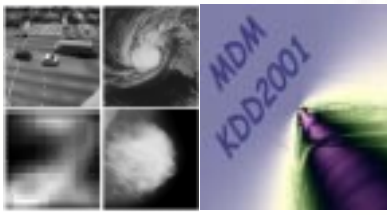
Second International Workshop on

Multimedia Data Mining

August 26th 2001
San Francisco, CA, USA

Edited by
Osmar R. Zaïane
Simeon J. Simoff



MDM/KDD 2001 in conjunction with
Seventh ACM SIGKDD International Conference on Knowledge Discovery & Data Mining

Proceedings

Second International Workshop on
Multimedia Data Mining

MDM/KDD'2001

Proceedings

Second International Workshop on
Multimedia Data Mining

MDM/KDD'2001

August 26th 2001
San Francisco, California, USA

In conjunction with

ACM SIGKDD
Seventh International Conference on
Knowledge Discovery and Data Mining

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001 (O.R. Zaï ane, S. J. Simoff, eds.)

The official workshop web site is:
http://db.cs.ualberta.ca/mdm_kdd2001/

An electronic version of the proceedings will be archived after the workshop at the ACM digital library archive site:
http://www.acm.org/sigkdd/proceedings/mdmkdd01/

Cover art production by Osmar R. Zaïane
Proceedings printed in Canada by Quality Color Press Inc. Edmonton

Foreword

Advances in multimedia delivery technologies have fuelled the rapid growth of research and development in multimedia computing. As emerging multimedia technologies set higher performance levels at competitive costs, they are starting to enable and proliferate intelligent multimedia solutions in a spectrum of commercial and laboratory projects. Such intelligent solutions are usually based on some data analysis techniques. The presentations at the 1st International Workshop on Multimedia Data Mining (MDM/KDD2000) held in conjunction with the Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining in Boston, Massachusetts, USA on August 20th, 2000 revealed that many researchers and developers in the areas of multimedia information systems and digital media turn to data mining and knowledge discovery methods for techniques that can improve indexing and retrieval in digital media (for more details about this workshop see Simoff, S. and Zaï ane, O. "Report on MDM/KDD2000: The 1st International workshop on multimedia data mining", SIGKDD Explorations, 2 (2), 2000, pp. 103-105). During the discussion at the end of the workshop participants identified that there is a need for

- Development and application of specific methods, techniques and tools for multimedia data mining; and
- Frameworks that provide consistent methodology for multimedia data analysis and integration of discovered knowledge back in the system where it can be utilized.

Consequently, the papers selected for presentation at the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001) held in conjunction with the 7th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining in San Francisco, CA, USA on August 26th, 2001, are grouped in the following streams:

- Frameworks for Multimedia Mining
- Multimedia Mining for Information Retrieval
- Applications of Multimedia Mining

The grouping reflects also the aim of the workshop - to bring together experts in analysis of digital media data, state-of-the-art data mining and knowledge discovery techniques, multimedia information retrieval, and knowledge engineers and domain experts from various applied disciplines with potential in multimedia data mining. The papers are of particular interest to the broad community of researchers in intelligent information technologies.

We would like to thank all those, who supported this year's efforts on all stages - from the development and submission of the workshop proposal to the preparation of the final program and proceedings. There were papers submitted from 10 different countries: Australia, Canada, China, France, Germany, India, Japan, Singapore, United Kingdom, and United States of America. All papers were extensively reviewed by at least three referees drawn from the program committee. Special thanks go to them for the final quality of selected papers depends on their efforts.

Osmar R. Zaï ane & Simeon J. Simoff
July 2001

Table of Contents

Workshop Chairs:

- Osmar R. Zaï ane, University of Alberta, Canada
- Simeon J. Simoff, University of Technology-Sydney, Australia

Program Committee

- Terry Caelli, University of Alberta, Canada
- Chabane Djeraba , University of Nantes, France
- Chitra Dorai, IBM Thomas J. Watson Research Center, USA
- Alex Duffy, University of Strathclyde, UK
- Max J. Egenhofer, University of Maine, USA
- William Grosky, Wayne State University, USA
- Howard J. Hamilton, University of Regina, Canada
- Jiawei Han, Simon Fraser University, Canada
- Alexander G. Hauptmann, Carnegie Mellon University, USA
- Wynne Hsu, National University of Singapore, Singapore
- Odej Kao, Technical University of Clausthal, Germany
- Nik Kasabov, University of Ottago, New Zealand
- Paul Kennedy, University of Technology-Sydney, Australia
- Latifur Khan, University of Texas, USA
- Flip Korn, AT&T Laboratories, USA
- Brian Lovell, University of Queensland, Australia
- Mark Maybury, MITRE Corporation
- Mario Nascimento, University of Alberta, Canada
- Gholamreza Nakhaeizadeh, DaimlerChrysler, Germany
- Monique Noirhomme-Fraiture, Institut d'Informatique, FUNDP, Belgium
- Vincent Oria, New Jersey Institute of Technology, USA
- Jian Pei, Simon Fraser University, Canada
- Simone Santini, University of California San Diego, USA
- Simeon J. Simoff, University of Technology-Sydney, Australia
- John R. Smith, IBM Thomas J. Watson Research Center, USA
- Duminda Wijesekera, George Mason University, USA
- Ian H. Written, University of Waikato, New Zealand
- Osmar R. Zaï ane, University of Alberta, Canada

Program for MDM/KDD2001 Workshop

Sunday, August 26, 2001, San Francisco, CA, USA

9:00 - 9:10 Opening and Welcome

9:10 - 10:00 Session 1 (Frameworks for Multimedia Mining)

- 09:10- 09:35 Image Mining: Issues, Frameworks and Techniques
  Ji Zhang, Wynne Hsu and Mong Li Lee
- 09:35-10:00 Multimedia Mining of Collaborative Virtual Workspaces: An Integrative Framework for Extracting and Integrating Collaborative Process Knowledge
  Simeon J. Simoff and Robert P. Biuk-Aghai

10:00 - 10:30 Coffee break

10:30 - 12:00 Session 2 (Multimedia Mining for Information Retrieval)

- 10:30-10:50 The PERSEUS Project: Creating Personalized Multimedia News Portal
  Victor Kulesh, Valery A. Petrushin and Ishwar K. Sethi
- 10:50-11:10 Automatic Feature Mining for Personalized Digital Image Retrieval
  Kyoung-Mi Lee and W. Nick Street
- 11:10-11:30 Rule and Visual Content-Based Indexing
  Chabane Djeraba
- 11:30-11:50 Semantic Content-Based Retrieval in a Video Database
  Pramod K. Singh and A.K. Majumdar

12:00 - 13:30 Lunch

13:30 - 15:30 Session 3 (Applications of Multimedia Mining)

- 13:30-13:50 An Interactive Environment for Kansei Data Mining
  Nadia Bianchi-Berthouze
- 13:50-14:10 Data Mining for Typhoon Image Collection
  Asanobu Kitamoto
- 14:10-14:30 Multimedia Data Mining for Traffic Video Sequences
  Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang and Jeff Strickrott
- 14:30-14:50 A Bayesian Learning Algorithm of Discrete Variables for Automatically Mining Irregular Features of Pattern Images
  Hanchuan Peng and Fuhui Long
- 14:50-15:10 Application of Data Mining Techniques for Medical Image Classification
  Maria-Luiza Antonie, Osmar R. Zaï ane and Alexandru Coman
- 15:10-15:30 A Computer-aided Visual Exploration System for Knowledge Discovery from Images
  Yusuke Uehara, Susumu Endo, Shuichi Shiitani, Daiki Masumoto and Shigemi Nagata

15:30 - 16:00 Discussion and Closure

16:30 Opening of SIGKDD 2001 Conference

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

IMAGE MINING: ISSUES, FRAMEWORKS AND TECHNIQUES

Ji Zhang Wynne Hsu Mong Li Lee
Department of Computer Science, School of Computing
National University of Singapore
Singapore, 117543

{zhangji, whsu, leeml}@comp.nus.edu.sg

Abstract

Advances in image acquisition and storage technology have led to tremendous growth in significantly large and detailed image databases. These images, if analyzed, can reveal useful information to the human users. Image mining deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the images. Image mining is more than just an extension of data mining to image domain. It is an interdisciplinary endeavor that draws upon expertise in computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence. Despite the development of many applications and algorithms in the individual research fields cited above, research in image mining is still in its infancy. In this paper, we will examine the research issues in image mining, current developments in image mining, particularly, image mining frameworks, state-of-the-art techniques and systems. We will also identify some future research directions for image mining at the end of this paper.

Keywords: Image mining, image indexing and retrieval, object recognition, image classification, image clustering, association rule mining.

1. Introduction

Advances in image acquisition and storage technology have led to tremendous growth in significantly large and detailed image databases [36]. The World Wide Web is regarded as the largest global image repository. An extremely large number of image data such as satellite images, medical images, and digital photographs are generated every day. These images, if analyzed, can reveal useful information to the human users. Unfortunately, there is a lack of effective tools for searching and finding useful patterns from these images. Image mining systems that can automatically extract semantically meaningful information (knowledge) from image data are increasingly in demand. The fundamental challenge in image mining is to determine how low-level, pixel representation contained in a raw image or image sequence can be efficiently and effectively processed to identify high-level spatial objects and relationships. In other words, *image mining deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the image databases.* It is an interdisciplinary endeavor that essentially draws upon expertise in computer vision, image processing, image retrieval, data mining, machine learning, database, and artificial intelligence [1]. While some of the individual fields in themselves may be quite matured, image mining, to date, is just a growing research focus and is still at an experimental stage. The main obstacle to rapid progress in image mining research is the lack of understanding of the research issues involved in image mining. Many researchers have the wrong impression that image mining is just a simple extension of data mining applications; while others view image mining as another name for pattern recognition. In this paper, we attempt to identify the unique research issues in image mining. This will be followed by a review of what are currently happening in the field of image mining, particularly, image mining frameworks, state-of-the-art techniques and systems. We will also identify possible research directions to bring image mining research to a new height.

The rest of the paper is organized as follows. Section 2 will discuss research issues that are unique to image mining. Section 3 discusses two possible frameworks for image mining: the functionality framework versus the information-driven framework. Section 4 gives an overview of the major image mining approaches and techniques used in image mining including object recognition, image indexing and retrieval, image classification and clustering, association rules mining, and neural networks. Finally, section 5 concludes with some future research directions for image mining.

2. Research issues in image mining

By definition, image mining deals with the extraction of image patterns from a large collection of images. Clearly, image mining is different from low-level computer vision and image processing techniques because the focus of image mining is in extraction of patterns from *large* collection of images, whereas the focus of computer vision and image processing techniques is in

understanding and/or extracting specific features from a *single* image. While there seems to be some overlaps between image mining and content-based retrieval (both are dealing with large collection of images), image mining goes beyond the problem of retrieving relevant images. In image mining, the goal is the discovery of image patterns that are significant in a given collection of images.

Perhaps, the most common misconception of image mining is that image mining is nothing more than just applying existing data mining algorithms on images. This is certainly not true because there are important differences between relational databases versus image databases.

(a) Absolute versus relative values.
In relational databases, the data values are semantically meaningful. For example, age is 35 is well understood. However, in image databases, the data values themselves may not be significant unless the context supports them. For example, a grey scale value of 46 could appear darker than a grey scale value of 87 if the surrounding context pixels values are all very bright.

(b) Spatial information (Independent versus dependent position)
Another important difference between relational databases and image databases is that the implicit spatial information is critical for interpretation of image contents but there is no such requirement in relational databases. As a result, image miners try to overcome this problem by extracting position-independent features from images first before attempting to mine useful patterns from the images.

(c) Unique versus multiple interpretations.
A third important difference deals with image characteristics of having multiple interpretations for the same visual patterns. The traditional data mining algorithm of associating a pattern to a class (interpretation) will not work well here. A new class of discovery algorithms is needed to cater to the special needs in mining useful patterns from images.

In addition to the need for new discovery algorithms for mining patterns from image data, a number of other related research issues also need to be resolved. For instance, for the discovered image pattern to be meaningful, they must be presented visually to the users. This translates to following issues:

(a) Image pattern representation.
How can we represent the image pattern such that the contextual information, spatial information, and important image characteristics are retained in the representation scheme?

(b) Image features selection.
Which are the important image features to be used in the mining process so that the discovered patterns are meaningful visually?

(c) Image pattern visualization.
How to present the mined patterns to the user in a visually-rich environment?

3. Image mining frameworks

Early work in image mining has focused on developing a suitable framework to perform the task of image mining. An image database containing raw image data cannot be directly used for mining purposes. Raw image data has to be first processed to generate the information usable for high-level mining modules. An image mining system is often complicated because it requires the application of an aggregation of techniques ranging from image retrieval and indexing schemes to data mining and pattern recognition. A good image mining system is expected to provide users with an effective access into the image repository and generation of knowledge and patterns underneath the images. To this end, such a system typically encompasses the following functions: image storage, image processing, feature extraction, image indexing and retrieval, patterns and knowledge discovery.

At present, we can distinguish two kinds of frameworks used to characterize image mining systems: function-driven versus information-driven image mining frameworks. The former focuses on the functionalities of different component modules to organize image mining systems while the latter is designed as a hierarchical structure with special emphasis on the information needs at various levels in the hierarchy.

3.1 Function-Driven Frameworks

The majority of existing image mining system architectures [8, 36] fall under the function-driven image mining framework. These descriptions are exclusively application-oriented and the framework was organized according to the module functionality. For example, Mihai Datcu and Klaus Seidel [8] propose an intelligent satellite mining system that comprises two modules:

(a) A data acquisition, preprocessing and archiving system which is responsible for the extraction of image information, storage of raw images, and retrieval of image.

(b) An image mining system, which enables the users to explore image meaning and detect relevant events.

Figure 1 shows this satellite mining system architecture.

Similarly, the MultiMediaMiner [36] comprises four major components:

(a) Image excavator for the extraction of images and videos from multimedia repository.
(b) A preprocessor for the extraction of image features and storing precomputed data in a database.
(c) A search kernel for matching queries with image and video features in the database.
(d) The discovery modules (characterizer, classifier and associator) exclusively perform image information mining routines to intelligently explore underlying knowledge and patterns within images.

3.2 Information-Driven Frameworks

While the function-driven framework serves the purpose of organizing and clarifying the different roles and tasks to be performed in image mining, it fails to emphasize the different levels of information representation necessary for image data before meaningful mining can take place. Zhang et.al. [18] proposes an information-driven framework that aims to highlight the role of information at various levels of representation. The framework, as shown in Figure 2, distinguishes four levels of information as follows.

(a) Pixel Level, also the lowest level, consists of the raw image information such as image pixels and the primitive image features such as color, texture, and shape;
(b) Object Level deals with object or region information based on the primitive features in the Pixel Level;
(c) Semantic Concept Level takes into consideration domain knowledge to generate high-level semantic concepts from the identified objects and regions;
(d) Pattern and Knowledge Level incorporates domain related alphanumeric data and the semantic concepts obtained from the image data to discover underlying domain patterns and knowledge.

The four information levels can be further generalized to two layers: the Pixel Level and the Object Level form the lower layer, while the Semantic Concept Level and the Pattern and Knowledge Level form the higher layer. The lower layer contains raw and extracted image information and mainly deals with images analysis, processing, and recognition. The higher layer deals with high-level image operations such as semantic concept generation and knowledge discovery from image collection. The information in the higher layer is normally more semantically meaningful in contrast to that in the lower layer.

4. Image mining techniques

Besides investigating suitable frameworks for image mining, early image miners have attempted to use existing techniques to mine for image information. The techniques frequently used include object recognition, image indexing and retrieval, image classification and clustering, association rules mining, and neural network.

4.1 Object Recognition

Object recognition has been an active research focus in field of image processing. Using object models that are known a priori, an object recognition system finds objects in the real world from an image. This is one of the major tasks in the domain of image mining. Automatic machine learning and meaningful information extraction can only be realized when some objects have been identified and recognized by the machine. The object recognition problem can be referred to as a supervised labeling problem based on models of known objects. Specifically, given a target image containing one or more interesting objects and a set of labels corresponding to a set of models known to the system, what object recognition does is to assign correct labels to regions, or a set of regions, in the image. Models of known objects are usually provided by human input a priori.

In general, an object recognition module consists of four components, namely, model database, feature detector, hypothesizer and hypothesis verifier. The model database contains all the models known to the system. The models contain important features that describe the objects. The detected image primitive features in the Pixel Level are used to help the hypothesizer to assign likelihood to the objects in the image. The verifier uses the models to verify the hypothesis and refine the object likelihood. The system finally selects the object with the highest likelihood as the correct object.

Recently, Jeremy S. De Bonet [17], aiming to locate a particular known object in an image or set of images, design a system that processes an image into a set of "characteristic maps". Michael C. Burl et al. [1] pursue an approach to generate recognizers automatically through learning techniques. The domain expert knowledge is

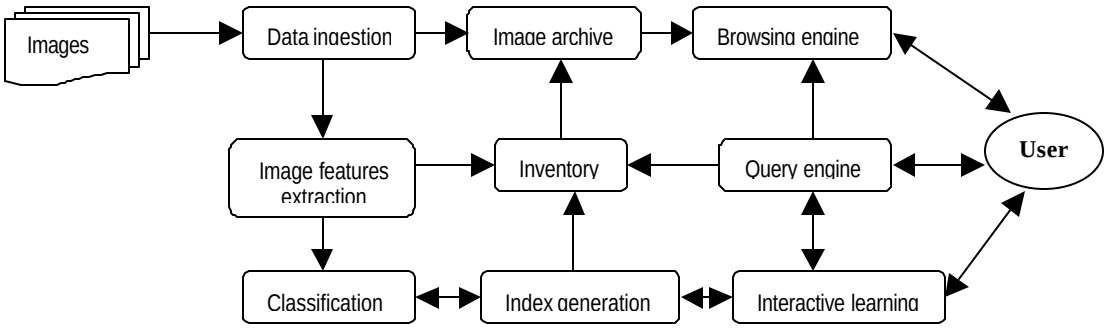


Figure 1. Functionality architecture of an intelligent satellite information mining system.

