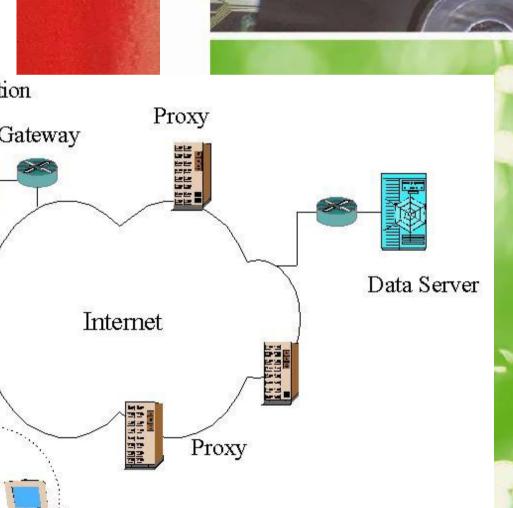
E-mail: wlee@cse.psu.edu. Manuscript received 2 Nov. 2004; revised 15 Mar. 2005; accepted 26 Apr.

resultant slow responses perceived by users have given outstanding requests. While push-based broadcast is useful

With the rapid growth of time-critical information increasing demand to support quality of service (QoS) in content distribution [15], [24], [28]. In many situations, user requests are associated with time constraints as a measure of QoS. These constraints can be imposed either by the users or the applications. For example, in wireless financial services, many users are interested in the up-to-minute (or even "second") stock quotes in order to react to dynamic and rapid market developments. As another example, in agreement (SLA) between a content/service provider and its users usually specifies the desired performance for Web requests, e.g., the response time of requests for CNN.com should not exceed 5 seconds [24]. In all the above cases, a deadline is associated with each request beyond which the serving of the request is useless (or less useful).

This paper focuses on on-demand data broadcast with time constraints, which we shall refer to as time-critical ondemand broadcast. A key issue in the design of an ondemand data broadcast system is the scheduling algorithm used to select and broadcast requested items from outstanding requests. While there has been significant work on developing on-demand broadcast scheduling algorithms (e.g., [1], [2], [10], [29]), none of them has considered the time constraints associated with requests. On the other hand, although some time-critical scheduling algorithms have been proposed for unicast-based real-time systems For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number TPDS-0268-1104. 1045-9219/06/\$20.00 © 2006 IEEE Published by the IEEE Computer Societ





# Gateway

#### eCX: a secure infrastructure for e-course delivery

Joe C.K. Yau Lucas C.K. Hui Bruce Cheung and

#### The authors

Joe C.K. Yau is a Software Engineer, Lucas C.K. Hui is Associate Professor and S.M. Yiu is a Teaching Consultant, all at the Department of Computer Science and Information Systems, The University of Hong Kong, Hong Kong, Bruce Cheung is a Lecturer at the School of Professional and Continuing Education, The University of Hong Kong,

#### Keywords

Security, Copyright, Applications software, Learning organizations, Distance learning, Online computing

#### Abstract

organizations offer online distance learning courses, the more and more critical, especially when the organizations and Cheung and Hui (1999). In particular, rely on the registration fees of students to maintain the smooth running of the courses. Provides a mechanism, the Secure e-Course eXchange (eCX), to protect the learning easily share the account with non-registered material from unauthorized dissemination, and shows how students. this mechanism can be integrated in the operation model of online learning course providers. The design of eCX is general enough to fit two operating models, namely the Institutional Server Model and the Corporate Server Model.

#### Electronic access

http://www.emeraldinsight.com/researchregister The current issue and full text archive of this journal is available at http://www.emeraldinsight.com/1066-2243.htm

Internet Research: Electronic Networking Applications and Policy Volume 13 · Number 2 · 2003 · pp. 116-125 MCB UP Limited · ISSN 1066-2243 DOI 10.1108/10662240310469060

Online education becomes one of the most important channels for students to acquire knowledge and learning material. Currently, there are about three types of courses provided through this education medium: (1) short courses without formal qualifications; (2) diploma, degree or even higher degree

courses with formal qualifications; and (3) courses which help students to take public examinations or to get a formal degree such as an external London University degree. Although many projects and researches have

been conducted on online distance learning, the issues of security have only been studied recently (Cheung et al., 1999a; Cheung and Hui, 1999; Furnell et al., 1998, 1999). In fact, there are quite a number of security concerns in this type of education system, for example, user authentication and access control. non-repudiation for critical actions like course registration, course tuition fee payment, confidentiality of user personal information, course material copyright protection, etc. For more information on what security issues an online learning system may consider, one can Online education has emerged as one of the major channels refer to the security framework given by Furnell for dissemination of learning materials. As more and more et al. (1998); and for more information on the problem of user authentication and access security concerns of these online education systems become control, one can refer to Cheung et al. (1999a) Cheung et al. (1999) provides a security model

such that a legitimately registered student cannot

Depending on the type of courses offered by

an organization, the security concerns may differ slightly. There is one security problem, the copyright protection problem, which is important to all kinds of e-courses, especially for type (1) and (3) courses mentioned above. The Emerald Research Register for this journal is available at Typical scenarios include the following: registered students infringing the copyrights of the course materials by passing the materials to non-registered students. Usually, the organization providing the course materials depends on the registration fee to maintain the operation of the organization. This copyright infringement severely jeopardizes the income of the organization.

#### Trustworthy Browsing - A Secure Web Accessing Model

Joe C.K. Yau<sup>1</sup>, Lucas C.K. Hui<sup>1</sup>, Bruce S.N. Cheung<sup>2</sup>, S.M. Yiu<sup>1</sup>, Y. Woo<sup>1</sup>, K.W. Lau<sup>1</sup>, Eric H.M. Li<sup>2</sup>

<sup>1</sup>Department of Computer Science, The University of Hong Kong {jckyau, hui, smyiu, ywoo, kwlau}@cs.hku.hk

<sup>2</sup>School of Professional And Continuing Education, The University of Hong Kong {bruce, ehmli}@hkuspace.hku.hk

#### Abstract

The web technology we are enjoying now is insecure, especially for accessing sensitive information. There is no solution that provides highly reliable user authentication to prove the identity of the information requestor to the server, nor a solution that securely protects the browsed information from being stolen.

To solve this problem, the Trustworthy Browsing system, based on a special browsing paradigm, is designed and developed. It employs a hardware-software hybrid solution and the key elements include the use of cryptographic hardware token for ensuring the user's identity, a customized browser to protect the contents, and an authentication protocol for verifying HTTP requests from the browser. The proposed solution can be integrated into applications for e-Education providers, e-Books publisher as well as electronic artwork publishers. In this paper, a detailed description and an in-depth security analysis for this system are given.

#### 1. Introduction

Nowadays, Internet has become one of the essentials to life. A research done by Nielsen//NetRatings shows that 75% Americans have Internet access [1], and the global Internet population is growing at a rate of 4% annually [2]. But as Internet becomes mature, users are becoming more concerned about security issues related to web access. From content provider's perspective, it is important that the content be delivered to eligible users only; and copyrighted contents

This research is supported in part by the UGC AoE Scheme (AoE/E-01/99), a RGC grant (HKU/7144/03E), and an ITF grant must be carefully protected from illegal copying. We need a web access model that is trustworthy. To ensure that the content is accessible only by eligible users, most systems require users to authenticate themselves using password before accessing the information. But if the content is confidential, authen-

ticating by password would not be secure enough. Another concern is that most of the web browsers available today allow users to save the content that they are viewing. But there are situations where content owners want to protect the contents from being copied, and only allow users to view the content online. Examples of such contents would be entertainment materials (e.g., movie or music), or information that is highly sensitive (e.g., a confidential document). With the popular web browsers we currently have, making copies of the content is easy. It is even possible for an adversary to dig into the browser's disk cache and retrieve the information. That brings threats to the content owners.