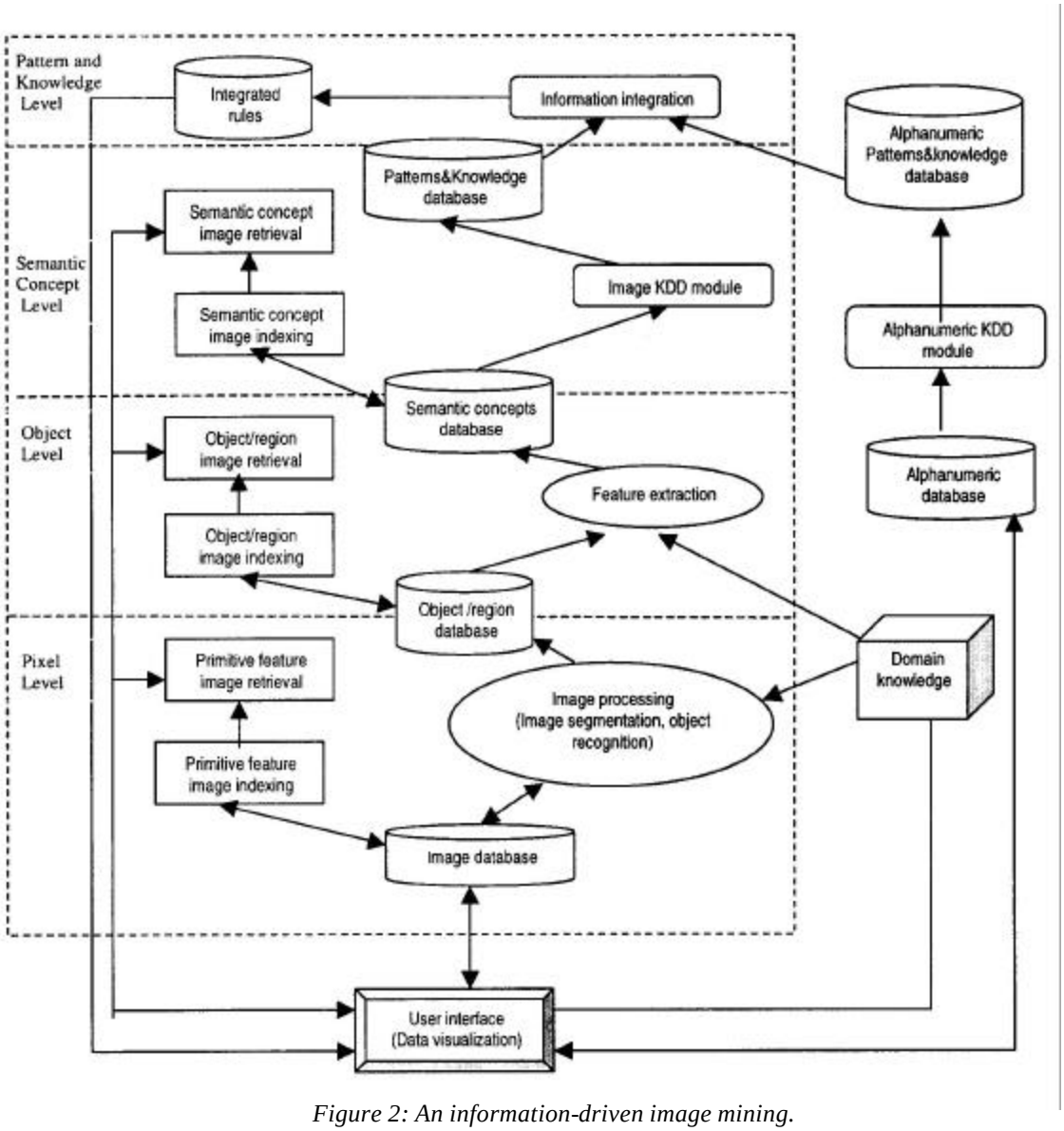


Figure 2: An information-driven image mining.

captured implicitly through a set of labeled examples. Stephen Gibson et al. [13] explore the possibility of finding common pattern in several images, which is an important part of image mining. Stephen Gibson develops and tests an optimal FFT-based mosaicing algorithm that has been shown to work well on all kinds of images.

4.2 Image Retrieval

Image mining requires that images be retrieved according to some requirement specifications. The requirement specifications can be classified into three levels of increasing complexity [1]:

(a) Level 1 comprises image retrieval by primitive features such as color, texture, shape or the spatial location of image elements. Examples of such queries are "Retrieve the images with long thin red objects in the top right-hand corner" and "Retrieve the images containing blue stars arranged in a ring"

(b) Level 2 comprises image retrieval by derived or logical features like objects of a given type or individual objects or persons. Examples include "Retrieve images of round table" and "Retrieve images of Jimmy"

(c) Level 3 comprises image retrieval by abstract attributes, involving a significant amount of high-level reasoning about the meaning or purpose of the objects or scenes depicted. For example, we can have queries such as "Retrieve the images of football match" and "Retrieve the images depicting happiness".

Rick Kazman and John Kominek [20] describe three query schemas for image retrieval: Query by Associate Attributes, Query by Description, and Query by Image Content. In Query by Associate Attributes, only a slight adaptation of conventional table structure is needed to tailor it to fit the image needs. The images are appended as extra field. Image retrieval is performed based on other associated attributes within the same table. In Query by Description, the basic idea is to store image descriptions, also known as labels or keywords, along with each image so that users can locate the images of interest using the descriptions. The image descriptions are normally generated manually and assigned to each image in the image preprocessing stage. It suffers from the drawbacks of the "vocabulary problem" [20] and non-scalability. In the early 1990's, because of the emergence of large-scale image repository, the two difficulties of vocabulary problem and non-scalability faced by the manual annotation approach became more and more acute. Content-based image retrieval is thus proposed to overcome these difficulties. There are three fundamental bases in content-based image retrieval, namely, visual information extraction, image indexing and retrieval system application [28]. Many techniques have been developed in this direction, and many image retrieval systems, both research and commercial, have been built.

In the area of commercial systems, IBM's QBIC system is probably the best known of all image content retrieval systems. It offers retrieval by any combination of color, texture or shape, as well as text keyword. It uses R*-tree indexes to improve search efficiency. More efficient indexing techniques, an improved user interface, and the ability to search grey-level images are incorporates in the latest version. Virage is another well-known commercial system. This is available as a series of independent modules, which system developers can build into their own programs. Excalibur, by virtue of its company's pattern recognition technology, offers a variety of image indexing and matching techniques. As far as the experimental systems, there have been a large number of such systems available. The representatives are Photobook, Chabot, VisualSEEk, MARS, Informedia, Surfimage and Synapse.

4.3 Image Indexing

Image mining systems require a fast and efficient mechanism for the retrieval of image data. Conventional database systems such as relational databases facilitate indexing on primary or secondary key(s). Currently, the retrieval of most image retrieval system is, by nature, similarity-based retrieval. In this case, indexing has to be carried out in the similarity space. One promising approach is to first perform dimension reduction and then use appropriate multi-dimensional indexing techniques that support Non-Euclidean similarity measures [28]. Indexing techniques used range from standard methods such as signature file access method and inverted file access method, to multi-dimensional methods such as K-D-B tree [26], R-tree [10], R*-tree [2] and R+-tree [29], to high-dimensional indexes such as SR-tree [19], TV-tree [21], X-tree [3] and iMinMax [24].

Other proposed indexing schemes focus on specific image features. [24] presents an efficient color indexing scheme for similarity-based retrieval which has a search time that increases logarithmically with the database size. [31] proposes a multi-level R-tree index, called the nested R-trees for retrieving shapes efficiently and effectively. With the proliferation of image retrieval mechanisms, [32] give a performance evaluation of color-spatial retrieval techniques which serves as guidelines to select a suitable technique and design a new technique.

4.4 Image Classification and Image Clustering

Image classification and image clustering are the supervised and unsupervised classification of images into groups respectively. In supervised classification, one is provided with a collection of labeled (pre-classified) images, and the problem is to label newly encountered, unlabeled images. Typically, the given labeled (training) images are used to do the machine learning of the class description which in turn are used to label a new image.

In image clustering, the problem is to group a given collection of unlabeled images into meaningful clusters according to the image content without a priori knowledge [15]. The fundamental objective for carrying out image classification or clustering in image mining is to acquire content information the users are interested in from the image group label associated with the image.

Intelligently classifying image by content is an important way to mine valuable information from large image collection. The classification module in the mining system is usually called classifier. [33] recognizes the challenge that lies in grouping images into semantically meaningful categories based on low-level visual features. Currently, there are two major types of classifiers, the parametric classifier and non-parametric classifier. [7] develops a variety of classifiers to label the pixels in a Landset multispectral scanner image. MM-Classifier, the classification module embedded in the MultiMedia Miner developed by Osmar R.Zaiane et al. [36], classifies multimedia data, including images, based on some provided class labels. James Ze Wang et al. [35] propose IBCOW (Image-based Classification of Objectionable Websites) to classify whether a website is objectionable or benign based on image content. [33] uses binary Bayesian classifier to attempt to perform hierarchical classification of vacation images into indoor and outdoor categories. An unsupervised retraining technique for a maximum likelihood (ML) classifier is presented to allow the existing statistical parameter to be updated whenever a new image lacking the corresponding training set has to be analyzed [4].

Image clustering is usually performed in the early stages of the mining process. Feature attributes that have received most attention for clustering are color, texture and shape. Generally, any of the three, individually or in combination, could be used. There is a wealth of clustering techniques available: hierarchical clustering algorithms, partition-based algorithms, mixture-resolving and mode-seeking algorithms, nearest neighbor clustering, fuzzy clustering and evolutionary clustering approaches. Once the images have been clustered, a domain expert is needed to examine the images of each cluster to label the abstract concepts denoted by the cluster. Edward Chang et al. [4] use clustering technique in an attempt to detect unauthorized image copying on the World Wide Web. [15] uses clustering in a preprocessing stage to identify pattern classes for subsequent supervised classification. Lundervold et al. [15] describe a partition-based clustering algorithm and manual labeling technique to identify material classes of a human head obtained at five different image channels (a five-dimensional feature vector).

4.5 Association Rule Mining

An association rule is an implication of the form $X \rightarrow Y$, where $X, Y \subset I$ and $X \wedge Y = \phi$. I is the set of objects, also referred as items. D is a set of data cases. X is called the antecedent and Y is called the consequent of the rule. A set of items, the antecedent plus the consequent, is call an itemset. The rule $X \rightarrow Y$ has support s in D if s% of the data case in D contains both X and Y, and the rule holds in D with confidence c if c% of the data base in D that support X also Support Y. Association rule mining generate rules that have support and confidence greater than some user specified minimum support and minimum confidence thresholds. A typical association rule mining algorithm works in two steps. The first step finds all large itemsets that meet the minimum support constraint. The second step generates rules from all the large itemsets that satisfy the minimum confidence constraint.

Association rule mining is a typical approach used in data mining domain for uncovering interesting trends, patterns and rules in large datasets. Recently, association rule mining has been applied to large image databases [25,22,36]. There are two main approaches. The first approach is to mine from large collections of images alone and the second approach is to mine from the combined collections of images and associated alphanumeric data [25]. C. Ordonez et al. [25] present an image mining algorithm using blob needed to perform the mining of associations within the context of images. A prototype has been developed in Simon Fraser University called Multimedia Miner [36] where one of its major modules is called MM-Associator. It uses 3-dimensional visualization to explicitly display the associations. In another application, Vasileios M. et al. [22] use association rule mining to discover associations between structures and functions of human brain. An image system called BRAin-Image Database has also been developed. Though the current image association rule mining approaches are far from mature and perfection compared its application in data mining field, this opens up a very promising research direction and vast room for improvement in image association rule mining.

4.6 Neural network

A neural network, by definition, is a massively parallel distributed processor made up of simple processing units, each of which has a natural propensity for storing experiential knowledge and making the knowledge available for use [14]. Neural networks are fault tolerant and are good at pattern recognition and trend prediction. In the case of limited knowledge, artificial neutral network algorithms are frequently used to construct a model of the data.

Even though there has been a lot of research work with regard to neural network and its applications, it is relatively new in the image mining domain. A noteworthy research work that applied neural network to image mining is the Artificial Neural Network (ANN) developed by G.G. Gardner et al [12] which provides a wholly automated approach to fundus image analysis. A Site

Mining Tools, based upon the Fuzzy ARTMAP neural network [6], provides an intuitive means by which an image analyst can efficiently and successfully mine large amounts of multi-sensor imagery for Feature Foundation Data (e.g. roads, rivers, orchards, forests) [30].

5. Conclusions

In this paper, we have highlighted the need for image mining in view of the rapidly growing amounts of image data. We have pointed out the unique characteristics of image databases that brought a whole new set of challenging and interesting research issues to be resolved. In addition, we have also examined two frameworks for image mining: function-driven and information-driven image mining frameworks. We have also discussed techniques that are frequently used in the early works in image mining, namely, object recognition, image retrieval, image indexing, image classification and clustering, association rule mining and neural network.

In summary, image mining is a promising field for research. Image mining research is still in its infancy and many issues remain solved. Specifically, we believe that for image mining research to progress to a new height, the following issues need to be looked at:

(a) Propose new representation schemes for visual patterns that are able to encode sufficient contextual information to allow for meaningful extraction of useful visual characteristics;
(b) Devise efficient content-based image indexing and retrieval techniques to facilitate fast and effective access in large image repository;
(c) Design semantically powerful query languages for image databases;
(d) Explore new discovery techniques that take into account the unique characteristics of image data;
(e) Incorporate new visualization techniques for the visualization of image patterns.

Acknowledgement

This research is supported by the National University of Singapore, Academic Research Fund, RP 991613.

References

[1] M. C. Burl et al. Mining for image content. In *Systemics, Cybernetics, and Informatics / Information Systems: Analysis and Synthesis*, (Orlando, FL), July 1999.

[2] N. Beckmann, H. P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. *In Proc ACM SIGMOD*, 1990.

[3] S. Berchtold, D. A. Keim and H. P. Kriegel. The X-tree: An index structure for high-dimensional data. *In Proceedings of the 22nd VLDB Conference*, pages 28-39, Mumbai, India, September 1996.

[4] L. Bruzzone and D. F. Prieto. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images, *IEEE Transactions on Geoscience and Remote Sensing*, Volume: 39 Issue: 2, pp 456 –460, Feb. 2001.

[5] E. Chang, C. Li and J Wang. Searching Near-Replicas of Image via Clustering. *SPIE Multimedia Storage and Archiving Systems VI*, Boston, MA, USA, 1999.

[6] G. A. Carpenter, S. Grossberg and J. H. Mrkuzon. Fuzzy ARTMAP: a neural architecture for incremental supervised learning of analog multimensional maps, *IEEE Transactions on Neural Networks*, 3(5), 698-713, 1688-1692,1998.

[7] R. F. Cromp and W. J. Campbell. Data mining of multi-dimensional remotely sensed images. *International Conference on Information and Knowledge Management (CIKM)*, 1993.

[8] M. Datcu and K. Seidel. Image information mining: exploration of image content in large archives. *IEEE Conference on Aerospace*, Vol.3, 2000.

[9] J. P. Eakins and M. E. Graham. Content-based image retrieval: a repot to the JISC technology applications program. *Northumbria Image Data Research Institute*, 1999.

[10] A. Guttman. R-tree: a dynamic index structure for spatial searching. *In Proc ACM SIGMOD*, 1984.

[11] D. Greene. An implementation and performance analysis of spatial data access. *In Proc ACM SIGMOD*, 1989.

[12] G. G. Gardner and D. Keating. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. *British Journal of Ophthalmology*, 1996.

[13] Gibson, S et al. Intelligent mining in image databases, with applications to satellite imaging and to web search, *Data Mining and Computational Intelligence*, Springer-Verlag, Berlin, 2001.

[14] S. Haykin. Neural Networks: a comprehensive foundation. *Prentice Hall International, Inc*. 1999.

[15] A. K. Jain, M. N. Murty and P. J. Flynn. Data clustering: a review. *ACM computing survey, Vol.31, No.3*, September 1999.

[16] R. Jain, R. Kasturi and B. G. Schunck. Machine Version. *MIT Press and McGraw-Hill Press*, 1995.

[17] J. S. D. Bonet. Image preprocessing for rapid selection in "Pay attention mode". *MIT*, 2000.

[18] J. Zhang, W. Hsu and M. L. Lee. An Information-driven Framework for Image Mining, *in Proceedings of 12th International Conference on Database and Expert Systems Applications (DEXA)*, Munich, Germany, September 2001.

[19] N.Katayama and S. Satoh. The SR-tree: An index structure for high-dimensional nearest neighbor queries. *In proceedings of the 1997 ACM SIGMOD*

Conference, pages 369-380, Tucson, Arizona, May 1997.

[20] R. Kazman and J. Kominek. Information organization in multimedia resources. *Proceedings of the 11th annual international conference on Systems documentation*, Pages 149 – 162,1993.

[21] K. Lin, H. V. Jagadish and C. Faloutsos. The TV-tree: An index structure for high-dimensional data. *The VLDB Journal*, 3 (4): 517-542,1994.

[22] V. Megalooikonomou, C. Davataikos and E. H. Herskovits. Mining lesion-deficit associations in a brain image database. *KDD*, San Diego, CA USA, 1999.

[23] W. Y. Ma and B. S. Manjunath. A texture thesaurus for browsing large aerial photographs, *Journal of the American Society for Information Science* 49(7), 633-648,1998.

[24] B. C. Ooi, K. L. Tan, C. Yu and S. Bressan. Indexing the Edges - A Simple and Yet Efficient Approach to High-Dimensional Indexing, *19th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 166-174, Dallas, Texas, May 2000.

[25] C. Ordonez and E. Omiecinski. Discovering association rules based on image content. *Proceedings of the IEEE Advances in Digital Libraries Conference (ADL'99)*, 1999.

[26] J. T. Robinson. The K-D-B tree: A search structure for large multidimensional dynamic indexes. *In Proceeding of the 1981 ACM SIGMOD Conference*, pages 10-18, June 1981.

[27] Y. Rui, K. Chakrabarti, S. Mehrotra, Y. X. Zhao, and T. S. Huang. Dynamic clustering for optimal retrieval in high dimensional multimedia databases. *In TR-MARS-10-97*, 1997.

[28] Y. Rui, T. S. Huang et al. Image retrieval: Past, present and future. Invited paper in *Int Symposium on Multimedia Information Processing*, Taipei, Taiwan, Dec 11-13, 1997.

[29] T. Sellis, N. Roussopoulos and C. Faloutsos. The R+ tree: A dynamic index for multi-dimensional objects. *In Proc 12th VLDB*, 1987.

[30] W. Strelilein and A. Waxman. Fused multi-sensor image mining for feature foundation data. *Proceedings of the Third International Conference on Information Fusion (FUSION 2000)*, Volume: 1, 2000.

[31] K. L. Tan, B.C. Ooi and L. F. Thiang. Retrieving similar shapes effectively and efficiently, *Multimedia Tools and Applications*, Kluwer Academic Publishers, The Netherlands, 2001.

[32] K. L Tan, B. C. Ooi and C. Y. Yee. An Evaluation of Color-Spatial Retrieval Techniques for Large Image Databases, *Multimedia Tools and Applications*, Vol. 14, No. 1, pp. 55-78, Kluwer Academic Publishers, The Netherlands, 2001.

[33] A.Vailaya, A. T. Figueiredo, A. K. Jain, and H. J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, Volume: 10 Issue: 1, pp 117 –130, Jan. 2001.

[34] D. White and R. Jain. Similarity indexing: Algorithms and performance. *In Proc. SPIE Storage and Retrieval for image and Video Databases*, 1996.

[35] J. Z. Wang, J. Li et al. System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms. *Interactive Distributed Multimedia Systems and Telecommunication Services, Proceedings of the Fourth European Workshop (IDMS'97)*, 1997.

[36] O. R. Zaiane, J. W. Han et al. Mining MultiMedia Data. *CASCON'98: Meeting of Minds*, pp 83-96, Toronto, Canada, November 1998.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

Multimedia Mining of Collaborative Virtual Workspaces: An Integrative Framework for Extracting and Integrating Collaborative Process Knowledge

Simeon J. Simoff
Faculty of Information Technology
University of Technology, Sydney
P.O. Box 123
Broadway, NSW 2007
Australia
simeon@it.uts.edu.au

Robert P. Biuk-Aghai
Faculty of Science and Technology
University of Macau
P.O. Box 3001
Macau S.A.R.
China
fst.robert@umac.mo

ABSTRACT

Collaborative virtual environments are becoming an intrinsic part of professional practices. In addition to providing collaboration support, they have the potential to collect vast amounts of multimedia data about the actions and content of such collaborative activities. The aim of this research is to utilize this data effectively, extract meaningful insights out of it and feed discovered knowledge back into the environment. The paper presents a framework for integrating multimedia data mining techniques with collaborative virtual environments, starting from early conceptual development. Discovered patterns are deposited in an organisational memory, which makes these available within the virtual environment. Some of the ideas are illustrated by an example from the application to collaborative spaces developed in LiveNet, a virtual workspace design system.