To better protect the content, we want to improve the security of web accessing. Specifically, we want a solution that has the following characteristics: (1) the content owner can be more certain of users'

identities who access the content; and

(2) securely protects the content being accessed. To solve this problem, we have designed and developed a hardware-software hybrid solution that can provide a secure browsing environment. In this paper, we introduce our solution - trustworthy browsing. In Section 2, possible applications for our solution are discussed. The problem is formally defined in Section 3. Section 4 gives a detailed description of our system, and Section 5 gives an indepth security analysis of it. Some concluding remarks and future research directions for this project are given in Section 6.

#### International Journal of Network Security, Vol.2, No.1, PP.21-28, Jan. 2006 (http://isrc.nchu.edu.tw/ijns/)

Towards a Secure Copyright Protection Infrastructure for e-Education Material: Principles Learned from Experience\*

Joe Cho-Ki Yau<sup>1</sup>, Lucas Chi-Kwong Hui<sup>1</sup>, Siu-Ming Yiu<sup>1</sup> and Bruce Siu-Nang Cheung<sup>2</sup> (Corresponding author: Joe Cho-Ki Yau)

Department of Computer Science, The University of Hong Kong<sup>1</sup> Pokfulam Road, Hong Kong. (Email: jckyau, hui, smyiu@cs.hku.hk) School of Professional and Continuing Education, The University of Hong Kong<sup>2</sup> Pokfulam Road, Hong Kong. (Email: bruce@hkuspace.hku.hk) (Received July 7, 2005; revised and accepted Aug. 9, 2005)

#### Abstract

1 Introduction

Copyright of e-Education material is valuable. The need concerns of e-Education. Among these security concerns, for protecting it is prominent. In the past two years, the copyright protection problem for the e-Education mawe have developed an infrastructure called e-Course eX- terial is one of the most important concerns that are vital change (eCX) for protecting the copyrights of e-Courses, to the operation of e-Education institutes. right from its development phase to its delivery phase. It has been adopted by an education institute, with a Almost in all e-Business sectors, the intellectual propuser-base of over 70,000 students, and has been receiving erty of material or content could be the most valuable positive feedback from students. To design a secure and asset of the business itself, and this is undoubtedly true effective copyright protection infrastructure is not trivial. for the e-Education sector. Many organizations rely on In particular, for efficiency purpose, one may allow students to retain a local copy of the e-Course material in Registered students infringing the copyrights of the course their own computers; on the other hand, we should make materials by passing the materials to non-registered stuto protect the material. In this paper, we summarize securely protected. some principles and knowledge we have gained through this project that should be observed for designing a seinciples would be useful to developers and researchers for designing and developing such a system. Keywords: Copyright protection, software protection, reverse engineering, e-Education, e-Learning

does not represent the official standpoint of HKU SPACE.

With the advent of the digital age, e-Education has be-

it difficult for them to make illegal copies of the material. dents can severely jeopardize the income of the organi-Only storing the material in encrypted form is not enough zation. Hence, copyrights of e-Course materials must be About three years ago, we initiated a study of the probcure copyright protection system. We believe that these lem of protecting the copyrights of e-Education materials. We proposed an infrastructure, called e-Course eXchange eCX [26, 27, 28, 29]. This infrastructure has been developed and deployed to a large group of users. In this paper, we will use eCX as a case study, and discuss the lessons we have learned from it. In Sections 2 and 3, we will take a closer look at e-Education, its security concerns, and, in particular, its domain specific requirements for copyright protection. In Section 4, we will give a brief survey on the existing copyright protection solutions. In Section 5, we will present the design of eCX. In Section 6, we will discome one of the most important channels for students to cuss the issues and design principles we learned from our acquire knowledge. Students of different levels are one experience with eCX. Section 7 then concludes the paper. way or the other making use of this new channel to learn, Although eCX may seem to be a solution specific to the and researchers are actively working on this area to make e-Education segment, the lessons we have learned are gen-\*This paper represents the personal opinion of the authors and eral enough and applicable to other copyright protection

the best use of it. However, as pointed out by Furnell

[9, 10], little attention has been devoted to the security



Q

PR1

# Pattern Recognition

IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 12, NO. 1, JANUARY 2003

#### Color Image Indexing Using BTC

Guoping Qiu

Abstract—This paper presents a new application of a well-studied image coding technique, namely block truncation coding the models visual keys with those of the query images. Because (BTC). It is shown that BTC can not only be used for compressing color images, it can also be conveniently used for content-based image retrieval from image databases. From the BTC compressed matrix (BCCM) and the other block pattern histogram (BPH). We use BCCM and BPH to compute the similarity measures of images for content-based image retrieval applications. Experimenta results are presented which demonstrate that BCCM and BPH are comparable to similar state of the art techniques.

Index Terms-Block truncation coding (BTC), color quantization, content-based image retrieval, image coding, image database.

T INTRODUCTION

coding/compression, which has been studied for more than effective and efficient image retrieval [4]. 30 years and significant advancements have been made. Many Block truncation coding (BTC) is a relatively simple image successful, efficient and effective image-coding techniques coding technique developed in the early years of digital imaging have been developed and the body of literature on image coding more than 20 years ago [5]. Although it is a simple technique, is huge. Well-developed and popular international standards, BTC has played an important role in the history of digital image e.g., [6], on image coding have also long been available and coding in the sense that many advanced coding techniques widely used in many applications.

the context of image database, which has also been actively re- BTC have long been surpassed by many newer image-coding searched by researchers from a wide range of disciplines in-techniques such as DCT (JPEG) [6] and wavelet [7], the comcluding those from computer vision, image processing, and tra-putational simplicity of BTC has made it and BTC-like image ditional database areas for over a decade [14]. One particularly coding techniques attractive in applications whereby real time promising approach to image database indexing and retrieval is fast implementation is desirable. Further more, with the rapid the query by image content (QBIC) method [1], whereby the vi-advancement in processor speed, storage device technology and sual contents of the images, such as color distribution (color histogram), texture attributes and other image features are extracted critical factor in many practical applications. Based on a certain from the image using computer vision/image processing tech-tradeoff, higher bit rate and complexity can be acceptable. On niques and used as indexing keys. In an image database, these the other hand, with very large image collections becoming visual keys are stored along with the actual imagery data and more and more common, effectively managing large image

Manuscript received March 9, 2001; revised October 2, 2002. The associate editor coordinating the review of this manuscript and approving it for publica-The author is with the School of Computer Science, The University of Nottingham, Nottingham NG8 IBB, U.K. (e-mail: qiu@cs.nott.ac.uk). Digital Object Identifier 10.1109/TIP.2002,807356

approach to QBIC is not efficient in terms of data storage. Not stream (without performing decoding), we derive two image content description features, one termed the block color co-occurrence matching/retrieval can only based on the pre-computed set of image features. Many image-coding methods developed over the years are essentially based on the extraction and retention of the most important (visual) information of the image. The retained important information, such as the DCT coefficients of JPEG can be used for

extra information has to be stored with the images, traditional

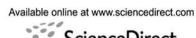
image indexing and object recognition [2]. However, since JPEG and other similar methods are not explicitly designed for image indexing purpose, models and features have to be derived from the transform coefficients, which generally involves complicated THE rapid expansion of the Internet and fast advancement and complex computation and also leads to an expansion of data. in color imaging technologies have made digital color images more and more readily available to professional and am- cue for image indexing [3], however, it is difficult to explicitly ateur users. The large amount of image collections available exploit color information from the transform coefficients without from a variety of sources (digital camera, digital video, scanner, decoding. On the other hand, nontransform based image coding the Internet, etc.) have posed increasing technical challenges to can have image features such as color more easily available. For computer systems to store/transmit and index/manage the image example, recent work on color image coding using vector quandata effectively and efficiently to make such collections easily tization has demonstrated that color as well as pattern information can be readily available in the compressed image stream The storage and transmission challenge is tackled by image (without performing decoding) to be used as image indices for

have been developed based on BTC or inspired by the success The challenge to image indexing/management is studied in of BTC. Even though the compression ratios achievable by database, making images easily accessible have becoming a challenge. Modern imaging systems not only require efficient coding, but also easy manipulation, indexing and retrieval, the so-called "fourth criterion" in image coding [18].

> In this paper we shall show another attractive feature of BTCtypes image coding methods. We shall demonstrate that BTC

1057-7149/03\$17.00 © 2003 IEEE









different applications. Even finding a starting query exam-

ple can be problematic because some visual forms are in-

trinsically hard to describe precisely. Second, current state

of the art technologies are not yet mature enough to realize

the ideal of CBIR. In particular, it is difficult to compute

image similarity measures that match the perceptual differ-

ences between images. Automatically retrieved images are

often not what the users expected, hence there is a gap be-

tween the retrieval results and users' expectation. Third and

fundamentally, the CBIR paradigm puts the user in a pas-

sive position in the sense that retrieval results are largely de-

termined by the computational algorithms; the user cannot

actively control the retrieval results. The immaturity of the

enabling computational algorithms has only made the task

Recent trends in CBIR have been to introduce brows-

ing, navigation and relevant feedback facilities to enable

users to interact with the retrieval system and to engage the

users, examples include Refs. [7-12]. The advantages of a

even more difficult.

#### Pattern Recognition 40 (2007) 1711-1721

#### Guoping Qiu\*, Jeremy Morris, Xunli Fan

Visual guided navigation for image retrieval

School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham NG8 1BB, UK

Received 12 April 2006; accepted 27 September 2006

In this work, we are interested in technologies that will allow users to actively browse and navigate large image databases and to retrieve images through interactive fast browsing and navigation. The development of a browsing/navigation-based image retrieval system has at least two challenges. The first is that the system's graphical user interface (GUI) should intuitively reflect the distribution of the images in the database in order to provide the users with a mental picture of the database content and a sense of orientation during the course of browsing/navigation. The second is that it has to be fast and responsive, and be able to respond to users actions at an interactive speed in order to engage the users. We have developed a method that attempts to address these challenges of a browsing/navigation based image retrieval systems. The unique feature of the method is that we take an integrated approach to the design of the browsing/navigation GUI and the indexing and organization of the images in the database. The GUI is tightly coupled with the algorithms that run in the background. The visual cues of the GUI are logically linked with various parts of the repository (image clusters of various particular visual themes) thus providing intuitive correspondences between the GUI and the database contents. In the backend, the images are organized into a binary tree data structure using a sequential maximal information coding algorithm and each image is indexed by an n-bit binary index thus making response to users' action very fast. We present experimental results to demonstrate the usefulness of our method both as a pre-filtering tool and for developing browsing/navigation systems for fast image retrieval from large image databases. © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Image database; Image retrieval; Browsing/navigation; Entropy; Information theory; Color

#### 1. Introduction

Managing large image database and providing effective tools for users to quickly find image items they are looking for is still a very challenging problem. In the past decade or so, the content-based image indexing and retrieval (CBIR) paradigm has dominated the research community, e.g. Refs. [1-6]. There are a number of intrinsic weaknesses associated with the traditional CBIR paradigm, which have hindered progress. First, the objectives of CBIR are ill defined. Although the idea of using visual examples to find similar images is sound and intuitive, it is problematic in practice. It is not clear in what circumstances/application scenarios users would want/prefer to search images by example. The definition of CBIR is too broad and vague; retrieval by example can mean different things to different users and in

\* Corresponding author. Tel.: +44 115 8466507; fax: +44 115 9514254. E-mail address: qiu@cs.nott.ac.uk (G. Qiu).

0031-3203/\$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

doi:10.1016/j.patcog.2006.09.020



#### Available at

www.ElsevierComputerScience.com POWERED BY SCIENCE @DIRECT\*

Pattern Recognition 37 (2004) 2177-2193



#### Compressing histogram representations for automatic colour photo categorization

Guoping Qiu\*, Xia Feng<sup>1</sup>, Jianzhong Fang

School of Computer Science, The University of Nottingham, Jubilee Campus, Nottingham NG8 1BB, UK Received 12 September 2003; received in revised form 8 March 2004; accepted 8 March 2004

Organizing images into semantic categories can be very useful for searching and browsing through large image repositories. In this work, we use machine learning to associate low level colour representations of digital colour photos with their high level semantic categories. We investigate the redundancy and performance of a number of histogram-based colour image content representations in the context of automatic colour photo categorization using support vector machines. We use principal component analysis to reduce the dimensionality of (high dimensional) histogram based colour descriptors and use support vector machines to learn to classify the images into various high level categories in the histograms subspaces. We present experimental results to demonstrate the usefulness of such an approach to organizing colour photos into semantic categories. Our results show that the colour content descriptors constructed in different ways perform quite differently and the performances are data dependent hence it is difficult to pick a "winning" descriptor. Our results demonstrate conclusively that all descriptors studied in this paper are highly redundant and that regardless of their performances, the dimensionalities of these histogram based colour content descriptors can be significantly reduced without affecting their classification performances. © 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Color histogram; Content-based indexing and retrieval; PCA; SVM

#### 1. Introduction

Fast advancement in digital imaging technology has resulted in the exponential increase of image data in both professional archives and personal leisure collections. Effectively managing large image repositories and making them easily accessible poses significant technical challenges. In the past decade, there has been significant research effort in content-based image retrieval (CBIR) [1]. In a CBIR system, a user can query the image repositories with a visual example and the system will return an ordered list of images that are similar to the query in some visual sense. Traditionally, image similarity is measured by some forms of

\* Corresponding author. Tel.: +44-115-8466507; fax: +44-115-951-4254. E-mail addresses: qiu@cs.nott.ac.uk (G. Qiu),

xxf@cs.nott.ac.uk (X. Feng), jzf@cs.nott.ac.uk (J. Fang). On leave from The Civil Aviation University of China and sponsored by the China Scholarship Council.

distance metrics in the feature space. However, similarity is a subjective concept, it is therefore not surprising and perhaps inevitable that there is a "gap" between a similarity measured by the CBIR systems and a similarity perceived by human observers. How to reduce this gap, often referred to as the 'semantic gap" in content-based image retrieval has received much attention in recent years, and technologies that provide solutions to reduce the semantic gap of CBIR are likely to play a pivotal role in content-based image indexing and

One direction pursued by researchers to narrow the semantic gap is based on the "learning by example" principle. Machine learning is an integral and essential part of human efforts to build intelligent machines, and the subject has been studied extensively in various disciplines of scientific and engineering pursues. In the context of CBIR, machine learning has been applied to classify collections of images into categories or classes of various descriptions [2]. In particular, image classification techniques based on various

0031-3203/\$30.00 © 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved. doi:10.1016/j.patcog.2004.03.006

Conservation laws for two (2+1)-dimensional differential-difference

Guo-Fu Yu\* Institute of Computational Mathematics and Scientific Engineering Computing, AMSS, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, P.R. CHINA and Graduate School of the Chinese Academy of Sciences, Beijing, P.R. CHINA

> Department of Computer Science, Hong Kong Baptist university, Kowloon Tong, Hong Kong, P.R. CHINA

Two integrable differential-difference equations are considered. One is derived from the discrete BKP equation and the other is a symmetric (2+1)-dimensional Lotka-Volterra equation. An infinite number

In recent years, much attention has been paid to studying a variety of intrinsic features shared by discrete integrable systems. Various methods have been developed to search for new discrete integrable systems, Lax pairs, soliton solutions, symmetries and conservation laws (CLs) etc. (See, e.g. [1]-[4] and references therein). Conservation laws play an important role in mathematics and engineering as well. The general approach to the search for a conservation law for a real-life problem is the variational method, where the Hamiltonian of the problem denotes a conservation law. It is easy to establish a variational formulation for a differential system by the semi-inverse method, but it is difficult to establish variational formulations for differential-difference systems [5]. Applications of the semi-inverse method can be found in the references

[6]-[7]. Concerning CLs of differential-difference equations, many results have been achieved by using some

successful methods. (See, e.g. [12]-[18].) However, to our knowledge, it seems that most examples of CLs of integrable differential-difference equations given in the literature are just (1+1) or (1+2)-dimensional (one discrete and two continuous ) cases [8]-[9]. Comparatively less (2+1)-dimensional (two discrete and one