Keywords

Multimedia data mining, knowledge extraction, collaborative virtual environments

1. INTRODUCTION

Collaborative virtual environments (CVE) have become increasingly popular in recent years. As a result, teams in many companies and organisations are conducting their projects via one or another form of such environments. The notion of CVEs span a broad range of distributed systems—from text-based virtual environments (text based MOO/MUD and Web-based WOO environments [15]) to desktop virtual worlds (an example of such environments with different interfaces and information organisation is presented in Figure 1 and Figure 2) and immersive virtual worlds (for an excellent taxonomy of the latter see [5]). There are numerous approaches and techniques for designing such environments to support collaborative projects [14]. These approaches and techniques use different ways of formalising the requirements for the environment and the design of the project workspace.

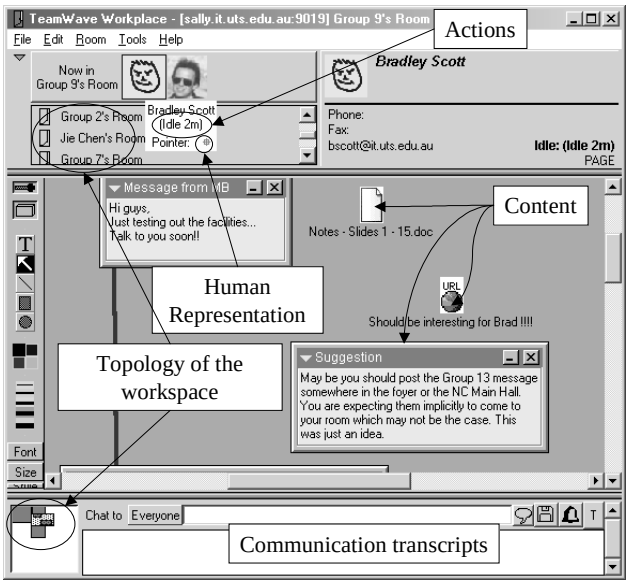


Figure 1. Example of a typical workplace in a collaborative virtual environment with 2D interface

Common design strategies are based on the utilisation of experts' knowledge in a top-down analysis cycle. At present, requirements formalisation methods for such environments are based on two major types of methodologies:

- Domain-dependent formalisation based on an ontology of a particular metaphor, usually coming from the domain in which the environment is expected to be applied. For example, Gutwin and Greenberg [10] recognized the importance of the ontology of a "place". Simoff and Maher [18] presented the architecture of a virtual "place" that follows the ontology of a university.

- Domain-independent formalisation of project (or business process) activities (and their attributes) based on soft-systems methodology [16].

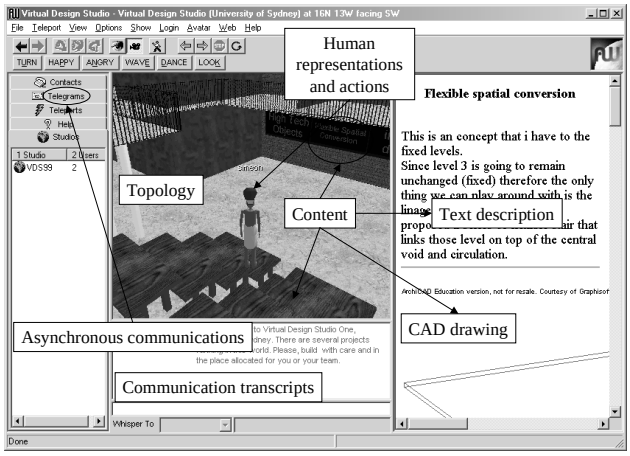


Figure 2. Example of a typical workplace in a collaborative virtual design studio (with 2D/3D interface)

Despite their variety and difference in functionality these environments have several key concepts in common:

- the concept of "*inserting people*" in the collaborative virtual environment, in other words, representing people as some entities, which range from the so-called "characters" or "avatars" in social virtual worlds [15] to sophisticated user profiles in collaborative e-market environments [17];

- the concept of "*structuring the space*" in the collaborative virtual environment, in other words, providing some way of configuring the layout of the space, separating and handling different information within the units of this structure, and some reference system for orientation and navigation;

- the concept of "*media usage*" which defines the feasible set of actions that can be performed in the collaborative virtual environment and the types of electronic media (file formats) that can be used and manipulated in the collaborative workspace [14]. The media usage defines to what extent the environment under consideration can be used for conducting collaborative projects in a particular domain.

1.1 People

Object representations of a person include characteristics such as a text description, messages about their movements in the place, and links to web pages to help establish their identity and personality. An important aspect of people's representation is the variety of "rights" that can be assigned to them. Different environments use different terms for this—privileges, roles, permissions. Thus, the representations are potential sources of preliminary information about a person's individuality. However, in collaborative projects it is important to be able to make judgments about the individual's working preferences and the collaboration styles of the people in a team or in different teams and to reuse such knowledge when forming teams in other collaborative projects. The preliminary information is not always sufficient for establishing successful work. Data mining techniques can be applied for extracting information about the individuals, the functioning of groups of individuals and to discover patterns of collaboration based on project communication between them. This knowledge can be reused for configuring groups in new projects.

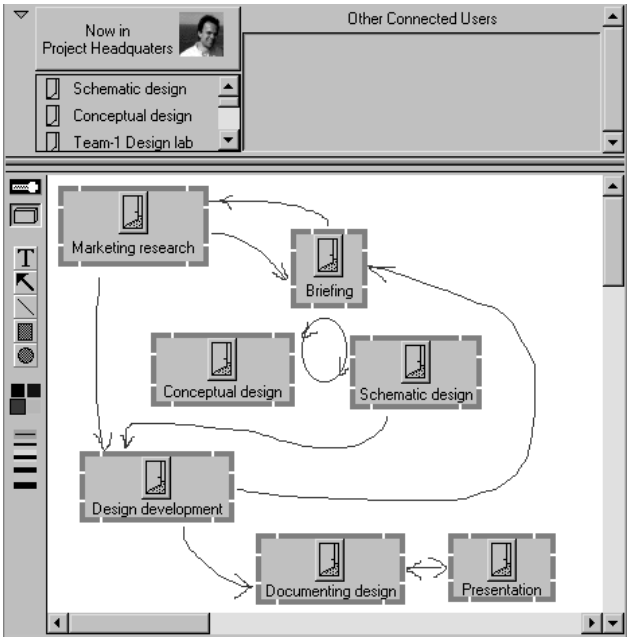


Figure 3. Pre-defined topology of a collaborative virtual environment for design projects

1.2 Space Structuring

The ways of structuring of the environment's space depend on a number of factors, including the ontology (what kind of place the environment is), purpose of the environment, the embedded functionality, the preferred communication and collaboration mode [14], underlying technologies and their integration [18]. Figure 3 illustrates the topology of a design environment, predefined by the model (ontology) of the design process. Each "doorway" leads to a room (workspace), which contains the relevant information for each design stage. The model used is derived from the existing models of the design process in the research literature (e.g. [13]). Further, such schema can be used as a prototypical workspace. However, such top-down approach does not capture the knowledge from the actual use of the virtual environment—which parts of it were used more intensively, what are the

"neighbouring" relations (e.g., co-visited rooms) and other relations, document contents during different phases, etc. Data mining techniques can be applied for discovering such relations. Discovered knowledge can be reflected in variations of the space structuring of the "design prototypes", resulting in building a library of such prototypes and reusing them according to the requirements of the new project space.

1.3 Media Usage

The ontology of the virtual environment can also provide substantial *a priori* knowledge not only about the possible navigation, but also about the set of feasible actions in such an environment. Usually the initial set of actions is derived from the design requirements. The real set of actions used in different projects, however, may vary substantially. The overlapping set of actions forms the common kernel set, and the remainder is the individual component. In the long term, this provides a potential for designing pro-active prototypes supporting different types of projects. Digital media is the basis on which project information can be shared across a network. In collaborative virtual environments the focus is on the following types of digital media: communication and action data, images, CAD models, text, links, audio and video components, and virtual reality representations. Data mining techniques can be applied for composing the action sets. Discovered action sets and space structures will form pro-active design prototypes, resulting in a library of such prototypes and their reuse according to the requirements of a new project. The content mining techniques delivers content features that then are related to the actions.

In this paper we present a framework for integrating multimedia data mining in the design framework of CVEs, in a way that facilitates not only the data collection and analysis, but also the application and integration of discovered knowledge. We illustrate some aspects of the application of the framework for monitoring and extracting knowledge from collaborative activities and incorporating that knowledge back into the collaborative environment.

2. THE FRAMEWORK

The presented framework embeds knowledge discovery in CVEs. Its two primary goals are:

1. To influence the design of CVEs so as to provide the data necessary for mining and analysis of collaboration.

2. To feed extracted knowledge back into the use of CVEs.

As a result, data design and design of the collaborative virtual environment are seen as complementary and parallel activities, offering the opportunity to control data collection to a greater extent. Knowledge obtained from collaboration data is a further contributor to CVE design. A number of related research efforts are underway in the direction of controlled data collection, carried out mainly in the field of e-commerce and Web data mining [19].

The framework for multimedia data mining in CVEs is shown in Figure 4. It includes four major groups of inter-woven components:

1. *Collaborative virtual environments*.

2. *Collaboration data*.

3. *Knowledge discovery*.

4. *Organisational memory*.

Moreover, the three components appearing in the upper part of the figure consist of three parts, at different levels of abstraction:

1. *Conceptual level*.

2. *Structural level*.

3. *Collaboration level*.

Below, we discuss the components of the framework in more detail.

2.1 Collaborative Virtual Environments

CVEs are the support systems within which collaboration is carried out. CVEs are becoming increasingly part of professional practice. Such environments aim to support certain work practices, hence are domain-specific. For each domain, an understanding of the domain-dependent requirements for the CVE has to be obtained.

On the conceptual level this activity identifies the concepts to be supported by the environment: the structuring metaphor employed, navigation facilities, representation of people and their abilities, artefacts and tools provided in the environment, etc. On the structural level, this initial step is followed by the actual design of the CVE when the relationship between the identified concepts is established and their detail is elaborated. Once designed (and implemented), the CVE is utilized by its users on the collaboration level.

2.2 Collaboration Data

The activities related to CVEs are paralleled by those related to collaboration data. Within the presented

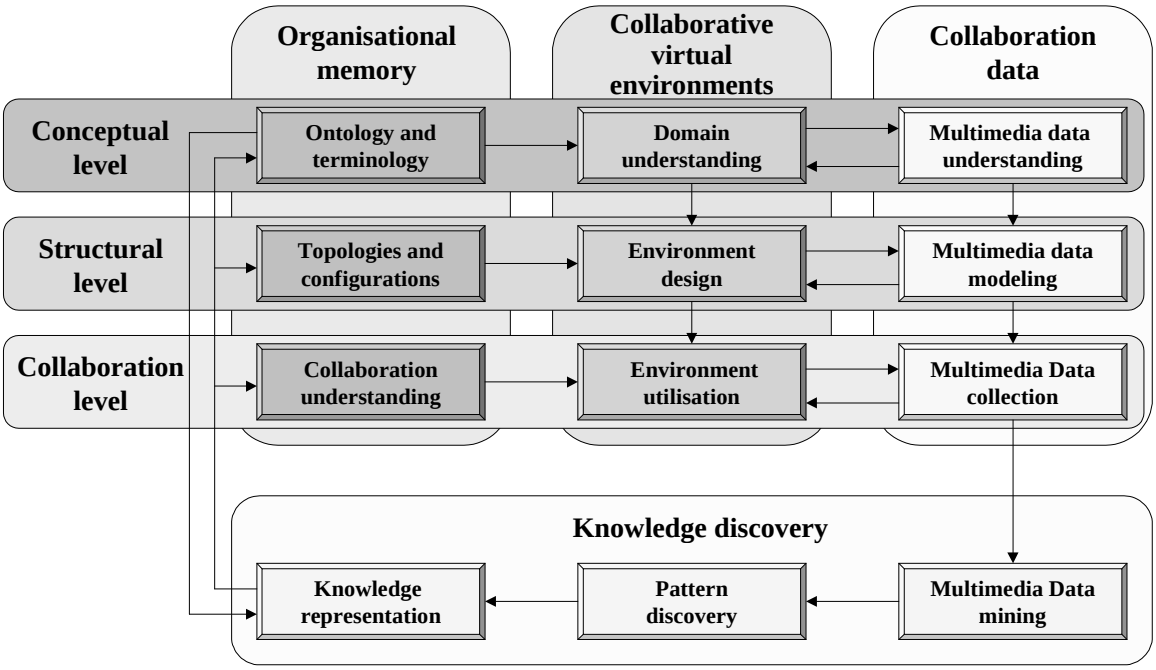


Figure 4. Framework for integrating data mining in the design and application of collaborative virtual environments and knowledge extraction from them

framework, collaboration data is that portion of data which facilitates knowledge discovery within the domain of collaboration, regardless of whether it is of direct use within the CVE. Traditionally, virtual collaboration systems did not provide any particular support for data collection aimed at knowledge discovery. Data was seen as an internal aspect of the system and only internally required data was maintained. The presented framework emphasizes the need for additional data that can enable knowledge discovery in the collaboration domain, and therefore within this framework collaboration data is treated separately from the CVE.

On a conceptual level, domain understanding within the CVE sphere and data understanding within the sphere of collaboration data are mutually complementary: once domain understanding identifies a concept to be supported, data understanding identifies the necessary data elements. Such data elements include fields of log files, the different multimedia document and file formats involved in collaboration, the transcripts from synchronous chat communications, the thread structure and content of asynchronous bulletin board discussions, the audio/video communications, and the integrating data descriptions, which define the relations between the different types of media data.

On the structural level, during environment design, data modeling identifies details of and relationships among the collaboration concepts and data.

Finally, on the collaboration level, the CVE is utilized in an actual project, which generates the collaboration data. These data are collected for subsequent data mining.

2.3 Knowledge Discovery

The knowledge discovery in this framework differs slightly from the classical schema [7]—the selection and data pre-processing stages are implicitly embedded in the data design. Therefore, collected data is expected to be ready for the application of multimedia (or rich media) data mining methods. As a further difference to the classical knowledge discovery schema, a step of knowledge representation is explicitly included at the end. Its purpose is to map discovered knowledge back into the CVE's representation.

Knowledge discovery in this framework aims to produce a better understanding of computer-mediated collaboration, and to enable the usage of discovered knowledge to improve structural features of the workspace configuration and media content when new projects are conducted in the environment. For example, through the analysis of the structuring of virtual environments, templates of structures of these environments can be collected, implying certain navigation behaviour. Collecting data about actual navigation within the environment can provide a source for discovering traversal patterns, which can provide indicators for improving the topology (structuring) of the environment.

Other possibilities for improvement of the environment exist according to particular collaboration and business

process needs. This is something difficult to know ahead of time. The development of such environments follows emergent and adaptive strategies, rather then predefined topologies. In both cases, some necessary indicators for improvement of the structure are required.

2.4 Organisational Memory

Over the past decade, the CSCW community and related areas have taken a keen interest in organisational memory (OM) [1,3,6]. This suggests that there is value in retaining and later drawing on historical records of virtual collaboration. Such records may be referenced when setting out on new virtual collaboration, to "see how others have done it", and perhaps to reuse and re-enact parts of those collaboration instances. Unlike conventional work settings where details of collaboration have to be collected manually through effort-intensive and sometimes intrusive methods, CVEs are an ideal source of data on collaboration, particularly when work is predominantly or entirely carried out virtually, as such environments can automatically record a great amount of detail on the collaboration.

While much work in organisational memory concerns itself with the content of collaboration, or the *declarative memory*, little work has been done on harnessing the *procedural memory*, or knowledge about how work has been carried out. The importance of utilising this aspect of organisational memory in groupware systems has been pointed out relatively early on [6], and again more recently within the context of virtual team effectiveness [8]. The presented framework makes the procedural portion of organisational memory an integral part of the collaboration support environment by maintaining knowledge extracted from collaboration environments and making it available within the environment.

On the collaboration level, this knowledge relates to an understanding of the collaboration. For example, it can identify what main types of activities were conducted within a virtual environment, how the activities were carried out over time, what differences exist in the activity of different people within the environment, etc. This knowledge can be utilized within the environment itself, leading for instance to an adaptation of the environment itself and/or its interface in order to facilitate the execution of predominant activities. It can also serve as a management and control instrument, which is of particular value when collaboration is completely virtual and traditional management methods are severely limited.

On the structural level, representations of the environment's topology are maintained. Where structural patterns are discovered in a set of environments, this too is deposited in the organisational memory in the form of different topologies and configurations available for reuse. Such information may feed back into environments in use, for instance to rearrange the environment's topology if its current arrangement is discovered to encumber work.

Use of CVEs may, over time, also lead to the emergence of new concepts, or an application of existing concepts in ways that were not previously anticipated. These are deposited on the conceptual level as modifications to the underlying ontology, and feed into the ongoing development of a CVE. An example of this is where an environment lacks a certain feature, but where users discover workarounds that, though cumbersome, allow the feature to be supported. Discovery of such cases can be of use in the development of the next version of the CVE to explicitly support the feature.

In the next section we present an example of the application of the framework for extracting knowledge about different aspects of computer-mediated collaboration.

3. EXTRACTING KNOWLEDGE FROM COLLABORATIVE ACTIVITIES

The example of the application of the framework is related to a particular system for developing collaborative virtual workspaces—LiveNet. The framework was particularly applied in the areas of CVEs, collaboration data, and knowledge discovery. We start by introducing the LiveNet system, then show how the framework was applied.

3.1 LiveNet as a Collaborative Virtual Environment

LiveNet is a virtual workspace design system, developed at the University of Technology, Sydney [12]. It supports mainly asynchronous collaboration of distributed groups of people, i.e. different-time, different-place interactions, although its design does not limit it from other modes of collaboration. A central server is accessed across the network through one of several client interfaces, most commonly through a Web interface (as shown in Figure 5). LiveNet provides virtual workspaces which bring together people, artefacts (e.g. multimedia documents), communication channels, awareness facilities, and a collection of tools, all tied together through a configurable governance structure. A simplified ontology of the basic CVE design unit in LiveNet is shown in Figure 6. In terms of the ontology, workspaces contain roles, occupied by participants (i.e. actual people), who perform actions. Some actions may operate on artefacts, others may be interactions with other workspace participants through discussions. However, most workspace elements such as documents, discussions and participants, may be shared between workspaces. Thus workspaces are not just stand-alone entities (units) but nodes in a network of inter-connected collaboration spaces. Neither are structures of workspaces in LiveNet collaboration space static—once

created, a workspace can be dynamically adapted to evolve—incorporate work subspaces, or attach to other workspaces, together with the project collaboration carried out in it. As a result entire "ecologies" of inter-connected workspaces can co-evolve. Mining the multimedia data in such evolving workspaces can provide invaluable insights about the development of such "ecologies".