In this paper, we will derive CLs for the following two (2+1)-dimensional (two discrete and one continuous differential-difference equations. The first one derives from the famous discrete BKP equation, which in the

 $[z_1 \exp(D_1) + z_2 \exp(D_2) + z_3 \exp(D_3) + z_4 \exp(D_4)]f \cdot f = 0,$ where  $D_1, D_2, D_3, D_4$  and  $z_1, z_2, z_3, z_4$  are bilinear operators and constants, respectively, satisfying  $D_1 + D_2 + D_3 + D_4 = 0,$   $z_1 + z_2 + z_3 + z_4 = 0.$ 

 $D_1 = \frac{1}{2}(\delta D_y + \epsilon D_x), \qquad z_1 = 1, \qquad D_2 = \frac{1}{2}(\delta D_y - \epsilon D_x), \qquad z_2 = -1 + \beta \delta \epsilon,$ 

 $D_3 = -D_n - \frac{1}{2}(\delta D_y + \epsilon D_x), \quad z_3 = -\alpha \epsilon - \beta \delta \epsilon, \quad D_4 = D_n - \frac{1}{2}(\delta D_y - \epsilon D_x), \quad z_4 = \alpha \epsilon,$ 

Spectral radius analysis of matrices and their

association with integrable systems\*

Honwah Tam<sup>1</sup> and Yufeng Zhang<sup>1,2</sup>

1 Department of Computer Science, Hong Kong Baptist University,

Hong Kong, P.R.China;

2 Information School, Shandong University of Science and Technology,

Qingdao Huangdao 266510, P.R.China

of conservation laws for the two differential-difference equations are deduced.

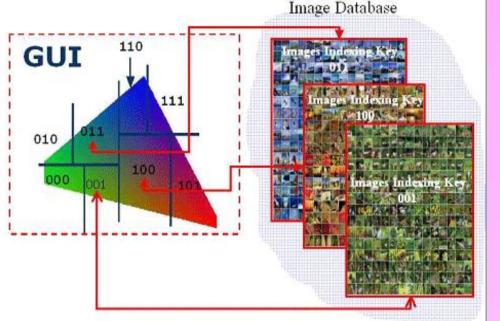
continuous) differential-difference equations have been considered for their CLs.

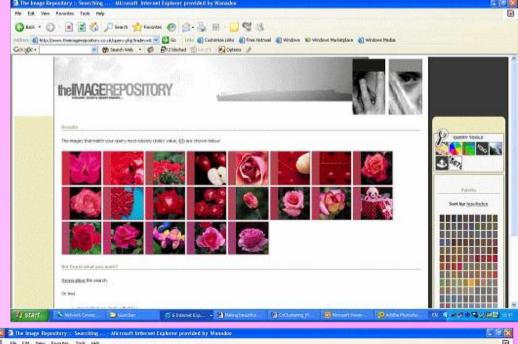
I Introduction

Hirota bilinear form is expressed by

\*Electronic mail:gfyu@lsec.cc.ac.cn









INSTITUTE OF PHYSICS PUBLISHING

J. Phys. A: Math. Gen. 39 (2006) 3367-3373

JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL doi:10.1088/0305-4470/39/13/014

#### On the nonisospectral Kadomtsev-Petviashvili equation

#### Guo-Fu Yu1,2 and Hon-Wah Tam

matics and Scientific Eng Academy of Mathematics and System Sciences, Chinese Academy of Sciences, PO Box 2719, Beijing 100080, People's Republic of China Graduate School of the Chinese Academy of Sciences, Beijing, People's Republic of China <sup>3</sup> Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China

Received 8 December 2005 Published 15 March 2006

Online at stacks.iop.org/JPhysA/39/3367

In this paper, we first present the Grammian determinant solutions to the nonisospectral Kadomtsev-Petviashvili (KP) equation. Then, by using the Pfaffianization procedure of Hirota and Ohta, an integrable coupled system is generated. Moreover, Gramm-type Pfaffian solutions to the Pfaffianized

PACS numbers: 02.30.Ik, 02.30.Jr, 05.45.Yv Mathematics Subject Classification: 35Q58, 37K40

#### 1. Introduction

In the early 1990s, Hirota and Ohta [1, 2] developed a procedure for generalizing equations from the Kadomtsev-Petviashvili (KP) hierarchy to produce coupled systems of equations, which we now call Pfaffianization. These Pfaffianized equations appear as coupled systems of the original equations and have soliton solutions expressed by Pfaffians. Such a procedure has been successfully applied to the DS equations [3], the discrete KP equation [4], the selfdual Yang-Mills equation [5], the two-dimensional Toda lattice [6], the semi-discrete Toda equation [7], the differential-difference KP equation [8], etc. In [9], Wronskian solutions of the nonisospectral KP equation [10]

 $4u_t + y(u_{xxx} + 6uu_x + 3\partial_x^{-1}u_{yy}) + 2xu_y + 4\partial_x^{-1}u_y = 0$ are derived by the Hirota method and the Wronskian technique. The solutions of the bilinear nonisospectral KP equation are expressed in Wronskian determinants. Since there are both Wronskian determinant and Grammian determinant solutions for the KP equation [11], we expect that there also exist Gramm-type expressions for (1). On the other hand, all the Pfaffianization procedures above are applied to isospectral systems. Hence it would be very interesting to consider Pfaffianization for nonisospectral systems.

0305-4470/06/133367+07\$30.00 © 2006 IOP Publishing Ltd Printed in the UK

#### Abstract

This paper begins with an isospectral problem and analyzes the spectral radius of its corresponding spectral matrix. This work enlightens us to set up a higher-dimensional isospectral problem whose compatibility condition gives rise to a (2+1)-dimensional zero curvature equation. From this equation a (2+1)-dimensional Lax integrable soliton equation hierarchy with constraints of potential functions, along with 5 parameters, is generated. The reduced cases of this hierarchy give three (2+1)-dimensional integrable systems, namely, the AKNS hierarchy, the Levi hierarchy and the D-AKNS hierarchy. Extending the above Lie algebra into more complicated ones, two integrable couplings of the (2+1)-dimensional hierarchy are derived. One of the two couplings has Hamiltonian structure by employing the quadratic-form identity. The corresponding integrable couplings of the reduced systems are also obtained. Finally, as a comparison study for generating expanding integrable systems, an antisymmetric Lie algebra and its corresponding loop algebra are constructed. From this Lie algebra a great many enlarging integrable systems can be generated. In addition, their Hamiltonian structures can be computed by the trace identity.

Keywords: spectral radius analysis, integrable couplings, Lie algebra, Hamiltonian structure

\*This work was supported by the Hong Kong Research Grant Council under grant number HKBU2016/05P and the National Science Foundation of China under grant number 10471139.

## Pattern Recognition

IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 16, NO. 5, MAY 2007

Skeleton has been defined in various ways in the different lit-

eratures. All kinds of definitions result in a variety of skeletion-

ization algorithms as well. However, it is generally agreed that a

appearance. In addition, the algorithm is fast and insensitive to

Generally, the skeleton of a character is the locus of the

midpoints or the symmetric axis of the character stroke [8].

skeletonization techniques have been proposed, among which

there are three well-known algorithms of the local symmetry

Simon [11] partitioned the character stroke into regular and

singular regions. The singular region corresponds to ends, inter-

sections and turns, and the regular region covers the other parts

of the strokes. As shown in Fig. 5, the white and shadow parts

represent the singular and regular regions respectively. Thus,

the skeletionization process of character based on the symmetry

analysis contains two main steps: 1) computing the primary

skeleton in the regular region and 2) amendment processing the

on the opposite side of the contour such that they form the

problem [6], [13], [14]. Unfortunately, its implementation often

suffers from the complicated computation, as well as the three

metrical approaches produce the skeleton with strong deviation

from the center of the underlying character strokes and rough

noise and affine transformations.

#### Wavelet-Based Approach to Character Skeleton

Xinge You, Member, IEEE, and Yuan Yan Tang, Fellow, IEEE

Abstract-Character skeleton plays a significant role in charmay consume less storage space and processing time, and is sufacter recognition. The strokes of a character may consist of two ficient for many applications. This representation is particularly regions, i.e., singular and regular regions. The intersections and junctions of the strokes belong to singular region, while the straight and smooth parts of the strokes are categorized to regular region.

Therefore, a skeletonization method requires two different processes to treat the skeletons in theses two different regions. All recognition, signature verification [7], etc. traditional skeletonization algorithms are based on the symmetry analysis technique. The major problems of these methods are as follows. 1) The computation of the primary skeleton in the regular region is indirect, so that its implementation is sophisticated and costly. 2) The extracted skeleton cannot be exactly located on the good skeleton should have some good properties. For instance, central line of the stroke. 3) The captured skeleton in the singular a skeleton should be homotopic to the underlying shape, and it region may be distorted by artifacts and branches. To overcome should be thin and well centered, it should have a pleasing visual these problems, a novel scheme of extracting the skeleton of character based on wavelet transform is presented in this paper. This scheme consists of two main steps, namely: a) extraction of primary skeleton in the regular region and b) amendment processing of the primary skeletons and connection of them in the singular region. A direct technique is used in the first step, where a new wavelet-based Different techniques of the local symmetry analysis may gensymmetry analysis is developed for finding the central line of the erate different symmetric axes (skeletons). A great deal of stroke directly. A novel method called smooth interpolation is designed in the second step, where a smooth operation is applied to the primary skeleton, and, thereafter, the interpolation compensathe primary skeleton, and, thereafter, the interpolation compensa-tion technique is proposed to link the primary skeleton, so that the skeleton in the singular region can be produced. Experiments are local symmetry (SLS) [9], and process-inferring symmetry conducted and positive results are achieved, which show that the analysis (PISA) [10]. proposed skeletonization scheme is applicable to not only binary mage but also gray-level image, and the skeleton is robust against

Index Terms-Interpolation compensation, modulus maxima, skeleton, wavelet transform.

I. INTRODUCTION

KELETONIZATION of character is one of the most active skeleton in the singular region. and important areas in image processing and pattern recognition. The skeleton is especially suitable for describing char-that it is difficult to decide the symmetric pairs from the boundacter since they have natural and inherent axes [1]-[3]. From aries of the character strokes in both continuous and discrete a practical point of view, the skeleton-based representation of domains [12]. Particularly, in the discrete domain, it may not characters by sets of thin curves rather than by a raster of pixels be possible from one side to find the symmetric counterparts

Manuscript received March 2, 2005; revised December 5, 2005. This local symmetries. Even if a pair of symmetric contour pixels work was supported in part by the National Science Foundation under Grant 60403011; in part by the Science and Technology Department, Hubei Province and Wuhan, China, under Grants 2006ABA023 and 20045006071-17, respectively. The technique of the regularity-singularity tively; in part by the Chongqing University of China; in part by RGC; and in part by the FRG of Hong Kong Baptist University. The associate editor coordinating the review of this manuscript and appropriate the review of the nating the review of this manuscript and approving it for publication was traditional symmetry analysis [8]-[10]. Moreover, the nonsym-

X. You is with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: xyou@comp.hkbu.edu.hk). Y. Y. Tang is with the Wuhan National Laboratory for Optoelectronics. Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: yytang@comp.hkbu.edu.hk)

Character strokes contain a variety of intersections and junctions, which belong to the singular region. For such cases, an-

1057-7149/\$25.00 @ 2007 IEEE

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 25, NO. 9, SEPTEMBER 2003

#### Skeletonization of Ribbon-Like Shapes Based on a New Wavelet Function

Yuan Yan Tang, Senior Member, IEEE, and Xinge You

Abstract—A wavelet-based scheme to extract skeleton of Ribbon-like shape is proposed in this paper, where a novel wavelet function plays a key role in this scheme, which possesses three significant characteristics, namely, 1) the position of the local maximum moduli of the wavelet transform with respect to the Ribbon-like shape is independent of the gray-levels of the image. 2) When the appropriate scale of the wavelet transform is selected, the local maximum moduli of the wavelet transform of the Ribbon-like shape produce two new paralle contours, which are located symmetrically at two sides of the original one and have the same topological and geometric properties as that of the original shape. 3) The distance between these two parallel contours equals to the scale of the wavelet transform, which is independent of the width of the shape. This new scheme consists of two phases: 1) Generation of wavelet skeleton—based on the desirable properties of the new wavelet function, symmetry analyses of the maximum moduli of the wavelet transform is described. Midpoints of all pairs of contour elements can be connected to generate a skeleton of the shape, which is defined as wavelet skeleton. 2) Modification of the wavelet skeleton-Thereafter, a set of techniques are utilized for modifying the artifacts of the primary wavelet skeleton. The corresponding algorithm is also developed in this paper. Experimental results show that the proposed scheme is capable of extracting exactly the skeleton of the Ribbon-like shape with different width as well as different gray-levels. The skeleton representation is robust against noise and affine transformation

Index Terms-Ribbon-like shape, skeletonization, wavelet transform, wavelet skeleton.

#### INTRODUCTION

character recognition, signature verification, understanding of paper-based graphics including maps and engineering drawings [11], [28], computer-assisted cartoon, fingerprint analysis, etc. Representation of a shape using a suitable form is essential in these recognition systems. An appropriate representation of a Ribbon-like shape is its skeleton [34]. The objective of this study is to extract skeletons from planar Ribbon-like shapes, which have the following properties:

1. conforming to human perceptions of the original 2. being centered inside the original shapes,

being efficiently computable, and 4. being robust against noise and geometric Skeletons have been defined in various ways in the different literatures. Generally, the skeleton of a shape is referred to as the locus of the symmetric points or symmetric axes of the local symmetries of the shape [1], in other words, different local symmetry analyses may result in different

Manuscript received 18 Apr. 2002; revised 29 Oct. 2002; accepted 24 Feb. 2003. Recommended for acceptance by S. Sarkar.

For information on obtaining reprints of this article, please send e-mail to:

is not shown further whether it is capable of skeleton extraction.

0162-6828/03/\$17.00 @ 2003 IEEE

Wavelet Theory and

Its Application to

**Pattern Recognition** 

ONE of the most significant topics in pattern recognition is analysis of Ribbon-like shapes. It can be applied to community, namely,

Blum's Symmetric Axis Transform (SAT) [1], Brady's Smoothed Local Symmetry (SLS) [3], and Leyton's Process-Inferring Symmetry Analysis

(PISA) [15]. Each method contributes greatly to skeletonization of the shapes. The difference in selection of the location of a symmetric point makes these methods distinguishable from each other. Nevertheless, the major problem of the SAT and PISA, as pointed out by Brady [3], is that the symmetric points of a local symmetry and, hence, a skeleton segment may lie in a perceptually distinct part of the underlying shape. Although the skeleton obtained from the SLS has a pleasing visual appearance, the major shortcoming of the SLS is that some perceptually irrelevant symmetric axes may be created. Some grouping rules can be used [8]. However, these rules may not be appropriate for a wide range of shapes. An alternative way is to divide the shape into several parts and, thereafter, the SLS axes are computed within each part [23]. Dividing a shape into suitable parts is crucial in this method; unfortunately, it is a difficult task. Another powerful measure

symmetric points and, hence, different skeletons are produced. There are three methods of the local symmetry of the symmetry proposed by Kovesi [12], [13] is based on the analysis of phase information. The mathematics of phase Y.Y. Tang is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong.
E-mail: yylang@comp.hkbu.edu.hk.
X. You is with the Faculty of Mathematics and Computer Science, Hubei University, Wuhan, China, 430062, P.R. China, and is also with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong. E-mail: xyou@comp.hkbu.edu.hk. rather awkward quantity to calculate. Moreover, it provides limited information about the overall shape of an object and it

Published by the IEEE Computer Society



Available at www.ElsevierComputerScience.com POWERED BY SCIENCE @DIRECT\* Pattern Recognition 38 (2005) 323-339