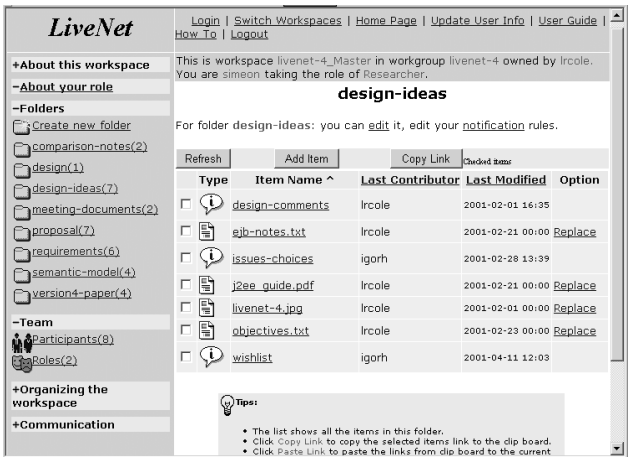


Figure 5. Typical LiveNet screen (web interface)

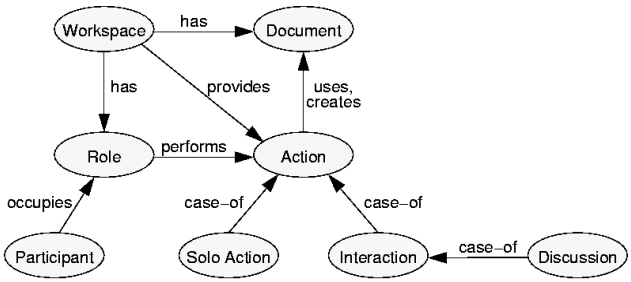


Figure 6. Simplified ontology of LiveNet

3.2 Integrating Collaboration Data

Early versions of LiveNet had not been developed with the support of data mining and knowledge extraction activities in mind. Consequently, only a limited amount of collaboration data was available, namely only data which was necessary for the internal operation of the system. While this allowed some structural-level features to be extracted, there was no data to support the extraction of collaboration-level or conceptual-level features.

Subsequently, the provision of suitable collaboration data was "retrofitted" onto an earlier version of LiveNet. Building on the existing domain understanding, conceptual data requirements were developed, followed by data modeling. These were integrated into LiveNet by appropriately adapting its design (and implementation)—corresponding to the flow from data modeling to environment design in our framework. Finally, collection of the new collaboration data and knowledge discovery followed.

The first iteration of this cycle lead to some knowledge extraction, both on the collaboration and the structural levels. However, analysis of the collaboration also revealed that certain data elements, which were not captured at the time would be needed to provide a more complete picture of the collaboration. This had not been accounted for in the first cycle of integration of collaboration data. Consequently, a second cycle was initiated in which data understanding and data modeling were refined, and environment design was brought up-to-date with the new data model. The subsequent data collection and data mining lead to a more comprehensive analysis of collaboration and a richer knowledge discovery. Following this second cycle, new data requirements are already emerging which, once implemented, will lead to a yet richer body of collaboration data. This confirms to us the validity of our framework in feeding discovered knowledge back into the ongoing development of the collaborative environment. It also highlights the fact that this is likely not achieved in a single effort, but is an iterative process, with insight from each iteration triggering a new iteration.

3.3 Knowledge Discovery in LiveNet

Collaboration data in LiveNet consists of two parts: a database contains the internal data of the CVE, maintaining the current state of all workspace elements (documents, roles, participants, etc.). The second part is a set of log files that are external to the system itself and which record all user actions carried out in the system over time. Although the vast majority of users interact with LiveNet through a web interface, the log records captured by the LiveNet server are on a semantically much higher level than those in the corresponding web access log. While a web log includes IP addresses, document names, timestamps and http request types, the LiveNet log records information in terms of the LiveNet CVE's conceptual model. Thus every record includes the name of the workspace and its owner, the name of the participant carrying out the action, his/her role name, the LiveNet server command requested, etc. This allows analysis to exploit metadata available in the application and to capture higher-level actions than a mere web log does (this corresponds to the approach of [2]).

The analysis we carried out focused primarily on the log of collaboration actions, and to a lesser extent on the workspace database. It involved pre-processing of the log, visualization of workspace data, and actual data mining. The pre-processing step normalizes session numbers, aggregates lower-level events into higher-level actions, and calculates session summaries. In this context, a session is the sequence of actions carried out by a user from login to logout time. Data pre-processing is

considered part of collaboration data collection and is usually automatically performed.

The data used originated from students and instructors of a number of courses at the University of Technology, Sydney, who used the LiveNet system both to coordinate their work, and to set up workspaces as part of the students' assignments. The data covers a three month period, with a total of 571,319 log records, They were aggregated into 178,488 higher-level actions in a total of 24,628 sessions involving 721 workspaces and 513 users.

3.4 Space Structuring

During knowledge discovery, using visualization certain of the relationships existing within and between workspaces can be discovered. This particularly aids exploratory analysis, when the purpose is to get an understanding of the structure of, and patterns in, the data. We selected data originating from students of one course who used LiveNet during the mentioned period. There were a total of 187 student users, organised into 50 mostly 3-5 person groups, whose use accounted for about 20% of the above-mentioned log data.

Initial visualization focused on networks of workspaces, to discover how individual student groups partitioned their work in terms of distinct workspaces, and to what extent these workspaces were linked to one another. This exploratory analysis revealed two distinct patterns: the majority of users preferred to use just one workspace to organise all their course work (such as posting drafts of assignment documents, discussing work distribution and problems, etc.). This workspace tended to contain many objects—or have a high *absolute workspace density* [4]. We term such groups *centralizers*. On the other hand, a few groups tended to partition their work across a collection of connected workspaces, usually with a separate workspace for each major course assignment. These workspaces tended to contain fewer objects (having a lower absolute workspace density) than the ones of the centralizers. We term these groups *partitioners*.

Figure 7 shows a map of LiveNet workspaces with colours highlighting absolute workspace density—lighter colour indicating lower density, darker colour indicating higher density. Branching out from the central node at the top are networks of workspaces for three groups. Nodes represent workspaces, edges represent hierarchical relationships between workspaces. What the map reveals is that the group on the right, Team40, has a very high density in the workspace used for facilitating its work (the workspace Team40_Master). Moreover, it uses only one workspace for this purpose. Thus the right group is a typical example of a centralizer. On the other hand, workspaces in the group at the centre have a much lower density. Out of the eight workspaces in this group, six are used for facilitating aspects of the group's work. This is indicative of a partitioner group.

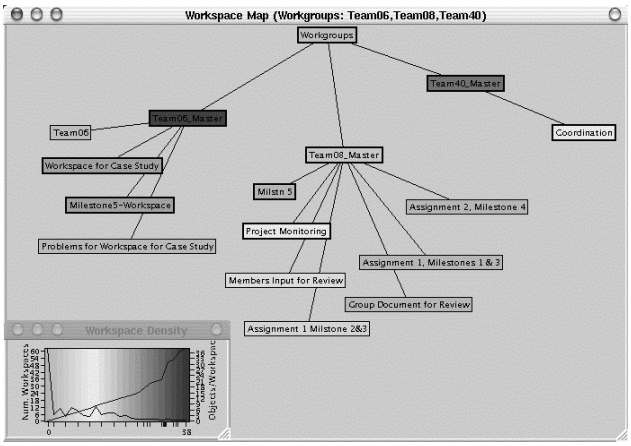


Figure 7. Workspace densities of three different groups

There are plausible explanations for both the centralizer and partitioner cases. Both approaches have their own advantages: in the centralizer case, it is convenience in not having to create multiple workspaces, to switch between them, and in addition to have everything available to all participants in a single location. In the partitioner case, the advantage is increased clarity, structuring according to task, and consequently reduced cognitive load. Furthermore, some groups may bring certain preferences as to the way to organise their work into workspaces and enact these preferences in the way they structure their virtual working environment. When such preferences are recognized during knowledge discovery, and deposited in the organisational memory, they can feed back into the design of new virtual collaboration environments, thus helping to offer more adequate support to cooperative groups with diverse working styles.

3.5 Media Usage

A further area we investigated was focused on identifying which actions different groups mainly carried out within LiveNet. All in all, 80 different actions are available in LiveNet. The majority of student groups used only about half of these. The major actions carried out are related to the main LiveNet conceptual elements: workspaces, roles, participants, documents, and discussions. A taxonomy of these actions is shown in Figure 8.

While all groups had been given the same task—to prepare a number of assignments and to set up a collection of workspaces to support a given process—the way they implemented this task varied markedly. This was evident in a number of aspects of their use of the LiveNet system, such as intensity of use, number of workspaces created, number and length of sessions, number of actions per session, etc. One area of our analysis focused on the proportional distribution of main actions. This revealed

that strong differences existed among different groups. To illustrate two examples, Figure 9 shows action distributions among the major high-level actions of the taxonomy of Figure 8 for one group whose distribution of actions was fairly even across categories (with the exception of the participant category): the five major action categories did not vary greatly, none of them exceeding 0.29 of the total (circle size signifies proportion out of the total).

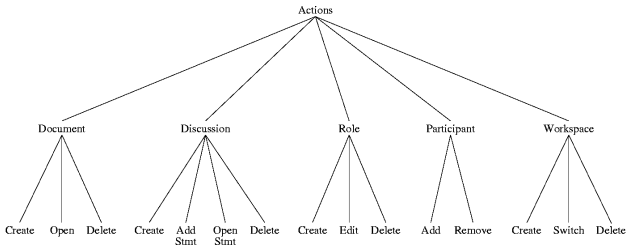


Figure 8. Taxonomy of major high-level LiveNet actions

Figure 10, on the other hand, shows a highly uneven distribution of actions in another group, where one action category (role) strongly dominates with 0.56 of the total, and two other action categories (document and discussion) barely register.

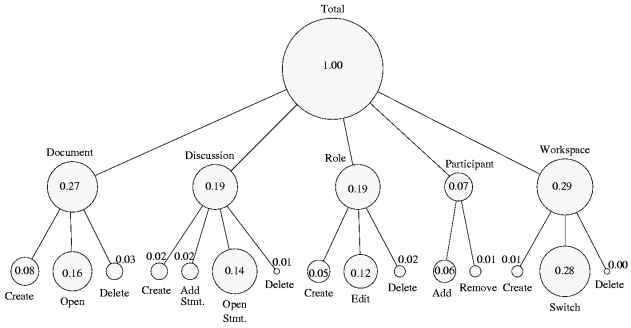


Figure 9. Relatively even distribution of actions in group 1

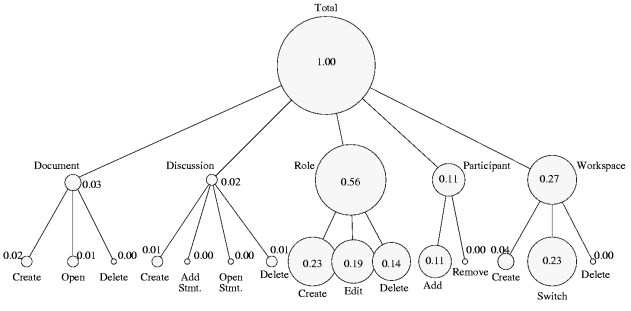


Figure 10. Highly uneven distribution of actions in group 50

This difference may be explained when considering that group 1 (Figure 9) had a total of 627 sessions consisting of a total of 7446 actions, while group 50 (Figure 10) had only 36 sessions and 633 actions. Not only did group 1 use LiveNet much more intensively, but they also made much greater use of the system to facilitate their own work (as manifested in the solid proportion of actions in the document and discussion categories). Thus the skew in action distribution towards role-related actions on the part of group 50 is caused by the under-utilization of other LiveNet features, not by an absolute high number of actions related to roles (in absolute terms, group 1 carried out 431 role-related actions, while group 50 carried out only 142 such actions). It should be noted that the choice of these two groups for illustration was not coincidental: group 1 was the best-performing group in the course, while group 50 was the worst-performing group, as measured in the marks obtained for their assignments in the course, one of which involved heavy use of LiveNet. The situation was comparable in other similarly scoring groups.

When such cases are identified and included in the organisational memory as part of a record of collaboration, they can be of use in evaluating virtual work. This can be particularly useful with fully virtual teams that never meet face-to-face, where conventional management methods for project monitoring and control are severely limited or absent. The organisational memory thus takes on the additional role of a management instrument.

The presented knowledge extraction related to media usage has already yielded interesting results through the application of relatively simple data analysis methods. As this research continues, we plan to more fully exercise a range of data mining methods on the available body of data to obtain further insights. For instance, clustering may assist us in identifying patterns of behaviour in different user groups, which can help categorize the usage of different collaboration spaces used for different purposes. It may also aid in the construction of "group profiles" which allows more personalized support to be provided.

4. INTEGRATING EXTRACTED KNOWLEDGE IN THE ORGANISATIONAL MEMORY

An important part of the framework is the way knowledge is returned back to the environment. An example of such feedback is the collection and generalisation over the workspace graph structure. The procedure to some extent is similar to building a case base of workspace configurations. Case indexing and retrieval is based on matching graph structures. The new collaborative process is formalised into a graph structure using concepts from a modified form of the soft systems methodology [11], with activities, roles and artefacts as node types and particular rules for the connections between the nodes (for example, a participant and an artefact cannot have a direct

connection). The formal representation is usually a result of a high-level (i.e. not detailed) description of the process. This representation is matched against the graph representation of the workspace configurations. Retrieved cases provide the initial configuration for further adaptation.

The framework also allows a feedback from the organisational memory towards modification of the knowledge representation schema, used for representation and incorporation of discovered knowledge. The detailed discussion of the issues related to the modification of the knowledge representation schema, however, are beyond the scope of this paper.

5. CONCLUSIONS

CVEs can provide researchers with enormous amounts of data about various aspects of computer-mediated collaboration. Unfortunately, the design of earlier environments did not pay much attention to the issues of data collection [9]. Thus, the application of multimedia data mining methods had to struggle with translating data collected for other purposes, for example, a server log used usually for correct recovery after a failure, into data useful for the goals of data mining. Consequently, the earlier application of multimedia data mining methods in CVEs has been focused mainly on the analysis of communication transcripts—whether recorded in synchronous collaborative sessions or over a bulletin board in asynchronous mode, and over project document content.

The framework presented in the paper looks at the integration of data mining technologies in CVEs at the early design stages of the virtual environment. A key issue at the design stage is the selection of the data that should be recorded. These records are complementary to the standard logs of the web server. They include activity log data, and the dynamics of media usage and changes in workspace content. Careful design and analysis of the activity log data have the potential to lead to improvements of the structure of the space and tuning the set of feasible actions with respect to the purpose of the environment. The applicability of the framework has been tested and demonstrated on a real environment. The new generation of environments has the potential to produce vast amounts of data about collaboration. The proposed combination of CVE and multimedia data mining technology will allow more coherent and consistent CVEs to be developed.

ACKNOWLEDGMENTS

This research was conducted at the Collaborative Systems Laboratory, University of Technology, Sydney. The research is supported by the Australian Research Council, the University of Technology, Sydney, and the University of Macau.

REFERENCES

1. Ackerman, M.S. Augmenting the organizational memory: A field study of Answer Garden. In *Proceedings of the Conference on Computer Supported Cooperative Work*. (Chapel Hill, NC, USA, 22-26 Oct. 1994), 243-252.
2. Ansari, S., Kohavi, R., Mason, L., and Zheng, Z. Integrating e-commerce and data mining: Architecture and challenges. In *WEBKDD 2000 Workshop: Web Mining for E-Commerce – Challenges and Opportunities*. (Boston, MA, USA, 20 Aug. 2000).
3. Bannon, L.J., and Kuutti, K. Shifting perspectives on organizational memory: From storage to active remembering. In *Proceedings of the Twenty-Ninth Hawaii International Conference on System Sciences*. (Hawaii, USA, 3-6 Jan. 1996), vol. 3, 156-167.
4. Biuk-Aghai, R.P., and Hawryszkiewycz, I.T. Analysis of virtual workspaces. In Kambayashi, Y., and Takakura, H., Eds., *Proceedings of the 1999 International Symposium on Database Applications in Non-Traditional Environments (DANTE '99)*. (Kyoto, Japan, 28-30 Nov. 1999), 325-332.
5. Capin, T.K., Pandzic, I.S., Magnenat-Thalman, N., and Thalman, D. *Avatars in Networked Virtual Environments*. John Wiley and Sons, Chichester, 1999.
6. Conklin, E.J. Capturing organizational memory. In Baecker, R.M., Ed., *Readings in Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration*. Morgan Kaufmann Publishers, 1993, 561-565.
7. Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery: An overview. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., Eds., *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press, Menlo Park, California, USA, 1996.
8. Furst, S., Blackburn, R., and Rosen, B. Virtual team effectiveness: A proposed research agenda. *Information Systems Journal 9*, 4 (1999), 249–269.
9. Greenhalgh, C. *Large Scale Collaborative Virtual Environments*. Springer-Verlag, London, UK, 1999.
10. Gutwin, C., and Greenberg, S. Design for individuals, design for groups: Tradeoffs between power and workspace awareness. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. (Seattle, WA, USA, 14-18 Nov. 1998), 207-216.

11. Hawryszkiewycz, I.T. Analysis for Cooperative Business Processes. In Zowghi, D., Ed., *Proceedings of the Fifth Australian Workshop on Requirements Engineering*. (Brisbane, Australia, 8-9 Dec. 2000), 3-11.

12. Hawryszkiewycz, I.T. Workspace networks for knowledge sharing. In Debrency, R., and Ellis, A., Eds., *Proceedings of AusWeb99, The Fifth Australian World Wide Web Conference*. (Ballina, Australia, 18-20 Apr. 1999), 219-227. (Available at: http://ausweb.scu.edu.au/aw99/papers/hawryszkiewycz/paper.html).

13. Lesley, H.G. and McKay, D.G. Towards an information and decision support system for the building industry. In Mathur, K.S., Betts, M.P. and Tham, K.W. (eds), *Management of Information Technology for Construction*, World Scientific, Singapore, 1993, pp. 101-111.

14. Maher, M.L., Simoff, S.J., and Cicognani, A. *Understanding Virtual Design Studios*, Springer-Verlag, London, UK, 2000.

15. Maher, M.L., Simoff, S.J., Gu, N., and Lau, K.H. Designing virtual architecture. In *Proceedings of CAADRIA2000*. (2000), 481-490. (Available at: http://www.arch.usyd.edu.au/~chris_a/MaherPubs/2000pdf/caadria2000.pdf)

16. Patching, D. *Practical Soft Systems Analysis*, Pitman, London, 1990

17. Schafer, J.B., Konstan, J., and Riedl, J., Electronic Commerce Recommender Applications. *Journal of Data Mining and Knowledge Discovery*, 5 (1/2), 115-152.

18. Simoff, S.J., and Maher, M.L. Loosely integrated open virtual environments as places. *Learning Technology 3*, 1 (Jan. 2001). (Available at: http://lttf.ieee.org/learn_tech/issues/january2001/ index.html#3).

19. Spiliopoulou, M., and Pohle, C. Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery 5*, 1-2 (Jan. 2001), 85-114.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

The PERSEUS Project: Creating Personalized Multimedia News Portal

VICTOR KULESH [1,2], VALERY A. PETRUSHIN [1], ISHWAR K. SETHI [2]

[1] Accenture Technology Laboratory, Accenture
3773 Willow Rd., Northbrook, IL 60062, USA
vkulesh@cstar.accenture.com
petr@cstar.accenture.com
[2] Intelligent Information Engineering Laboratory
Department of Computer Science and Engineering
Oakland University
160 Dodge Hall, Rochester, MI 48309, USA
isethi@oakland.edu

ABSTRACT

This paper describes the Perseus project, which is devoted to developing techniques and tools for creating personalized multimedia news portals. The purpose of a personalized multimedia news portal is to provide relevant information, selected from newswire sites on the Internet and augmented by video clips automatically extracted from TV broadcasts, based on the user's preferences. To create such an intelligent information system several techniques related to textual information retrieval, audio and video segmentation, and topic detection should be developed to work in accord. The approaches to event mining and tracking on the Internet, commercial detection and recognition in video and audio streams, and selection of relevant news video fragments, based on closed captioning and audio transcripts, are described.

KEY WORDS: intelligent information systems, video analysis, audio analysis, multimedia news portal, event tracking, user profiling.