PATTERN RECOGNITION

Directed connection measurement for evaluating reconstructed stroke sequence in handwriting images

> K.K. Lau\*, Pong C. Yuen, Yuan Y. Tang Department of Computer Science, Hong Kong Baptist University, Hong Kong Received 28 May 2003; received in revised form 15 July 2004; accepted 15 July 2004

Generally speaking, the performance of on-line handwriting recognition systems is better than that of off-line handwriting recognition systems. Many researchers then started to extract/estimate on-line/temporal information from a static handwritten image to improve the performance of off-line recognition systems. More specifically, various algorithms have been developed to extract the stroke sequence, which is an important temporal information, from a static handwritten image in the last decade, However, to the best of our knowledge, there are no methods to evaluate the performance of these algorithms. In view of this limitation, this paper presents a directed connection measurement for evaluating the performance of these methods. The measurement is designed in terms of direction and connection. The direction measurement is based on the consecutive arrangements of the items, while the connection measurement is based on both the number of disconnections and the Feigin and Cohen model (FCM) in ranking analysis. We replace the Kendall's distance, the key metric component in FCM, by a new connection metric. We have also proved that the new direction and connection metrics satisfy all axioms for being a valid metric. The two correlations are combined to form a directed connection correlation, which is applicable in comparing the performance of different stroke sequence recovery algorithms. © 2004 Published by Elsevier Ltd on behalf of Pattern Recognition Society.

Keywords: Ranking analysis; Handwriting recognition; Stroke sequence estimation

Handwriting recognition has been an active and challenging research area since the 1980s and attracts a lot of researchers working on this area. Basically, handwriting recognition can be divided into two problem domains. namely on-line and off-line. In off-line handwriting recognition, what we have is a static image. The typical applications include bank check processing and postal address recognition. Over the two decades of development, the performance

\* Corresponding author. E-mail addresses: kklau@comp.hkbu.edu.hk (K.K. Lau), pcyuen@comp.hkbu.edu.hk (P.C. Yuen), yytang@comp.hkbu.edu.hk (Y.Y. Tang).

has greatly improved. However, due to the large variations on writing images, the performance is still not fully satisfied. In on-line handwriting recognition, in addition to the static handwritten image, we have on-line writing information, such as stroke sequence, pressure, velocity and acceleration. Thus, on-line handwriting recognition system usually performs better than that of an off-line one. A very good review can be found in Ref. [1].

Govindaraju and Krishnamurthy [2] suggested that although each person has his/her own personal handwriting

style, the basic pen-tip movements for two different persons writing the same script are close. Besides, Plamondon and Srihari [1] stated that the recovered temporal information may be helpful in improving the recognition performance. In view of this, some researchers started to think about the

0031-3203/\$30.00 © 2004 Published by Elsevier Ltd on behalf of Pattern Recognition Society. doi:10.1016/j.patcog.2004.07.006

Optical Engineering 44(5), 057002 (May 2005)

#### Component-based subspace linear discriminant analysis method for face recognition with one training sample

Jian Huang Pong C. Yuen Hong Kong Baptist University Department of Computer Science Hong Kong, China E-mail: jhuang@comp.hkbu.edu.hk

Department of Mathematics

Wen-Sheng Chen

Jian Huang Lai ZhongShan University Department of Mathematics

Abstract. Many face recognition algorithms/systems have been developed in the last decade and excellent performances have also been reported when there is a sufficient number of representative training samples. In many real-life applications such as passport identification only one well-controlled frontal sample image is available for training. Under this situation, the performance of existing algorithms will degrade dramatically or may not even be implemented. We propose a component-based linear discriminant analysis (LDA) method to solve the one training sample problem. The basic idea of the proposed method is to construct local facial feature component bunches by moving each local feature region in four directions. In this way, we not only generate more samples with lower dimension than the original image, but also consider the face detection localization error while training. After that, we propose a subspace LDA method, which is tailor-made for a small numper of training samples, for the local feature projection to maximize the discrimination power. Theoretical analysis and experiment results show that our proposed subspace LDA is efficient and overcomes the limita tions in existing LDA methods. Finally, we combine the contributions of each local component bunch with a weighted combination scheme to draw the recognition decision. A FERET database is used for evaluating the proposed method and results are encouraging. © 2005 Society of Photo Optical Instrumentation Engineers. [DOI: 10.1117/1.1900104]

Subject terms: face recognition; component-based approach; subspace linear discriminant analysis; weighted combination scheme. Paper 040215 received Aug. 31, 2004; revised manuscript received Nov. 18, 2004; accepted for publication Nov. 19, 2004; published online May 13, 2005.

1 Introduction Face recognition research started in the late 1970's and has

been one of the active and exciting researches in computer science and information technology areas since 1990.1,2 Recently, many face recognition algorithms have been developed, such as eigenface, <sup>3-6</sup> linear discriminant analysis (LDA), <sup>7-9</sup> and elastic bunch graph matching. <sup>10</sup> It is well known that LDA-based methods have proven to be an effective appearance-based approach for developing practical face recognition systems. Its superior performance has been reported when there are sufficient and representative samples for training in many literatures in recent years. However, many real-life face recognition applications could only offer one training sample image per person, such as identity card verification, passport verifica-

tion in customs, and law enforcement scenarios. In such situations, we only have one frontal image per person captured under controlled lighting conditions. Training a face recognition system with one sample is hard, as it is difficult to learn the other image variations. Fortunately, in such applications, the testing images are usually captured under a controlled environment such that the light variations will not change dramatically and the face images are nearly frontal view. But still, a general LDA algorithm cannot be

0091-3286/2005/\$22.00 © 2005 SPIE

such applications where there is only one sample per person for training. The first problem is that the within-class scatter matrix S, cannot be calculated when there is only one training sample per person, so the LDA method cannot be applied directly. Now suppose we could offer more than one training sample image per person. Will the LDA-based algorithm always work? A general LDA algorithm may still have some problems. This is the second problem that needs to be solved. It is also known as the small sample size (S3) problem. In recent years, some LDA-based algorithms have been developed to solve the small sample size problem5. and can be divided into two categories, namely, the twostage approach: dimension reduction followed by the LDA and subspace approach. This work combines the concept of the component

implemented in such a situation and no LDA-based method

is designed for solving one training sample problem as fa

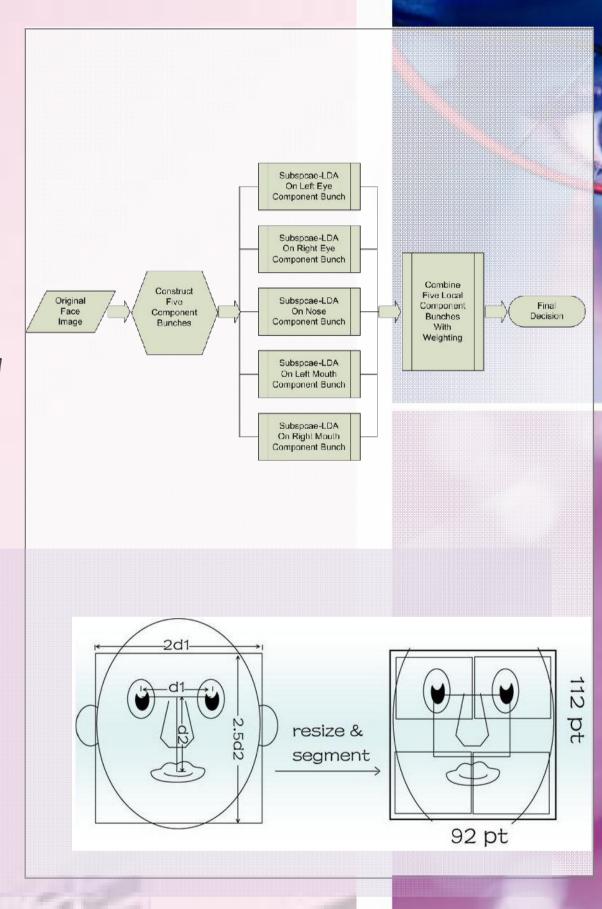
as we know. There are two problems that need to be solved

before the LDA-based method can be directly applied in

based approach and LDA method, and develops a component-based LDA method for face recognition with one training sample. The basic idea of the proposed method is to construct local facial feature component bunches by moving each local feature region in four directions. In this way, we not only generate more samples at lower dimension than the original image, but also consider the face

May 2005/Vol. 44(5)

### Wavelet Theory and Its Application Pattern Recognition MULTIMODAL INTERFACE FOR HUMAN-MACHINE COMMUNICATION to MACHINE PERCEPTION P. C. Yuen, Y. Y. Tang ARTIFICIAL INTELLIGENCE & P. S. P. Wang MACHINE PERCEPTION ARTIFICIAL INTELLIGENCE Volume 48 **World Scientific**



IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 35, NO. 5, OCTOBER 2005

#### GA-Fisher: A New LDA-Based Face Recognition Algorithm With Selection of Principal Components

Wei-Shi Zheng, Jian-Huang Lai, and Pong C. Yuen

uncorrelated optimal discriminant vectors.