1. INTRODUCTION

Today a growing number of people rely on the Internet as their primary source of information, especially news. Most broadcasting companies have their own web sites and update them with new information as soon as it becomes available. It has also become evident that such vast amount of information is often too much for people to sift through in order to get to that one news they are interested in. Major news portals such as Excite.com or Yahoo.com significantly alleviate the problem. These portals classify the news articles by topic and allow registered users to select the topics they are most interested in. The articles, that news portals provide, contain mostly text and pictures and to see a video of some event the user has to either watch the TV or try to locate video clips somewhere on the Internet. In order to enhance users' experience we decided to develop technology for creating personalized multimedia news portals. The technology integrates the achievements of information retrieval, user profiling, and streaming media indexing techniques.

How does the personalized multimedia news portal work? Figure 1 presents a Perseus prototype. Imagine that you have your personalized text-based news portal such as the Exite.com's NewsTracker. It knows your favorite news providers and events you are interested in. Every day it brings you fresh news articles relevant to your profile. It also lists some other events and you can add them to your profile. If you are interested in a broad topic then several key words could be sufficient to specify your interest (as in NewsTracker), but if you are interested in particular event then, probably, the whole article could be the best and the easiest way to indicate your preference. Next day the portal brings you more relevant articles and you can add more documents to strengthen tracking this event. The system may accept your feedback on how relevant the articles it brought are. The next step is to extend the portal to bring video information that is relevant to events that you are tracking. There are some videos on the Internet, but most of them are outdated, and the rest is badly specified (the manually crafted XML and MPEG-7 will eventually improve the situation though). The most up-to-date and reliable source of video information is TV news broadcasts. Imagine your news portal can communicate with an agent (it could look like a TiVo box or a Web-based service) that knows TV schedule and can analyze video and audio streams to clip relevant video stories, digitize them and send back to your news portal.

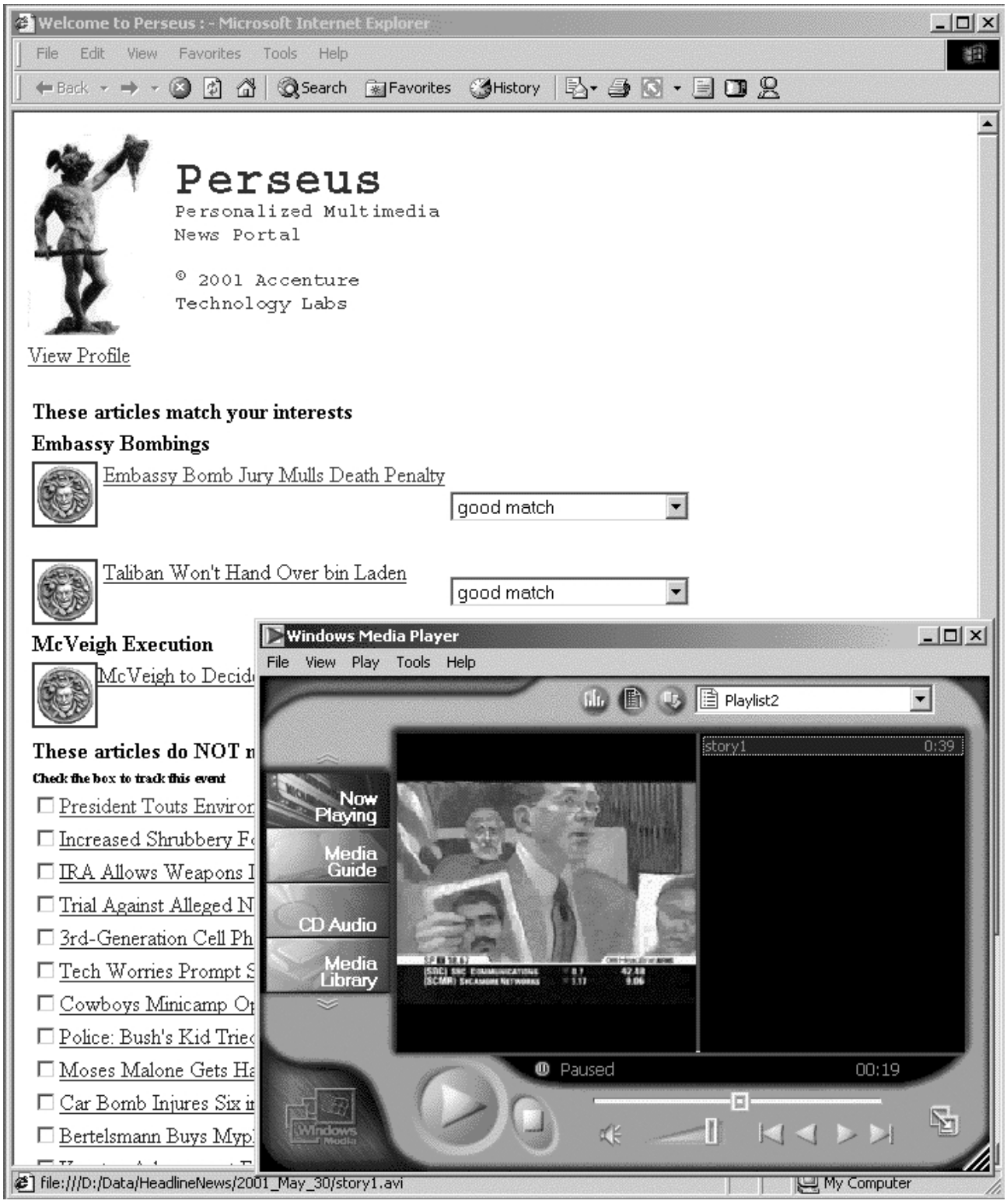


Figure 1. Perseus: personalized multimedia news portal.

Now beside some of your news articles you can see a button that brings a relevant video clip.

An event has its life span. Some events last for days, the others - for months. Organizing multimedia databases according to events and relationships among events brings more meaningful order than using only time and spatial dimensions. This type of organizing allows creating summaries for texts and video films about an event automatically.

To implement personal multimedia news portal the following problems should be solved.

- How do we model the user's interests or in other words what constitutes the user's profile? The challenge here is to achieve high system performance while keeping the user's direct involvement in the profile maintenance to a minimum.
- How newswire articles can be matched to the user's interests? The focus must be on low miss and false alarm rates.
- How audio/video data can be segmented into cohesive stories and matched with the corresponding stories in the user's profile and newswire articles? The major problem here is the correct video clip extraction.
- How user's feedback can be treated and what to do if none is provided?

In the following sections of the paper we shall survey the existing approaches to user profiling, topic tracking and streaming media segmentation, describe the architecture of personalized multimedia news portal, and overview the current state of research on the Perseus project.

2. RELATED WORKS

In this section we shall briefly discuss the current state of art in the following research areas: user profiling, event or topic tracking and audio and video segmentation.

2.1. User Profiling

There are two major paradigms behind user profiling for the purpose of information filtering: content-based and collaborative ones [3], [14]. The main difference between the two is the source of data for the profile. The content-based approach uses the content of desired information for building a profile for a particular user. Usually, a profile consists of a set of weighted words or short phrases. In case of collaborative methods a group profile is constructed for a group of users that have similar interests. In order to build such a profile some members of the group have to provide input to the system, for instance, a review of the book they had bought or movie they had seen. If their feedback is positive then it is assumed, that people with similar interests will like that item too. The advantage of the collaborative filtering approach is in its high precision. But there are several disadvantages: first, people should be highly motivated to provide feedback; second, it takes time to collect enough responses to make a reliable decision, and, third, the system must have access to a network of respondents. This analysis inclined us toward using a content-based approach for user profiling.

In the content-based framework several approaches to creating user's profiles are developed. In [14] three-descriptor representation was introduced. Each category of interest is described by positive, negative and long-term descriptors, which are feature vectors learned from the user's feedback. The advantage of this approach is a fairly good representation of user's interests, however, it heavily relies on the user's feedback, which many users might not be willing to provide. Another technique is using implicit feedback from the user by observing the user's behavior [10]. However it is more difficult to implement. For this reason we decided to utilize the concept of sample document(s) for representing user's interests in the profile. The profile is modified only when explicit feedback from the user is provided to the system.

2.2 Event Tracking

Traditional information filtering methods deal with assigning a topic to each document. Impressive results were achieved in this area using variety of techniques [7], [12]. Recently, however, the focus has shifted towards event tracking rather than topic tracking [1],[2]. The difference is that the notion of topic is more general than the notion of event. For example, "Genetics" is a topic, while "Dolly cloning" is an event, because it occurred in one specific lab at some concrete time. In general, a topic comprises of one or more events and when it comes to news the users tend to favor tracking specific events rather than quite broad topics. In [2] and [11] a framework for event detection and tracking was described. We shall use the same terminology and similar methodology for tracking the documents from the newswire sources. For evaluation purposes we use the Detection Error Tradeoff (DET) curves, introduced in [9] and adopted by the Topic Detection and Tracking (TDT) initiative [1] for evaluation purposes of event tracking techniques.

2.3 Video Segmentation

A lot of research have been done on multimedia content analysis (see a survey [13]). There are several levels of segmentation, but, usually, the following two levels are of concern [8]:

1. Segmentation of audio/video stream into commercials and actual news broadcast parts.
2. Segmentation of news parts into stories.

A similar definition of video segmentation is used in the famous Informedia project for indexing multimedia databases [5]. The authors use an ad hoc but quite successful approach that relies on black frame, scene change frequency and time statistics for commercial detection. Story boundaries are detected by using some heuristics from the closed captioned text which is aligned with the audio transcript for video clip extraction. This technique achieves a low false alarm rate, however, the miss rate still needs some improvement.

3. MULTIMEDIA NEWS PORTAL ARCHITECTURE

A personalized multimedia news portal brings news in accordance to a user's profile and augments them with video clips extracted from TV broadcasts. The system consists of two agents: a Web News Agent (WNA) and a Streaming News Agent (SNA) (Figure 2).

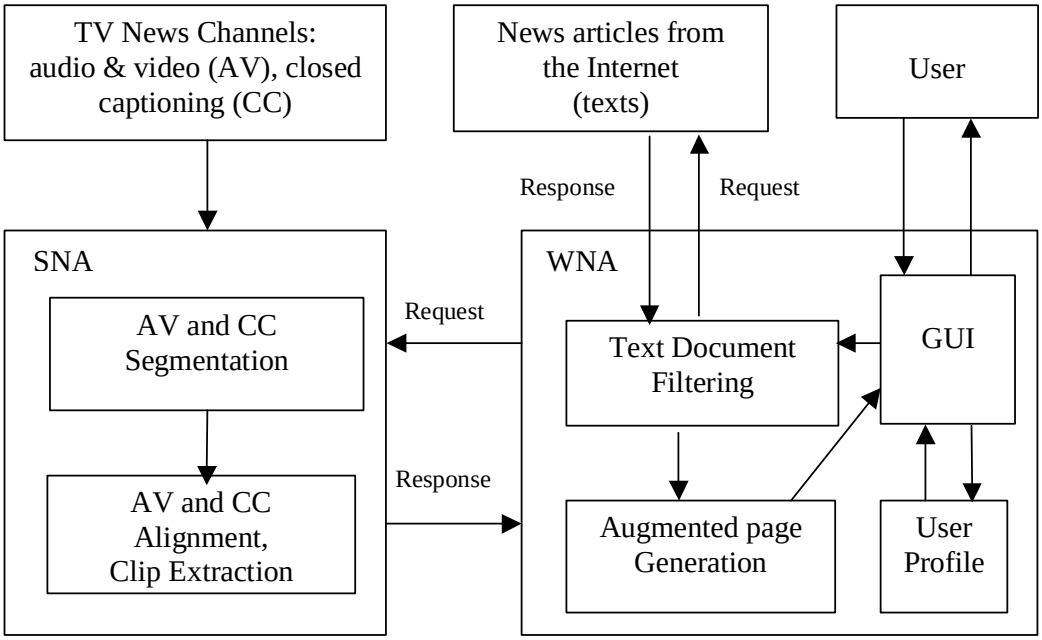


Figure 2. Personalized multimedia news portal architecture.

The Web News Agent performs the following tasks:

- Maintains user's profile.
- Selects relevant text documents on the Internet.
- Communicates with the SNA for requesting and obtaining video clips.
- Generates augmented Web pages for the user.

The Streaming News Agent performs the following tasks:

- Does segmentation of video and audio streams.
- Selects video clips according to requests.
- Communicates with WNA for receiving requests and sending video clips.

Below we describe each block of the system in detail and provide the results of our experiments.

3.1. Information sources and document representation

The textual news documents we used in the experiments come from the Associated Press (AP) in HTML format. A software agent downloads the newswire articles automatically. On average about 20 documents are downloaded from the AP web site every day. In general, the number of documents related to the same event does not exceed two, however, some documents are being updated with the latest additions to the development of the story.

The preprocessing stage includes extracting the actual text of the news article. The automatically retrieved HTML file contains significant amount additional information such as HTML tags, links to other sites, advertisements, etc., which is noise for our system. Stop-word removal and stemming algorithms are applied to the clean text document. Finally the classical term-frequency-inverse-document-frequency (TFIDF) features are computed, where TF stands for term frequency and is computed as the number of times a particular term is present in a document, while IDF is the inverse document frequency and tells us in how many documents the term is found. In order to compute statistically sound IDF values a fairly large collection from the application domain is needed. In our case though we use the IDF values obtained from the documents comprising the user's profile and the news documents of daily catch. Lastly, TFIDF is a term-by-term product of TF and IDF values. Each document is represented by a single TFIDF feature vector. A user's profile is a set of events, where each event is represented by one or several documents. In those cases when more than one document is available for some event in the profile the event feature vector is obtained by averaging all document vectors that describe the event.

3.2. Tracking the events using newswire documents

For tracking events the similarity measure is used. In our experiments we used three similarity measures: the traditional weighted cosine of two vectors, extended Jaccard measure, and Dice measure. To make decision on relevancy of a document the similarity value was compared to a threshold. The threshold was initially computed as to minimize the cost function with the cost of miss and false alarm being equal.

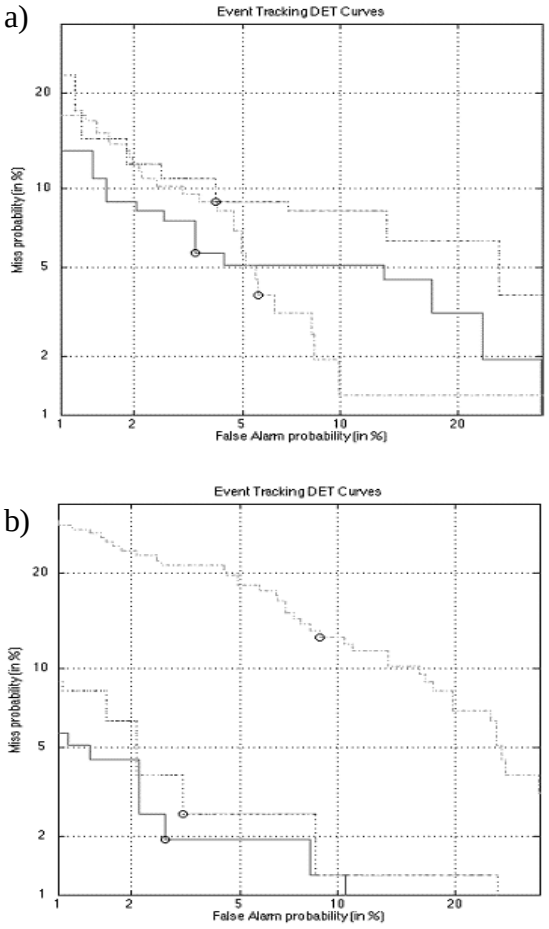


Figure 3. Event tracking DET curves for different similarity measures.

In order to evaluate how the number of documents describing each event in the profile influences the miss and false alarm rates, we created a dataset of events by collecting news articles about those events over a period of several weeks. The dataset contains 14 events. For each event there are four news articles that are used for training purposes (profile documents) and from 3 to 16 news articles for testing purposes. A total of 148 relevant documents and 1839 irrelevant were used during the testing phase. We ran four tests starting with one document per event in the profile and increasing the number by one for each test. Even though the larger number of documents per event is likely to produce better results, we decided to limit out tests to maximum four documents because (1) in most cases four documents are not available for initial seeds, and (2) it is very unlikely that a user will take the trouble of submitting more than one or two documents for an event that may only last a few days.

Figures 3 shows the evaluation results as DET curves for the different similarity measures. Figures 3a and 3b present DET curves for three similarity measures (solid line corresponds to the cosine measure, dash-dotted line - to the Dice, and dotted line - to extended Jaccard measures) and one and four seed documents correspondingly. For one seed document (Figure 3a) all three measures gives comparable results with the cosine measure working better in low false alarm probability domain, and the Dice measure - in low miss probability domain. But for four seed documents (Figure 3b) both cosine and Jaccard measures significantly outperform the Dice measure, and the cosine measure produces the best results.

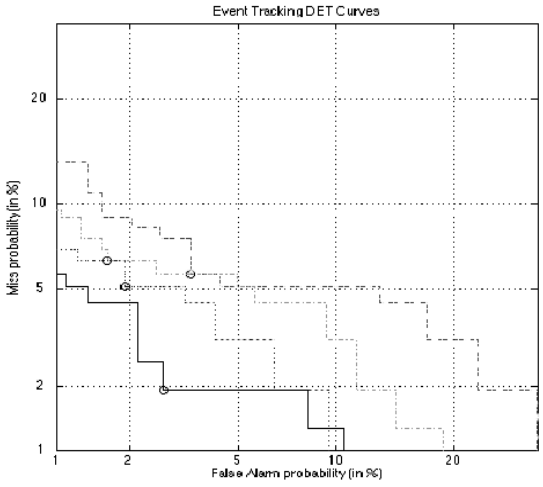


Figure 4. Event tracking DET curves for different numbers of seed documents.

Figure 4 allows comparing the performance of cosine measure for different numbers of seed documents in the event's profile. The double dotted, dash-dotted, dotted, and solid curves correspond to tests with 1, 2, 3 and 4 seed documents per event respectively. The performance improves with adding more documents to the event's profile. The total error is reduced from 4.57% to 2.3%.

3.3. Streaming media segmentation

The first step in audio and video segmentation is to separate commercials from actual news broadcasting. There are several clues that allow detect commercials such as black frames between commercials and frequent changes of shots in video stream, and simultaneous presence of speech and music in audio stream. A closer look allows noticing that some commercials are the short variations of some basic commercials. A basic

commercial is often sliced and stitched to produce a new clip that still conveys the same message, uses the same video shots and audio segments, but some shots are missing from the original clip or appear in a different sequence. The above variability makes difficult the recognition of a known commercial even if the basic commercial is available. These observations led us to an idea of modeling the video stream by a Hidden Markov Model (HMM) of special architecture and modeling the audio stream by a Gaussian Mixture Model (GMM). We use color histograms to represent video frames and mel-frequency cepstrum coefficients (MFCC) for audio data. For our initial evaluation we recorded 65 basic commercials that we used to train the models and 28 commercials that we used for testing.

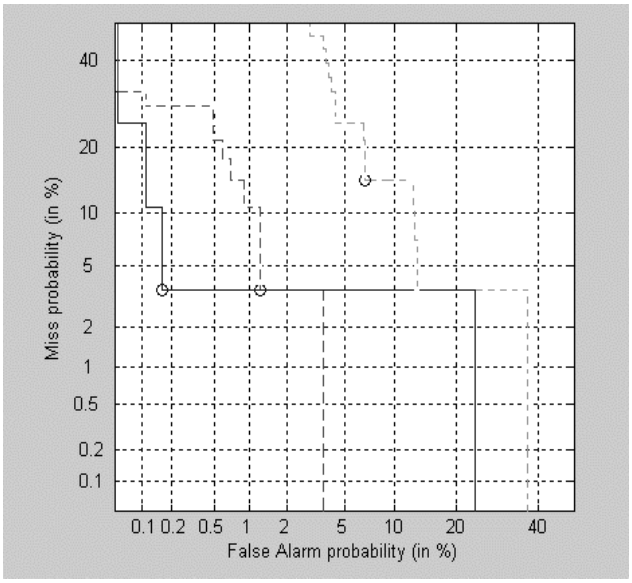


Figure 5. DET curves for commercial recognition.

Figure 5 shows the DET curves of our test with short-dashed, long-dashed and solid lines representing video, audio and audio plus video channels respectively. The small circles on each curve show the optimal operating point of the system. The cost of false alarm to miss ratio is 2:1. For these particular parameters the optimal operating points are shown in Table 1. It can be seen that using audio signal alone gives pretty good results. Using both audio and video models can slightly improve the performance.

Table 1. Optimal operating points.

	Miss Probability (%)	False Alarm(%)
Video only	14.29	6.73
Audio only	3.57	1.24
Video and Audio	3.57	0.16

3.4. Matching newswire documents with AV clips

Now the system has a set of newswire documents that match the user's profile and the next step is to find the corresponding video clips from TV broadcasts. For doing this we use closed captioning, audio, and video streams. Providing closed captioning is the industry standard for news broadcasting. Closed captioning has valuable information about topic and speaker changes, but it is loosely synchronized with audio and video streams. Moreover, most commercially available closed captioning extraction software does not put time stamps into text except for the beginning of recordings. Audio and video streams are synchronized much better. After detecting blocks of commercials the AV stream consists of news segments of length from 5 to 15 minutes. Each segment may contain from one to twenty stories. The system analyzes each AV segment using the described below algorithm.

The algorithm for extracting video clips goes through the following steps:

- *Topic matching.* Text of each topic is extracted from the closed captioning. It is stop-listed, stemmed, and compared, using cosine similarity measure, to each selected newswire article. If the similarity value exceeds the threshold then the topic is marked as relevant.
- *Audio analysis.* Audio stream is cut into chunks that roughly correspond to phrases using largest pause detection algorithm. The chunks have length from 4 to 10 seconds with average about 6 seconds. These chunks are fed into the IBM's Via Voice speech recognition engine, which produces transcripts. Each transcript has a time stamp associated with it. The transcripts are locally aligned on letter level with relevant closed captioning using a sequence alignment algorithm [4], [6]. Using letter level instead of word level alignments allows slightly improve similarity score for imperfect transcripts. Then time stamps for beginning and ending words of the topic are calculated based on time stamps for audio chunks.
- *Video analysis.* The purpose of video analysis is to extract a relevant AV clip based on the topic's beginning and ending time stamps, and time stamps for video shots. This is achieved by searching for the nearest shot cut in the video stream in the interval of 10 video frames. If no shot cut is found then the cut in video is made according to the corresponding time stamp for the topic.

Our approach is similar to one presented in [5] with the major difference that we do not have any time stamps in close captioning except the starting time.

The proposed approach has been tested on a half-hour CNN Headline news broadcasting. Video stories were

labeled manually and then extracted automatically. The 6.8% miss error rate and 22% false alarm rate were obtained. The above approach is very sensitive to the quality of audio stream that in turn influences the accuracy of transcription. Most of the miss and false alarm errors were produced by the stories that have noisy audio signals, such as live footage in a noisy environment or speech over music.

4. CONCLUSION

This paper describes the architecture of a personalized multimedia news portal that provides relevant information, selected from newswire sites on the Internet based on the user's preferences, and augments it by video clips automatically extracted from TV broadcasts. The implementation of such application requires developing technology that combines textual information retrieval with audio and video analyses. The system consists of two agents: a Web News Agent and a Streaming News Agent. The Web News Agent maintains a user's profile, searches for relevant news articles, and makes requests to the Streaming News Agent for video clips. The Streaming News Agent processes TV broadcasts and extracts relevant video clips using closed captioning, speech transcripts and video analysis. The above proposed techniques for streaming media analysis proved to be useful. As the direction for the future work we consider story extraction based only on transcripts.

REFERENCES

1. J. Allan, J. Carbonelly, G. Doddington, J. Yamron, and Y. Yangy, "Topic Detection and Tracking Pilot Study Final Report. Final report." In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

2. J. Allan, V. Lavrenko, and R. Papka Event Tracking, UMASS Computer Science Department, *CIIR Technical Report IR-128*, 1998.

3. D. Billsus and M. J. Pazzani. Learning collaborative information filters. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 46-54, Madison, WI, 1998. Morgan Kaufman.

4. Dan Gusfield. *Algorithms on string, tree, and sequences. Computer Science and Computational Biology.* Cambridge University Press, 1997.

5. A. G. Hauptmann and M. J. Witbrock, "Story Segmentation and Detection of Commercials In Broadcast News Video", 1998.

6. Huang, X. and Miller, W., A Time-efficient, Linear-Space Local Similarity Algorithm, *Advances in Applied Mathematics*, 12:337-357, 1991

7. H. Jin, R. Schwartz, S. Sista, F. Wall Topic Tracking for Radio, TV Broadcast, and Newswire, In *Proceedings of the DARPA Broadcast News Workshop*, 199-204, 1999.

8. R. Lienhart, Ch. Kuhmunch, and W. Effelsberg On the detection and recognition of television commercials. In *Proc. IEEE Conf. on Multimedia Computing and Systems*, Ottawa, Canada, June 1997, pp. 509 - 516.

9. Martin, A., Doddington, G., Kamm, T., Ordowski, M. and Przybocki, M. The DET Curve in Assessment of Detection Task Performance. In *Proceedings of Eurospeech* 1998.

10. D. Oard and J. Kim. Implicit Feedback for Recommender Systems. In *Proceedings of the AAAI Workshop on Recommender Systems*, July 1998.

11. R. Papka *On-line new event detection, clustering and tracking*, PhD Dessertation, University of Massachusetts, Amherst, 1999.

12. M. Stricker, F. Vichot, G. Dreyfus, F. Wolinski Vers la conception automatique de filters d'informations efficacies. In *Proc. RFIA 2000*, Paris, February 2000.

13. Y. Wang, Z. Liu and J. Huang, Multimedia content analysis using both audio and visual clues, *IEEE Signal Processing Magazine*, November 2000.

14. Dwi Hendramo Widyantoro *Dynamic Modeling and Learning user profile in personalized news agent*, M.S.Thesis, Texas A&M University, 1999.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

Automatic Feature Mining for Personalized Digital Image Retrieval

Kyoung-Mi Lee
Department of Computer Science
The University of Iowa
Iowa City, IA 52242, USA
email: klee1@cs.uiowa.edu

W. Nick Street
Department of Management Sciences
The University of Iowa
Iowa City, IA 52242, USA
email: nick-street@uiowa.edu

ABSTRACT
Computing the distance between two features often measures similarity between objects. Unfortunately, the feature space cannot always capture the notion of similarity in human perception. So, most current image retrieval systems use weights measuring the importance of each feature. However, the similarity does not vary with equal strength or in the same proportion in all directions in the feature space. In this paper, we present an incremental method to automatically obtain feature weights based on both the clustered database and on relevance feedback. We show that using cluster information for an initial search gives better results than using the standard distance. In contrast to existing image database systems, the system can store user feedback into feature weights and reduce redundant search steps. We present shape-based indexing and retrieval results that demonstrate the efficacy of our technique.

KEY WORDS
Image retrieval, Incremental clustering, Weighted feature distance, Prototype refinement, Shape-based query

1. Introduction

In general, queries based on content similarity to an example (object or images) in terms of features (such as color, texture, shape, etc) are known as Query-by-Example. These features are extracted from each example to represent its content and then the image retrieval procedure looks for the most relevant examples in a multidimensional space defined by the features under a given similarity metric. However, features are unequal in their differential weight for computing the similarity between examples. So most current image retrieval systems use weights measuring the importance of each feature. However, the similarity does not vary with equal strength or in the same proportion in all directions in the feature space emanating from a given query.

Extracted features cannot always capture similarity in human perception. To give the user more control over the search criteria, some systems allow the user to weight the features [6, 2]. This manual adjustment is not generally applicable because it is time consuming and the user does not usually have a detailed understanding of the features.

Recently, relevance feedback has been the most commonly applied method to calculate feature weights using the user's preference. An initial search is done in the database using the original query input by the user. Upon being presented the results of this search, the user labels some of these results as relevant or irrelevant according to the user's information needs. Rui *et al.* used a vector space model where a new query feature vector is generated as a weighted linear combination of the original feature vector and the feature vectors of the images that were labeled as relevant or irrelevant by the user [8]. Peng *et al.* proposed probabilistic feature relevance learning to update weights iteratively, also based on relevance feedback [7]. Aksoy *et al.* used the ratio of standard deviations of the feature values both for the whole database and also among the images selected as relevant by the user [1].

In this paper, we propose an incremental learning to automatically obtain feature weights based on both the clustered database and on relevance feedback. Feature weights with cluster information give better initial search than flat feature weights. This feature weights is updated and adapted using relevance feedback by clustering relevant examples into the corresponding cluster. Because users specify their particular preferences by providing feedback, the adapted feature weights can provide personalized retrieval. Section 2 presents the proposed method: incremental clustering, weighted feature distance, and prototype refinement. The indexing and retrieval performance of the shape-based query system are presented in Section 3. Finally, a conclusion will be presented in Section 4.

2. Method

In this section, we propose a novel method that estimates each feature's weight in the distance metric. The target of this paper is to get personalization with feature weights by incorporating relevant feedback.

2.1 Incremental Clustering

A major issue that databases are now facing is the extremely high rate of update. Typically, clustering is used to accelerate query processing by considering only a small number of representatives of the clusters, rather than the entire database. However, existing clustering algorithms are not suitable for maintaining clusters in ever-increasing

volume of data, and they have been struggling with the problem of updating clusters without frequently performing complete reclustering [3].

A given set of N data, $\mathbf{V} = \{\mathbf{v}^1,\mathbf{v}^2,\cdots,\mathbf{v}^N\}$ will be partitioned into C clusters $\{\mathbf{P}^1,\mathbf{P}^2,\cdots,\mathbf{P}^C\}$ such that data within the same cluster have a high degree of similarity, while shapes belonging to different clusters have a high degree of dissimilarity. Each of these clusters is represented by a prototype $(\mathbf{P}^c)^{c=1\cdots C}$. The ith feature of the prototype, $\mathbf{P}^c = (P_1^c, P_2^c, \cdots, P_I^c)$ is the center of the ith features of data in the corresponding cluster and can be computed by averaging the features that belong to the cluster **c**. To measure the spread of a set of data around the center of the data in the cluster, we use the standard deviation (σ^c).

In this paper, we use an incremental learning method [4] as follows. If an input data is similar to an existing cluster or is grouped by the user explicitly, the input data is assigned to one of the existing clusters. If the input data is dissimilar to all existing clusters or is designed as a new cluster by a user, the new cluster is created with the input data.

Incrementally, whenever a new datum **u** is indexed and assigned to an existing cluster **c**, P_i^c and $\sigma_i^{c'}$ are updated as follows:

$$P_i^{c'} = \frac{N_c P_i^c + u_i}{N_c + 1} \quad \text{and } \sigma_i^{c'} = \sqrt{\frac{N_c s_i + (P_i^{c'} - u_i)^2}{N_c + 1}},$$

where N_c is the number of data in the cluster **c** before assigning **u** and is increased by 1. Also $s_i = (\sigma_i^c)^2 + (P_i^{c'}$.

2.2 Weighted Feature Distance

In this section, we consider the problem of similarity measurement used for retrieval. The most common measure of similarity between data, **v** and **u**, is the distance between them as $\mathbf{d}_p(\mathbf{v},\mathbf{u}) = \|\mathbf{v} - \mathbf{u}\|_p$ where $\|\cdot\|_p$ is, e.g., the Manhattan (p=1), Euclidean (p=2), or max (p=∞) norm. In the I dimensional space, the distance between a prototype $\mathbf{P}^c$ and a given query **q** is computed as

$$\mathbf{d}_p(\mathbf{P}^c, \mathbf{q}) = \left(\sum_{i=1}^{I} |P_i^c - q_i|^p\right)^{\frac{1}{p}} \tag{1}$$

where P_i^c and q_i are the ith features of $\mathbf{P}^c$ and **q**, respectively.

However Eq. (1) does not take into account the influence of the scale of each feature in the distance computation. The scale of a feature can change the overall contribution of that feature in the distance computation [7]. Thus, each feature should be rescaled so that all features contribute equally to the distance computation in Eq. (1).

To capture such scale information as weights, we propose a measure using the standard deviation. The weight for a feature P_i^c can be defined as

$$w_i^c = \frac{|P_i^c - q_i|}{\sigma_i^c}.$$

That is, a feature's importance inversely proportional to the relative distance between the query and the prototype along that dimension. It is also important to normalize each feature in the the shapes to the same range to ensure that each individual feature receives equal weight in determining the similarity between two shapes [8]. Therefore, we propose the following exponential weight

$$w_i^c = \begin{cases} 1 - \exp\left(-\dfrac{|P_i^c - q_i|}{\sigma_i^c}\right), & \text{if } \sigma_i^c \neq 0 \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

It is evident that $0 \leq w_i^c \leq 1$, where $w_i^c = 1$ indicates that q_i is too far from σ_i^c or σ_i^c is 0, and thus q_i isn't similar to P_i^c at all. On the other hand, $w_i^c = 0$ states P_i^c and q_i are identical, and so q_i is totally similar to P_i^c. Values in between show the degrees of weights that q_i exerts at i. In the distance function used as a similarity measure, it is desired to give a larger weight to a dissimilar shape. Thus, Eq. (2) can be used as weight for the distance between a prototype and a datum. The distance between $\mathbf{P}^c$ and **q** can finally be computed as

$$\mathbf{d}_w(\mathbf{P}^c, \mathbf{q}) = \left(\sum_{i=1}^{n} |w_i^c (P_i^c - q_i)|^p\right)^{\frac{1}{p}}. \tag{3}$$

These weights enable the distance computation to elongate dissimilar features, and, at the same time, to constrict the closest ones.

In addition to $\mathbf{d}_w(\mathbf{P}^c,\mathbf{q})$, for the actual retrieval, we need to compute the distance between **q** and data **v** in $\mathbf{P}^c$ which is similar to **q** as

$$\mathbf{d}_w(\mathbf{v}, \mathbf{q}) = \left(\sum_{i=1}^{n} |w_i^c (v_i - q_i)|^p\right)^{\frac{1}{p}}. \tag{4}$$

Given **q**, selection of data sharing similar characterizing elements starts by comparing **q** to the set of prototypes using Eq. (3). If some prototypes are selected as being similar to **q**, only the data in those prototypes calculate the distance with the query using Eq. (4).

2.3 Prototype Refinement

In the information retrieval literature, it has been well established that retrieval performance can be significantly improved by incorporating the user as part of the retrieval loop. the most popular approach is query refinement which is the process of automatically adjusting an existing query using information fed back by the user about the relevance of previously retrieved examples. We want to use relevance feedback to give the user more control on the search criteria. We also want to allow the database system to learn the user's need. When the user looks for the image he found previously, the system can give results using information learned by his feedback, instead of starting an initial search.

1. Initialization: m = 0,
   $N_{c_0} = N_c$, $\mathbf{P}^{c_0} = \mathbf{P}^c$, and $\sigma^{c_0} = \sigma^c$, where m = 0 means there is no relevance feedback to **q**.
2. Retrieval
   (a) Calculate $w_i^{c_m}$ using Eq. (2).
   (b) Search similar prototypes using Eq. (3).
   (c) Search relevant data in similar prototypes $\mathbf{P}^{c_m}$ using Eq. (4).
3. User relevance feedback: Select relevant data R^j from retrieved results.
4. Prototype refinement: Modify each new prototype and standard deviation using Eq. (5).
5. $m = m + 1$ and go to step 2.

Algorithm 1. Feature weights are updated incrementally by user relevance feedback

In this paper, we refine prototypes by specifying over time a set of relevant results. At each step, the user labels the retrieved data as relevant or irrelevant to the query **q**. The system then clusters the relevant data to the corresponding prototype $\mathbf{P}^c$ and makes it satisfy the user's need better than the original prototype

$$
\begin{aligned}
N^{c'} &= N_c + N_c^R, \\
P_i^{c'} &= \frac{N_c P_i^c + \sum_{j=1}^{N_c^R} R_i^j}{N_c + N_c^R}, \quad \text{and} \\
\sigma_i^{c'} &= \sqrt{\frac{N_c s_i^c + \sum_{j=1}^{N_c^R} (P_i^c - R_i^j)^2}{N_c + N_c^R}},
\end{aligned}
\tag{5}
$$

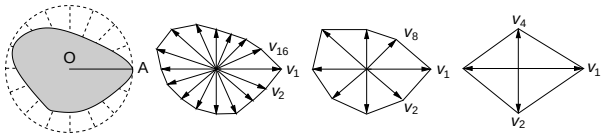
where N_c is the number of shapes and N_c^R is the number of relevant shapes and R_i^j is the ith feature of the jth relevant data, in the cluster **c**, respectively. Also, $s_i^c = (\sigma_i^c)^2 + (P_i^{c'} - P_i^c)^2$. This process is detailed in Algorithm 1.

3. Application to Shape-based Query

We present experiments using the well-known Columbia database which contains 1440 images of 20 different objects: 72 images per object taken at 5^o in pose [5]. The performance in the experiments is evaluated using two measures: precision and recall. Recall is the ratio of relevant images retrieved to the total number in the database; it measures the ability of a system to present all relevant images. Precision is the ratio of relevant images retrieved to the total number of images retrieved; it measures the ability of a system to present only relevant images.

3.1 Shape-based Query

In this paper, we use only shape for indexing and retrieval and adopts the centroid-radii model [4, 9] to represent the shape. The shape is represented as a vector, **v** = $(v_1, v_2, \cdots, v_I)$, = $(q_{\theta_1}, q_{\theta_2}, \cdots, q_{\theta_I})$, where I is the number of shape features and v_i is the ith radius from the centroid to the boundary of the shape. With a sufficient number of radii, dissimilar shapes can be differentiated from each other. We used a 36-dimensional data space for two databases. Shapes are normalized for translation, size, orientation and reflection during the clustering.



(a) Query (b) 16D (c) 8D (d) 4D

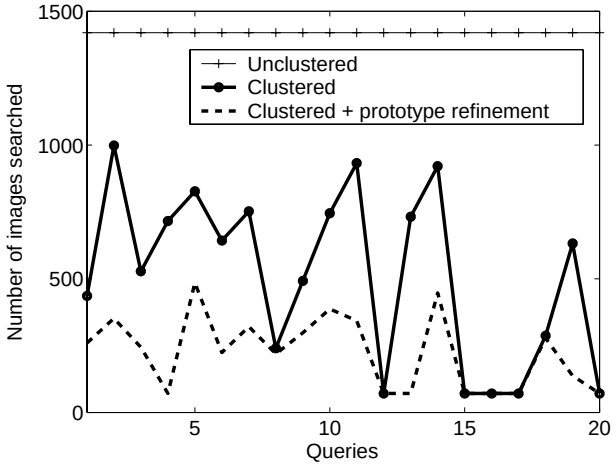
Figure 1. Shape-based query

The three representations in Figure 1 ((b), (c) and (d)) show this model can facilitate multiresolution representation and multiresolution querying. For example, for a database whose shapes are represented using I features, a shape query with $\frac{I}{2}$ features can be of the form $(v_1, v_3, \cdots, v_{I-1})$. This is a less precise description of the query. To facilitate query retrieval with such query, the query can be mapped to a I-dimension vector where values of those unspecified features are filled with "*don't care*" values [9]. Clearly, more shapes match such query than when all the features are specified.