A number of LDA-based recognition algo-

superior performance of LDA on face recognition has been

reported in many literatures and Face Recognition Technology

(FERET) evaluation test [9]. Despite the advantages of using

the number of training samples is less than total number of

pixels in an image. Under this situation, the within-class scatter

matrix will be singular. In turn the inverse of within-class

scatter matrix cannot be calculated. S3 problem always occur

Basically, there are at least four approaches proposed to

overcome the S3 problem. The most well-known technique is

the Fisherface [1], which combines the techniques of Principal

[7] before LDA. The basic idea of PCA for dimension reduc-

tion is to keep the (top n) largest nonzero eigenvalues and the

corresponding eigenvectors for LDA. The idea of this approach

is correct from image compression point of view; keeping

the largest nonzero principal components means that we keep

most of the energy (information) of that image by projecting

into lower dimension subspace. However, from the pattern

classification point of view, this argument may not be true. The

main reason is that, in pattern classification, we would like to

find a set of projection vectors that can provide the highest

discrimination between different classes. Therefore, choosing

the largest principal components as the bases for dimension-

ality reduction may not be optimal. Along this line, Zhao et al.

[16] and Pentland et al. [21] proposed to remove the largest

three eigenvalues for representation. A lot of experiments have

been performed to support this argument, but the theoretical

justifications are not enough. The second approach is adding

on LDA\QR is equivalent to the one on generalized LDA. The

nullspace method [3] is another approach. The basic idea of

Abstract—This paper addresses the dimension reduction problem in Fisherface for face recognition. When the number of discriminant vectors [5], and in 2001, Jin et al. [4] presented training samples is less than the image dimension (total number of -class scatter matrix (Sw) in Linear Discriminant (PCA) is suggested to employ in Fisherface for dimension reduction of Sw so that it becomes nonsingular. The popular method is technique is the Fisherface, which combines the techniques o select the largest nonzero eigenvalues and the corresponding of Principal Component Analysis (PCA) with the LDA. The eigenvectors for LDA. To attenuate the illumination effect, some researchers suggested removing the three eigenvectors with the largest eigenvalues and the performance is improved. However, as far as we know, there is no systematic way to determine which eigenvalues should be used. Along this line, this paper proposes a theorem to interpret why PCA can be used in LDA and an small sample size (S3) problem. The problem happens when be used in LDA using a Genetic Algorithm (GA). A GA-PCA is also be used as part of the basis for dimension reduction. Using the GA-PCA to reduce the dimension, a GA-Fisher method is designed and developed. Comparing with the traditional Fish- in face recognition applications. erface method, the proposed GA-Fisher offers two additional advantages. First, optimal bases for dimensionality reduction are derived from GA-PCA. Second, the computational efficiency of LDA is improved by adding a whitening procedure after dimension reduction. The Face Recognition Technology (FERET) and Component Analysis (PCA) with the LDA (it is also known as Carnegie Mellon University Pose, Illumination, and Expression PCA+LDA). It performs the dimension reduction by PCA [6], (CMU PIE) databases are used for evaluation. Experimental results show that almost 5% improvement compared with Fisherface can be obtained, and the results are encouraging. Index Terms-Dimension reduction, face recognition, GA-PCA, genetic algorithms, LDA, PCA.

I. INTRODUCTION

INEAR DISCRIMINANT analysis (LDA) [1], [16] has been one of the popular techniques employed in the face recognition. The basic idea of the Fisher Linear Discriminant is to calculate the Fisher optimal discriminant vectors so that the ratio of the between-class scatter and the within-class scatter (Fisher Index) is maximized. In addition to maximizing the Fisher Index, some restrictions, when finding the optimal vectors, are added to reduce the error rate in face recognition.

Manuscript received May 6, 2004; revised November 26, 2004. This work was supported in part by the NSFC under Grant 60144001, the NSF of Guang-Dong, China under Grant 021766, the RGC Earmarked Research Grant HKBU-2119/03E, and the Key (Key grant) Project of Chinese Ministry of Education under Grant 105134. The Associate Editor recommending this paper was Vitters Mexico. torro Marino.

W.-S. Zheng and J.-H. Lai are with the Mathematics Department, Sun Yal-sen
University (ZhongShan University), Guangzhou 510275, China (e-mail: Sunny-

WeiShi@163.com; stsljh@zsu.edu.cn).

University, Hong Kong (e-mail: pcyuen@comp.hkbu.edu.hk).
Digital Object Identifier 10.1109/TSMCB.2005.850175

approach is utilizing the pseudo-inverse of the within-class and it has been proved that the solution to the eigenproblem P. C. Yuen is with the Department of Computer Science, Hong Kong Baptist

#### A Novel Incremental Principal Component Analysis and Its Application for Face Recognition Haitao Zhao, Pong Chi Yuen, Member, IEEE, and James T. Kwok, Member, IEEE

eigenface system was developed in 1991. In 1997, PCA was

Usually, PCA is performed in batch mode. It means that

all training data have to be ready for calculating the PCA

once the training data have been fully processed. If we want

projection matrix, the matrix has to be retrained with all training

data. In turn, it is hard to scale up the developed systems. To

overcome this limitation, an incremental method is a straight-

Incremental PCA (IPCA) has been studied for more than

two decades in the machine-learning community. Many IPCA

methods have also been developed. Basically, existing IPCA

algorithm developed by Weng et al. is the most recent work.

(SGA) [17], the CCIPCA algorithm generates "observations" in

PCs. Suppose that sample vectors are acquired sequentially,

e.g.,  $u(1), u(2), \ldots$ , possibly infinite. The first k dominant PCs

a) if i = n, initialize the *i*th PC as  $v_i(n) = u_i(n)$ ;

 $v_1(n), v_2(n), \ldots, v_k(n)$  are obtained as follows [13].

For n = 1, 2, ..., do the followings steps.

2) For  $i = 1, 2, ..., \min(k, n)$ , do:

1)  $u_1(n) = u(n)$ .

a complementary space for the computation of the higher order

recognition technology.