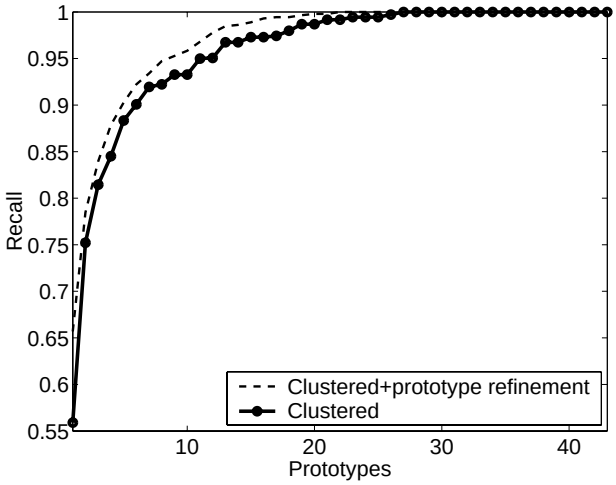
3.2 Shape-based Indexing

We used 1420 images for indexing and 20 images (45^o for each object) for query. For object-based images, locating and recognizing objects from the images is a challenging goal. There is no general solution to the difficult task of object segmentation, and successful solutions are limited to specific domains. In cases where good segmentation is possible, the object shape is a characteristic which can contribute enormously in further analysis. In this paper, we use an adaptive thresholding algorithm to extract object boundaries. The object database then was incrementally clustered and created 45 prototypes. These prototypes are used as

an indexed key to reduce search time. To evaluate indexing effectiveness, we tested the retrieval performance after clustering.



(a)



(b)

Figure 2. Indexing Effectiveness of Columbia database: (a) Number of images searched for recall of 1.0 per query and (b) Average retrieval performance vs. the number of prototypes

Figure 2 (a) shows the number of images searched for a recall of 1.0 in the database. A recall of 1.0 is the case of finding 71 images of each object, though there may be different shapes at some angles. Instead of searching and ranking all shapes, the system searches only a small set of clusters and ranks shapes in the clusters. If some positive examples are missed, the system searches the next most relevant prototypes. Figure 2 (a) shows that the clustering method reduces search range. Figure 2 (b) shows how the number of prototypes affects the average performance for 20 queries performed on the database of the 1420 images. With only the one most relevant prototype, images from the same object were judged as most similar 56.7% of the time in average. As the number of prototypes reached six, performance reaches approximately 90%.

3.3 Shape-based Retrieval

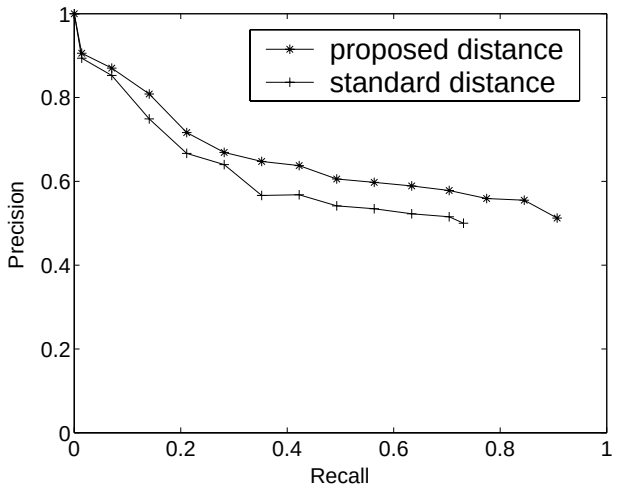


Figure 3. Average precision-recall of Columbia database

The proposed system first finds six most similar prototypes and then ranks shapes in these six clusters based on weighted distance from the queries. We measure the precision and recall over a number of images in six clusters. The performance using the standard distance measure is also presented by directly calculating $\mathbf{d}_p(\mathbf{q},\mathbf{v})$, the unweighted distance between a query $\mathbf{q}$ and a shape $\mathbf{v}$ in database. Figure 3 shows the average result of 20 queries on the database at the initial search without feedback. It shows that the proposed weighted distance is better than the standard distance, in both precision and recall. Figure 5 (a) and (b) show examples to compare the two distances.

In addition to weighted distance, relevance feedback is an important issue in this paper. In the experiment of prototype refinement, we tested the same query with and without refinement. Figure 2 (a) and (b) in Section 3.2 shows that our prototype refinement give a better guide for search, thus the system works more efficiently as it learns the shapes.

For retrieval experiments, the system performs a new search in the database and retrieves 21 images in every incremental step. In order to take into account only the feedback effect, the system eliminates the shapes that have already been learned by the user in the first retrieval. The purpose of the experiment is to compare the retrieval performance with relevance feedback vs. the performance without relevance feedback. Figure 4 shows the average results over 20 queries of the proposed relevance feedback algorithm and shows that the precision keeps getting better

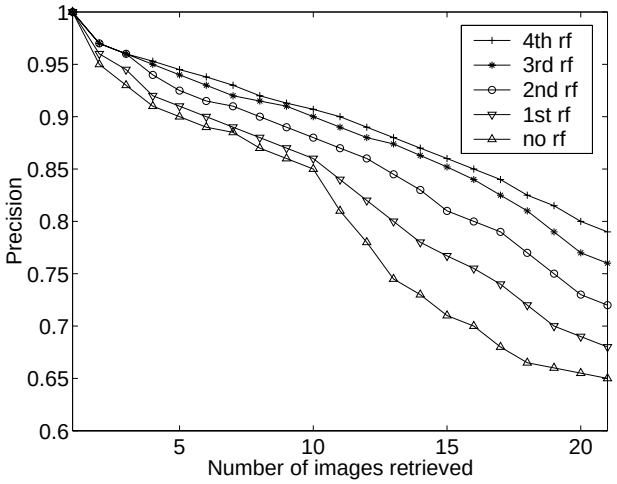
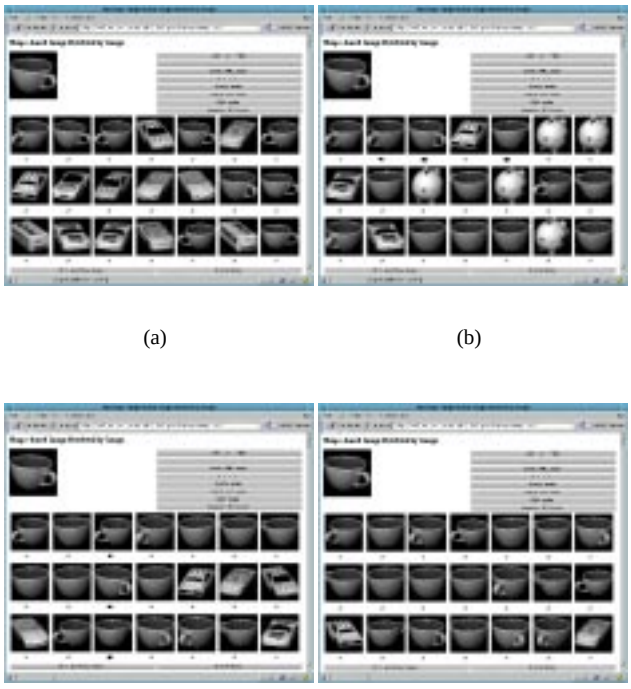


Figure 4. Average precision of Columbia database for the first two relevance feedbacks (rf)

than with no user feedback. Figure 5 (a) and (b) show an example of results with two distances.

Figure 5 presents the retrieval result of a query (the upper left image) on the database. The user selected what he wanted to retrieve (small black boxes) in Figure 5 (b). Figures 5 (c) and (d) show the results based on the user relevance feedback. One important thing of this paper is the proposed system returns Figures 5 (d) when the user wants to find the same query after the second relevance learning, not starting again with Figure 5 (b).



(a) (b)

(c) (d)

Figure 5. Retrieval results of the Columbia database: Images are ranked from left to right, top to bottom by increasing distance: (a) with the standard distance and (b) with the weighted distance for initial search, respectively, (c) with first relevance learning, and (d) with the second relevance learning

4. Conclusion

This paper presents an incremental feature weight learning method based on both the clustered database and on relevance feedback for efficient shape-based image indexing and retrieval. The experimental results shows convincingly that the weighted distance using cluster information achieves better results even in the initial search. We also showed that learning feature weights based on user feedback can indeed improve retrieval performance of the proposed database system.

References

[1] S. Aksoy, R. M. Haralick, F. A. Cheikh, and M. Gabbouj. A weighted distance approach to relevance feedback. In *Proceedings of the IAPR International Conference on Pattern Recognition*, pages 812–815, 2000.

[2] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. F. Shu. The Virage image search engine: An open framework for image management. In *Proceedings of SPIE*, volume 2076, pages 76–87, 1996.

[3] B. B. Chaudhuri. Dynamic clustering for time incremental data. *Pattern Recognition Letters*, 15(1):27–34, 1994.

[4] K.-M. Lee and W. N. Street. Automatic segmentation and classification using on-line shape learning. In *Proceedings of the 5th IEEE Workshop on the Application of Computer Vision*, pages 64–70, 2000.

[5] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical Report CUCS-005-96, Columbia University, 1996. http://www.cs.columbia.edu/CAVE/research/softlib/coil-20.html.

[6] W. Niblack, R. Barber, W. Equitz, M. Flicker, E. Glasman, D. Petkovic, P.Yanker, and C. Faloutsos. The QBIC project: Query images by content using color texture and shape. In *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Databases V*, volume 1908, pages 173–187, 1993.

[7] J. Peng, B. Bhaunu, and S. Qing. Probabilistic feature relevance learning for content-based image re-

trieval. *Computer Vision and Image Understanding*, 75(1/2):150–164, 1999.

[8] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in MARS. In *Proceedings of the IEEE International Conference on Image Processing*, volume II, pages 815–818, 1997.

[9] K.-L. Tan, B. C. Ooi, and L. F. Thiang. Indexing shapes in image databases using the centroid-radii model. *Data and Knowledge Engineering*, 32(3):271–289, 2000.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

Relationship extraction from large image databases

Chabane Djeraba
IRIN, Ecole Polythechnique de l'Université de Nantes,
2 rue de la Houssinière, BP 92208 - 44322 Nantes Cedex 3, France
E-mail: djeraba@irin.univ-nantes.fr

ABSTRACT
We highlight an algorithm that discovers hidden relationships between image features. These relationships accurate image classes. The best relationships are selected on the basis of confidence measures. To reduce the combinatory explosion of relationships, because images of the database contain very large numbers of colors and textures, we consider a visual thesaurus that group together similar colors and textures. Thus the visual thesaurus summarizes the image features. The visual thesaurus is created by an algorithm based on a clustering strategy. The relationships discovered permit the automatic categorization of images during their insertion into image databases, and return accurate and relevant results.

Keywords
Image, databases, relationship, discovery.

1. INTRODUCTION
In lots of content-based indexing and retrieval such as [18], [11], [8], [13], knowledge are weakly supported. Queries based on visual content are not powerful enough to specify semantics in queries, such as "finding images of animal in mountains". Such queries are not easy, because the links, between the content and the semantic in the user's mind, are not easy to declare. There are differences between the image features of the examples (ex. color, texture) and the semantic (ex. flowers in front of a lake) the user is looking for. It is not easy, because the user has to define the semantic he is looking for in terms of visual descriptions. In the query, the features mentioned earlier may be useful (animals have specific textures and specific colors), but they are not sufficient. What is missing is the knowledge extraction capability. "Animals in mountains" have specific colors, textures and relationships between visual features that are very difficult to specify explicitly in the query. These relationships are hidden. The user's mind discriminates "images that contain animals in mountains" from "images that contain animals in front of lakes". This is possible because there are several hidden relationships between visual features that discriminate the two databases of images: "animals in mountains" and "animals in front of lakes".

The challenge question is: how can the system extract not only the visual features (ex. colors, textures), but also the hidden relationships between the features in order to make possible semantic search? The answer to this challenge question highlights new solutions in which the relationships are extracted automatically, that are necessary to efficient semantic search.

An average way to deal with this challenge is combine textual and visual features in indexing and retrieval. Based on experiments, it is proved that queries based on textual and visual features are more efficient than queries based on textual or (exclusive) visual queries [3], [7]. Visual features are extracted automatically, however textual features are annotated manually. Although the effort in "manual" annotation is less important than in classical information retrieval [21], [20], the "manual" property of the annotation is a serious "speed limit" to the exploitation of such algorithm. Furthermore, the manual annotation is quite subjective and ambiguous, and it is very difficult to capture the visual content of an image using words.

One algorithm, to deal with this problem, organizes the digital databases in a meaningful manner using image categorization. Image categorization classifies images into semantic databases that are manually pre-categorized. The categorization of images into databases can be helpful in semantic organization of digital databases. The categorization of images is quite difficult in general. Images in the same semantic databases may have large variations with dissimilar visual descriptions (images of persons, images of industries, etc.), and images from different semantic databases might share a common background (some flowers and sunset have similar colors). And this limits the efficiency of automatic categorization based exclusively on visual content of images (texture, color).

In this paper, we propose a new scheme for automatic hierarchical image categorization. We assume a training set of images with known database labels available. We use low level features – autocorrelogram of colors and Fourier descriptors of texture – which are, together, efficient for content-based image retrieval. Using low level features for the training images, relationships among visual features are extracted automatically. These relationships discriminate image databases. The best

relationships are selected on the basis of two confidence measurements (conditional probability and implication intensity). Once the categorization tree is obtained, any new image can be classified easily. Furthermore, query results are obtained with semantics because they belong to semantic databases. We prove that discovering hidden relationships contributes to making the content-based retrieval more efficient. The hierarchical algorithm of image categorization has several advantages: - easy browsing and navigation through the databases, - efficient retrieval, - and ergonomically friendly presentation of databases.

In this paper, we present, in section 2, the related works. In section 3, we present the advanced framework of the algorithm. In the section 4, we present how to create visual thesaurus from digital images. In section 5, we present image features. In section 6, we highlight the interest of the relationship discovery algorithm, and show how the relationship discovery engine is used to learn discriminating characteristics that are fundamental to categorization. In section 7, we present result evaluations.

1. RELATED WORKS

Very few studies have considered relationship extraction and data categorization on the basis of image features, and very few of them, [12] [2], concern relationship extraction and data categorization in the context of image indexing and retrieval.

More recently, a method, for data resource selection in distributed visual information systems [2], has been proposed. The method is based on a metadatabase at a query distribution server. The metadatabase records a summary of the visual content of the images in each database through image templates and statistical features. Image templates and statistical features characterize the similarity distributions of the images. The selection of the databases is driven by searching the metadatabase using a ranking algorithm that uses query similarity to a template and the features of the database associated with the template. Two selection algorithmes have been implemented, termed by mean-based and histogram-based algorithmes. The template summarizes the visual content-based on basic features (histogram of color and mean of colors). An important difference with our algorithm concerns the use of mean and histogram of colors; our algorithm uses autocorrelograms for colors and Fourier coefficients for textures. Finally, the content of databases is summarized by visual templates and statistical features, however, in our algorithm, the content of databases is summarized by relationships among visual features.

Another algorithm [12] proposes a new scheme for automatic hierarchical image categorization. It assumes that a training set of images with known database labels is available. It uses image descriptors: banded color correlograms. Using banded color correlograms, the algorithm models the feature data using singular value decomposition and constructs a categorization tree. Once the categorization tree is obtained, any new image can be classified. The method has been tested on eleven databases (aviation, photography, British motor car collection, etc.). The categorization tree obtained seems to be conform to the semantic content of the eleven databases. An interesting point of this algorithm is the use of correlograms. The results suggest that correlograms have more latent semantic structures than histograms. The technique used extracts a certain form of relationship to classify images. Using a noise-tolerant SVD [5] description, the image is classified in the training data using the nearest-neighbor with the first neighbor dropped. Based on the performance of this categorization, the databases are partitioned in sub-databases, and the inter-class dissociation is minimized. This is accomplished using normalized cuts. The sub-databases and the training images that were correctly classified with respect to the sub-databases are worked upon recursively to obtain a hierarchical categorization tree.

Compared to this algorithm [12], our algorithm uses two basic visual features: autocorrelograms of colors and Fourier descriptors of the texture. These two descriptors make the description of the image content accurate. So our algorithm extracts automatically the relationships between these two descriptors in each database of images. The extraction of relationships is not possible in this algorithm as the content of images are represented by only one descriptor: banded color correlograms.

In the general context of content-based image retrieval, except a few cases such as those presented above, the state of the art is weak. Although many visual information systems have been developed recently [8], [18], none of these systems operate by considering relationship extracted from image databases. The selection of image databases for a given query discussed in this paper offers a new algorithm to design a relationship-content-based retrieval system.

2. ADVANCED FRAMEWORK

Basically, the framework is composed of two important components: indexing and retrieval. In indexing, the image contents are extracted automatically. Methods, whether aided by the user or not, may identify relevant regions in images and compute features such as color and texture of the regions or compute the visual features of the whole image. The extracted contents are represented as or transformed into suitable models and data structures, and then stored in the database. In an optimal framework, the representation of the contents is managed by a virtual memory, in which the most frequently accessed contents are kept in the high speed memory (e.g. central memory) and the remaining contents are kept in lesser speed memory (e.g. secondary memory). The retrieval consists in searching images by selecting target images or content properties such as color, texture of image regions, or combinations of these. The system includes a visual query tool that lets users form a query by drawing, sketching and selecting textures and colors. Finally, the retrieval process computes distances between source and target features, and sorts the most similar images. The basic properties of the framework presented so far are classical and shared by lots of visual information systems [8], [13], [11]. In the context of relationship-content-based indexing and retrieval, we propose a suitable framework in which three extensions are concerned:

-Firstly, on the basis of semantic databases (Birds, Flowers, Water, Buildings, etc.), the system extracts image features, and discovers the relationship (relationships shared by images) that discriminates each database. The relationships describe relationships between visual features (color and textures of

images). Each set of relationships linked to a database summarizes image contents of the database. Relationships contribute to database discriminations. The features and relationship extracted are saved in the database.

Secondly, when an image is inserted into the database, it is classified "automatically" in the database hierarchy. At the end of the categorization process, the image is added to a specific database. In this case, the distance between the image and the relationship associated to the database is the shortest one, compared to the distance between the image and the other databases. Otherwise, the instantiation relationship between the image and the database, is not considered.

-Thirdly, the architecture supports efficient retrievals and browsing through databases. For example, the user may specify queries such as "find images similar to the source image but only in People databases" or "find all images that illustrate the bird database with such colors and such texture". In the retrieval task, when the final database is selected, features (colors, textures) of the query specification are compared with features of the image database to determine which images correctly match (are similar to) the given features. The matching task is based on computing the distance between target and source image regions. When mixing several features, such as colors and textures, the resulting distance is equal to the *Sum* taking into account the fuzzy values of the considered features. The resulting images are sorted, the shortest distance corresponds to the most similar images. Ideally, we should develop a fully-automated system that, after extracting and storing visual features of images, clusters together similar images in databases. Each database is automatically obtained and should correspond to a semantic database of the application field. However, this is not realistic as some tasks inevitably require the user's intervention such as database semantic identification and validation.