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, VOL. 36, NO. 4, AUGUST 2006

Abstract-Principal component analysis (PCA) has been proven not only reduce the image dimension, but also provides a to be an efficient method in pattern recognition and image analysis. Recently, PCA has been extensively employed for facerecognition algorithms, such as eigenface and fisherface. The have been reported and discussed in the literature. Many PCA-based face-recognition systems have also been analysis and a new algorithm, namely fisherface, was develdeveloped in the last decade. However, existing PCA-based face-oped. After that, PCA has been extensively employed in facerecognition systems are hard to scale up because of the computational cost and memory-requirement burden. To overcome this imitation, an incremental approach is usually adopted. Incremental PCA (IPCA) methods have been studied for many years in the machine-learning community. The major limitation of existing projection matrix during training stage. The learning stops IPCA methods is that there is no guarantee on the approximation error. In view of this limitation, this paper proposes a new IPCA to incorporate additional training data into an existing PCA method based on the idea of a singular value decomposition (SVD) updating algorithm, namely an SVD updating-based IPCA (SVDU-IPCA) algorithm. In the proposed SVDU-IPCA algorithm, we have mathematically proved that the approximation error is bounded. A complexity analysis on the proposed method is also forward approach. presented. Another characteristic of the proposed SVDU-IPCA orithm is that it can be easily extended to a kernel version. The proposed method has been evaluated using available public databases, namely FERET, AR, and Yale B, and applied to existing face-recognition algorithms. Experimental results show that the algorithms can be divided into two categories. The first category difference of the average recognition accuracy between the proosed incremental method and the batch-mode method is less than the covariance matrix [13], [16], [17]. To the best of our 1%. This implies that the proposed SVDU-IPCA method gives a close approximation to the batch-mode PCA method.

knowledge, the candid covariance-free IPCA (CCIPCA) [13] Index Terms-Error analysis, face recognition, incremental Instead of estimating the PCs by stochastic gradient ascent

principal component analysis (PCA), singular value decomposition (SVD).

I. INTRODUCTION RACE RECOGNITION has been an active research area in the computer-vision and pattern-recognition societies [1]-[6] in the last two decades. Since the original input-image space has a very high dimension [2], a dimensionality-reduction technique is usually employed before classification takes place. Principal component analysis (PCA) [1], [7] is one of the most popular representation methods for face recognition. It does

Manuscript received February 7, 2005; revised July 18, 2005. This project was supported in part by a Faculty research Grant of the Hong Kong Baptist University and in part by the Research Grant Council under Grant HKBU-2119/03E. This paper was recommended by Associate Editor H. Qiao. H. Zhao is with the Institute of Aerospace Science and Technology, Shanghai Jiaotong University, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR (e-mail: zhaoht@

$$\begin{split} v_i(n) &= \frac{n-1-l}{n} v_i(n-1) + \frac{1+l}{n} u_i(n) u_i^{\mathrm{T}}(n) \frac{v_i(n)}{\|v_i(n)\|} \\ u_{i+1}(n) &= u_i(n) - u_i^{\mathrm{T}}(n) \frac{v_i(n)}{\|v_i(n)\|} \frac{v_i(n)}{\|v_i(n)\|} \end{split}$$

Since the PCs are obtained sequentially and the computation i. Since the PCs are obtained sequentially and the computation of the i-th PC, the error will i. T. Kwok is with the Department of Computer Science, The Hong Kong i. T. Kwok is with the Department of Computer Science, The Hong Kong i. The Hong Kong i is the propagated and accumulated in the process. Although the University of Science and Technology, Hong Kong, SAR (e-mail: jamesk@ estimated vector  $v_i(n)$  converges to the ith PC, no detailed error Digital Object Identifier 10.1109/TSMCB.2006.870645 analysis was presented in [13]. Further, Weng et al. [13] defined

#### Kernel Machine-Based One-Parameter Regularized Fisher Discriminant Method for Face Recognition Wen-Sheng Chen, Pong C. Yuen, Jian Huang, and Dao-Qing Dai

IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS -- PART B: CYBERNETICS, VOL. 35, NO. 4, AUGUST 200

Abstract—This paper addresses two problems in linear dis- Fisherface [3], was reported in 1997. After that, a number of criminant analysis (LDA) of face recognition. The first one is the LDA-based face recognition algorithms/systems are then develproblem of recognition of human faces under pose and illumina-tion variations. It is well known that the distribution of face images good result under controlled conditions and nonlinear. The traditional linear methods, such as LDA, will not give a satisfactory performance. The second problem is the LDA is a linear method and is hard to solve nonlinear problem, small sample size (S3) problem. This problem occurs when the while the second is the small sample size (S3) problem. number of training samples is smaller than the dimensionality of For the first problem, it is well known that the distributions feature vector. In turn, the within-class scatter matrix will become of face image variations under different pose and illumination singular. To overcome these limitations, this paper proposes a new kernel machine-based one-parameter regularized Fisher discriminant (KIPRFD) technique. KIPRFD is developed based on based methods, the performance of LDA-based method will degrade under pose and illumination variations. To overcome nalysis method and the well-known kernel approach. Therefore, this drawback, kernel method is employed. The basic idea is parameter and kernel parameter. This paper further proposes a

with the existing LDA-based methods, the proposed method gives Index Terms—Face recognition, pose and illumination variations, RBF kernel function, regularized discriminant analysis, small sample-size problem.

I. INTRODUCTION

ACE recognition research started in the late 1970s and has P become one of the most active and exciting research areas in computer science and information technology since 1990. A number of algorithms/techniques have been developed in the last decade. Two very good reviews can be found in [1] and [2]. Among various techniques, the linear (Fisher) discriminant analysis (LDA) based[3] method is one of the promising approaches in face recognition. LDA is theoretically sound and its objective is to find the most discriminant feature for classification. Therefore, it is good for pattern recognition problem. The first well-known LDA-based face recognition algorithm,

supported by the Science Faculty Research Grant of Hong Kong Baptist University, RGC Earmarked Research Grant HKBU-2119/03B and NSFC (60144001, Shenzhen 518060, China (e-mail: chenws@szu.edu.cn).

10101013, 60175031, 10231040). This paper was recommended by Associate comp.hkbu.edu.hk; jhuang@comp.hkbu.edu.hk). D.-Q. Dai is with the Department of Mathematics, Sun Yat-Sen (ZhongShan) University, Guangzhou 510275, China (e-mail: stsddq@zsu.edu.cn). Digital Object Identifier 10.1109/TSMCB,2005.844596

to apply a nonlinear mapping  $\Phi: x \in \mathbb{R}^d \to \Phi(x) \in F$  to the input data vector x in input space  $R^d$  and then to perform kernel and regularized parameter in within-class scatter matrix the LDA on the mapped feature space F. This method is the multaneously based on the conjugate gradient method. Three so-called kernel Fisher discriminant (KFD) [4], [5]. The feature databases, namely FERET, Yale Group B, and CMU PIE, are space F could be considered as a linearization space. However, selected for evaluation. The results are encouraging. Comparing its dimensionality could be arbitrarily large and possibly infinite. Direct applying LDA method to feature space is impossible. Fortunately, the exact mapping function  $\Phi$  does not need to be used directly. The feature space can become implicit by using kernel trick, in which the inner products  $\langle \Phi(x_i), \Phi(x_i) \rangle$ in F can be replaced with a kernel function  $K(x_i, x_i)$ , i.e.,  $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle = \Phi(x_i)^T \cdot \Phi(x_j)$ , where  $x_i, x_j$ are input pattern vectors in input space  $R^d$ . So, the nonlinear mapping  $\Phi$  can be performed implicitly in input space  $\mathbb{R}^d$ . Owing to the nonlinear property of KFD and its strong theoretical background, it has been applied in face recognition to solve the pose and illumination problems since 2000 [6]–[11]. It is found that kernel-based methods [6]-[8] outperform linear methods and ICA. Liu et al.[9] performed a good experimental analysis on the KFD and showed that KFD, in general, gives better performance than that of KPCA. Lu et al. proposed [10], [11] to solve the pose problem using kernel direct LDA (KDDA). They experimentally showed that five individuals with different poses could be linearly separated in feature space while these images are overlapped in input space. Moreover, KDDA is outperformed than that of KPCA and general dis-Manuscript received June 14, 2004; revised October 25, 2004. This work was criminant analysis (GDA) [12]. For the second limitation, the S3 problem occurs when the number of training samples is smaller than the dimensionality of feature vector. In turn, the within-class scatter matrix will W.-S. Chen is with the Department of Mathematics, Shenzhen University, become singular. Some algorithms, such as PCA + LDA [3], Sneuznen 318050, Chana (e-mail; chenws@szu.edu.cn).

P. C. Yuen and J. Huang are with the Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong (e-mail: (RDA) method [15], [16], have been developed to solve the S3 problem. However, PCA+LDA and Direct LDA are designed in the subfeature-space, not in the full feature space. So it

may loss some useful discriminant information. Dai et al. [15]