These extensions improve system performance compared to previous versions [6] of the system. For example, in the retrieval task, when the user gives an example image (called source image) to formulate his query, and asks "find images similar to the source image", the system will not match the source image with all images in databases. It will match the source image features with only the target image features of suited databases. If the relationship that characterizes a database is globally respected by the source image, then the considered database is the suited one. Then, the system focuses the search on the sub-databases of the current one. In the target database, the search is, generally, sequential. Another advantage of these extensions is the richness of the description contained in the results of queries since the system presents both similar images and their databases. For example, the user specifies the query: "find images that are visually similar to query images". The query images represent waterfalls. The system matches the query images with both database centroids and relationship, in the form of relationships of the different databases. The image belongs to waterfall database, because the distance between the image and the class centroid of the image database is the shortest one, and the relationships associated with the database are globally verified. The retrieval process, then, matches all the images of this database in a sequential order. The first images returned contain waterfall. All the images are visually similar to the example images. This example illustrates "query by examples" that are based on combination of the visual features. "Query by examples" specifies a query that means "find images that are similar to those specified". The query may be composed of several images. Several images make the "efficiency" of retrieval accurate. For example, several images of "waterfall" give us accurate description of the waterfall. This property enables possible the refinement of retrieval based on the feed backs (results of previous queries).

3. VISUAL THESAURUS

The creation of the visual thesaurus is a fundamental pre-processing step necessary to extract relationships. It is not possible to extract useful relationships without the pre-processing step in which similar features are clustered. The centroids of the feature clusters constitute the visual thesaurus. Without the visual thesaurus, we have to consider all features of all images, and then, we obtain very few features shared by images, and then very few relationships, that discriminate databases. In such cases, the relationships extracted will be too weak to discriminate the databases.

Based on the learning set of length equal to T (image databases contain T images), the algorithm creates a visual thesaurus of colors and textures. The length of the visual thesaurus is equal to L. The visual thesaurus contain the most representative colors and textures of images. In our experiments, $L = 256$ and $T = 27331$ images. In other words, the algorithm clusters together similar numerical representations of color and texture. The algorithm result summarizes the color and texture features of the image databases. The difference between the color and texture clustering algorithms concerns the features (autocorrelogram for colors and Fourier coefficients for texture) and distances used, respectively Euclidean for texture and quadratic for the color. The features of databases are composed of features of unknown probability density.

The principle of the algorithm is distinguished by three steps:

- The first step, categorization, clusters features of the learning set around the initial visual thesaurus features that are the most similar. The objective is to create the most representative partition of the feature space. In other words, based on the visual thesaurus $Y_s = \{Y_{s,k}\ k=1,..,L\}$ and the learning set, the system extracts the partition $Database_s = \{Database_{s,k};\ k = 1, ..., L\}$, in which $distance(x, y)$ is minimal. So: $x_t \in Database_{i,k}$ when $distance(x_t, y_k) \leq distance(x_t, y_j)\ \forall j \neq k$. The distortion $D_s = 1/T \sum_{t=1,T} min_Y distance(x_t, y),\ y \in Y_s$. If the feature is a color then $distance = quadratic\_distance$. If the feature is a texture, then the $distance = texture\_fourier\_distance$.

The quadratic distance takes into account the color similarity between the autocorrelogram bins by using the symmetrical similarity matrix A. The matrix weights are normalized to obtain $0 \leq a_{pq} \leq 1$. So, the matrix diagonal is equal to 1, since any color is identical with itself ($a_{pp}=1$). A coefficients (a_{pq}) close to 0, represents a dissimilarity between p and q bins. For example, in QBIC [8], the quadratic distance between two color autocorrelograms is used with a similarity matrix A whose elements are defined by: $a_{ij} = (1 - d_{ij}/d_{max})$, with $d_{max} = max_{ij}(d_{ij})$, d_{ij} being Euclidean distance between the colors i and j in any color space. The two distributions H and I, may also be

normalized in order that $0 \le h_{c_p}, i_{c_p} \le 1$ and $\sum_p h_{c_p} = \sum_p i_{c_p} = 1$; $D_{L2} = \sqrt{(\sum_{l=1,n} (h_{cl} - i_{cl})^2)}$. The quadratic distance $D_Q(H, I) = \sqrt{((H-I).A.(H-I)^T)}$ with A the similarity matrix $(n \times n)$, $A = [a_{pq}]$, a_{pq} weight of the similarity between the p and q bins. This distance makes it possible to obtain satisfactory results since it appreciates color similarity correctly. However, its major drawback is that it is time-consuming compared to the other distances. Euclidean distance results from the quadratic distance where A matrix is the identity matrix (no correlation between the autocorrelogram bins).

For the texture, we extend the Euclidean distance to Fourier coefficients; we call it « *texture_Fourier_distance* ». Thus, the matching distance between the Fourier descriptors of texture *t'* of an image *image'* and the Fourier descriptors of the texture *t* of an image « image », is triggered by computing the distance between *t* and *t'*, namely: $d(t, t') = \sqrt{(\sum_{n=1,N} (|T'_n - K.|T_n|)^2)}$, $N=10$, for *t* and t' textures, we have a positive constant K, and for any $n \neq 0$, $|T'_n| = K*|T_n|$, where $Z_n = \sqrt{(|X_n|^2 + |Y_n|^2)} = \sqrt{(a_n^2 + b_n^2 + c_n^2 + d_n^2)}$. That is to say, the textures are identical except for one geometric transformation. The translation, scale and rotation have no effect on the module of Fourier coefficients. $K = 1/N*(\sum_{n=1,N} (|T'_n|/|T_n|))$ is an estimation of K which minimizes the error on the first N (e.g. 11) coefficients of Fourier.

- The second step, optimization, permits the correct adaptation of the visual thesaurus clusters. The gravity centers of the clusters created in the previous step are computed. The gravity centers are replaced by the most similar feature of the database. This operation also characterizes *k-medoid* algorithms [15]. The replacement of the gravity centers by the most similar image features will avoid the empty clusters.

The algorithm is reiterated in the new visual thesaurus in order to obtain a new partition. The algorithm converges to a stable state by developing at each iteration the distortion criteria. Each iteration of the algorithm should reduce the mean distortion. The choice of the initial visual thesaurus will influence the local minimum that the algorithm will achieve, the global minimum corresponds to the initial visual thesaurus. The creation of the initial visual thesaurus is inspired from the splitting technique [10].

Distortion is a statistical measure used in the clustering algorithm. The experimental results have shown that the distortion values decrease quickly as the splitting number rises. After a quick initial decrease, the distortion values decrease very slowly. Conversely, the entropy increases quickly as the number of splitting rises, and then, it increases very slowly.

- The third step, splitting, decomposes features Y_k of the visual thesaurus into two different features $Y_{k-\varepsilon}$ and $Y_{k+\varepsilon}$, where ε is a random vector of weak energy, and its distortion depends on the distortion of the split vector. The algorithm is then applied to the new visual thesaurus in order to optimize the reproduction features. The splitting step reduces not only the dependency between the initial visual thesaurus and the final visual thesaurus, but also increases the entropy of the visual thesaurus. It means that, the visual features of image databases are grouped homogeneously around the features of the visual thesaurus.

4. IMAGE FEATURES

The color is the first feature considered to describe the image content. It is represented by an autocorrelogram of 256 elements. Each element of the autocorrelogram answers to the question: what is the probability of two pixel separated by k pixels to have the same color. The texture is an important aspect of human visual perception, and it is the second important feature extracted automatically from image regions. The algorithm considered implements Fourier model [22]. Fourier model has very interesting advantages: - the texture can be reconstructed from the features. – it has a mathematical description rather than a heuristic one. - And finally, the model supports the robustness of description to translation, rotation and scale transformations. An important contribution of our representation is our extension of Fourier model to texture description. This extension considers the matching process, and particularly the similarity measure. In this extension, we consider *texture(t)* to be composed of two functions: *x(t)* and *y(t)*. So *texture(t) = (x(t), y(t))*. *x(t)* represents the different level of gray of *x*, and *y(t)* represents the different level of gray of *y*. *t* indicates the different indices of the signal texture. $t = 0, N-1$. N is the period of the function, and N = number of *x* values and *y* values = length of the normalized image. So, we have two series of coefficients $S(a_n, b_n)$ and $S(c_n, d_n)$ that represent Fourier coefficients of *x(t)* and *y(t)* respectively. We consider only eleven coefficients of Fourier that select the lowest frequencies of the sub-band $k \in [0\text{-}10]$. In this extension, we modify the similarity measures (Euclidean distance) in order to consider the coefficients of the two signals *x(t)* and *y(t)*. The choice of implementing Fourier descriptors is based on the fact that it represents any complex texture with only few parameters, for N harmonics, we have $2+N*4$ coefficients.

5. RELATIONSHIP EXTRACTION

Based on the visual thesaurus, the algorithm discovers relationships. Relationships are relationship shared by images of the same databases (*Birds, Animals, Aerospace, Cliffs, etc.*). The extraction of relationships is done in two steps: symbolic clustering and relationship discovering. In the first step, the description of the content of images are transformed into a symbolic form defined in the visual thesaurus. The features of the image are replaced by the most similar features defined in the visual thesaurus. To each feature of the visual thesaurus is associated a symbolic representation in order to manipulate it easily. In the second step, the relationship discovery engine automatically determines common relationships among the image features. These relationships are in the form of *Premise => Conclusion* with a confidence. These relationships are called statistical as they accept counter-examples. Confidence is very important in order to estimate the quality of the relationships induced. In order to estimate the confidence of relationships, we implement two confidence measures: the conditional probability and the implication intensity. The user should indicate the threshold above which relationships discovered will be kept (relevant relationships). In fact, the weak relationships are relationships that are not representative of the shared relationship. Conditional probability allows the system to determine the

discriminating characteristics of considered images. However, this measure presents several drawbacks. That is why we also implemented the intensity of implication [9]. For example, implication intensity requires a certain number of examples or counter examples. When the doubt area is reached, the intensity value increases or decreases rapidly contrary to the conditional probability that is linear. In fact, implication intensity simulates human behavior better than other confidence measures and particularly conditional probability.

In the retrieval task, when the user specifies an image (called source image) as the base of the query, and asks "find images similar to the source image", the system will not match the source image with all images of the database. It will match the source image features with all the target images of the suited databases. The suited databases contain relationships globally respected by the source image. For example, if we have a source image that contains a *"head"* texture, but it does not globally verify the relationships associated to this database, we can deduce the weakness relationship between the source image and the database. The system matches the source image with databases through their learned and stored relationships.

6. "EFFICIENCY" EVALUATION

We have conducted extensive experiments of varied data sets to measure the efficiency of the relationship-content-based query. First, we present the metrics used to measure retrieval performance followed by a description of the data sets. Finally, we present the results along with our observations.

6.1 Evaluation Method

The retrieval system can be evaluated by considering its capacity to effectively retrieve information relevant to a user. It is called the retrieval efficiency. Retrieval efficiency is measured by recall and precision metrics [20], [21]. For a given query and a given number of images retrieved, recall gives the ratio between the number of relevant images retrieved and the total number of relevant images in the collection considered. Precision gives the ratio between the number of relevant images retrieved and the number of retrieved images. $Precision = |relevant \cap results| / |results|$; $Recall = |relevant \cap results| / |relevants|$

Recall and precision values for a system can be represented in a recall and precision graph [19], where the precision of the system is plotted as a function of the recall. This representation enables, for example, measurement of the precision at different recall points.

6.2 Data Sets

We have conducted experiments on a data set which is a collection of images covering a wide range of databases including animals, panorama, archive, flowers, scenery, people, nature, etc. This collection contains around 27331 images and requires 1 Go for data storage and 110 MB for meta data (index), and 36 MB for index after compression.

The collection was hand-picked to provide an assortment of breathtaking, high quality photographs. Included are: - 3,500 high resolution photos, professional quality JPG photos at 240 DPI and 16 million of colors. - 15,000 medium resolution 8 bit color photos (jpg), a stunning variety of photos and special effects. - 7,500 black and white historical photo clips (jpg), amazing shots from memorable moments in history. It includes shots of people, nature, technology, sports, food and more. Historical images portray movie stars, politicians, royalty, sports, heroes, memorable events and everyday life. – 1331 for different resolutions and different semantics of images.

All images are catalogued into broad databases and each image carries an associated description. In this case, manually separating the collection into relevant and non relevant sets was unfeasible due to the size. It is not simple to obtain the exact number of positive results in the large image databases. Instead, we estimate the cardinal of the result set of each query. The pre-categorization of images in semantic databases such as panorama, flowers, etc. simplifies our estimation process. For certain images, we pre-determine by hand and by exhaustive browsing a set of relevant image results. For other images, we start the search by issuing the query involving textual descriptions or visual features, to obtain the first set of image results. The set is refined and expanded by using query by example and exhaustive browsing.

6.3 Results and Analysis

The recall and precision values for our system are computed as follows. *82* reference («query») images are selected from the test collection. The sub-set of images is selected per database (waterfalls, fires, panorama, etc.). For an image reference, we associate a relationship-content-based query, and we associate a content-based query that doesn't use the relationship associated to databases. The threshold used to select multiple relationships is set at *95%* and *90%* for respectively conditional probability and implication intensity confidences. The threshold ε used to determine the stability of the distortion mean is set at *0.05* for both local and global clustering. So, the local clustering algorithm is stopped when *(distortion* D_{s-1} *at iteration s-1 – distortion* D_s *at iteration s) / distortion* D_s at iteration $s < \varepsilon$. Note that in order to perform comprehensive examinations of the number of features in the visual thesaurus, we purposely choose *256* color features and *256* texture features. So the length of the visual thesaurus L is equal to *512* features. The δ offset used to determine the number of split is defaulted to $L/2 = 256$ for texture and color.

Since it is not possible to retrieve all relevant images, our experiment evaluates only the first 50 ranked images. For experimental simplification, we don't use the similarity ranges. In the content-based queries, each query is decomposed into texture and color sub-queries.

Judging from our experiment results, it is obvious that the use of relationship leads to improvements in both precision and recall over the majority of queries tested. The average improvements of relationship-content-based queries over content-based queries are *23%* for precision and *17 %* for recall. Precision and recall are better for relationship-content-based queries (queries that mix visual features and relationship associated to databases) than for queries that use only visual features (colors and textures). We observe that for a sub-set of images, the precision of content-based queries is better than the precision of relationship-content-based queries. A possible explanation for this

is that images of the same semantic databases may have large variations with dissimilar visual descriptions (images of people, images of holidays, etc.), and images from different semantic databases might share common visual features. So relationship-content-based retrieval that selects only one database may forget images that are visually similar to the query examples, and belong, semantically, to several databases. We observed that the general principle of «the larger the retrieved set, the higher the recall, and the lower the precision» is observed.

7. CONCLUSION

We presented the interest of relationship discovery in the traditional architecture of content-based indexing and retrieval systems in large image databases. The relationship discovery engine extracts relationships that characterize image databases. From the features of images belonging to the same database, the system finds the pattern of interest in the form of relationships which are qualified on the basis of two confidence measures (conditional probability and implication intensity). These induced relationships are very helpful for: - the comprehension of the considered database, - the automatic categorization of new images during their insertion in large databases, - obtaining results with more semantics, and improving the retrieval process. More generally, we believe, strongly, that discovery of hidden relationships from multimedia will play in the future a revolutionary role in the content-based multimedia indexing and retrieval.

REFERENCES

[1] Chang S. F., Smith J. R., Beigi M., Benitez A., « Visual information retrieval from large distributed online databases », Communications of the ACM, 40:12, pages 63-67, 1997.

[2] Wendy Chang, Gholamhosein Sheikholeslami, Jia Wang, Aidong Zhang, « Data Resource Selection in Distributed Visual Information Systems », IEEE Transactions on Relationship and Data engineering, 10(6): 926-946, November-December, 1998.

[3] Chua T., Teo K., Ooi B., Tan K., "Using domain relationship in querying image databases", Multimedia Modeling, Towards the information Superhighway, France, Toulouse, 12-15 November, 1996.

[4] Danzig P., Li S., Obraczk K., «Distributed Indexing for Autonomous Internet Services, Technical Report, Dept. Of Computer Science, University of South California, June 1992.

[5] Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R. « Indexing by latent semantic analysis », Journal of the American society for Information Science, 41:391-407, 1990.

[6] Djeraba C., Bouet M., "Digital Information Retrieval", ACM CIKM'97, Las Vegas, USA, November 1997.

[7] Djeraba C., Bouet M., Briand H. "Concept-Based Query in Visual Information Systems". IEEE ADL 1998: 299-308.

[8] Faloutsos C., Flickner M., Niblack W., Petrovic D., Equitz W., Barber R. "Efficient and Effective Query by Image Content". Research report, IBM Alameda Research Center, 1993

[9] Gras Régis, THE EISCAT CORRELATOR, EISCAT technical note, Kiiruna 1982, EISCAT Report 82/34, 1982.

[10] Gray R. M. « Vector Quantization », IEEE ASSP Mag., pages 4-29, April 1984.

[11] Amarnath Gupta, Ramesh Jain "Visual Information Retrieval", A communication of the ACM, May 1997/Vol. 40, N°5.

[12] [12] Huang J., Kumar R., Zabih R., "An automatic Hierarchical Image Categorization Scheme", ACM MM-98, Bristol, 10 pages, 1998.

[13] Ramesh Jain: Content-based Multimedia Information Management. ICDE 1998: 252-253.

[14] Kahle B., Medlar A., « An Information System for Corporate Users: Wide Area Information Servers. ConneXions – The Interoperability Report, 5(11):2-9, November 1991.

[15] L. Kaufman and P. J. Rousseeuw. "Finding Groups in Data: an Introduction to Cluster Analysis". John Wiley & Sons, 1990.

[16] Koster M. « ALWEB: Archie-like Indexing in the Web. Computer Networks and ISDN Systems, 27(2):175-182, 1994.

[17] Linde Y., Buzo A., Gray R. M. « An algorithm for Vector Quantizer Design », IEEE Trans. On Comm., Vol. COM-28, N° 1, pages 84-95, January, 1980.

[18] Pentland A., Picard R. W., Sclaroff S. "Photobook: Tools for Content-Based Manipulation of Image Databases", in Proc. of SPIE-94, pages 34-47, Bellingham, Washington, 1994.

[19] Raghavan, V., Jung, G., and Bollman, P., "A Critical Investigation of Recall and Precision as Measures", ACM Transactions on Information Systems 7(3), page 205-229.

[20] C. J. Keith van Rijsbergen "Information retrieval", Second edition, London: Butterworths, 1979.

[21] Salton Gerard "Automatic Information Organization and Retrieval", McGraw Hill Book Co, New York, 1968, Chapter 4.

[22] C. T. Zahn, R. Z. Roskies, « Fourier descriptors for plane closed curves », IEEE Trans. On Computers, 1972.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

SEMANTIC CONTENT-BASED RETRIEVAL IN A VIDEO DATABASE

PRAMOD K. SINGH
Faculty of Information Technology, University of Technology Sydney (UTS), PO Box 123, Broadway, NSW 2007, Australia

A.K. MAJUMDAR
Department of Computer Science & Engineering, Indian Institute of Technology (IIT), Kharagpur, India

Abstract

To take advantage of the rich information content of video data, data management systems that allow efficient and effective storage, indexing, and retrieval of these data are essential. The need of managing temporal information of video data is common to many application areas. Among them, we focus on echocardiogram video data management. In this paper we describe an approach of Semantic Content-Based Retrieval (SCBR) of video data using object state transition (OST) data model which allows storage and indexing of echo video at different levels of abstraction based on semantic features of video objects.

Keywords: Echocardiogram, video data, video segmentation, object extraction, object state transition.

1. Introduction

The key features of a video data are the spatial and temporal semantics associated with it. Semantic Content-Based Retrieval (SCBR) requires multidisciplinary research effort in areas such as computer vision, image processing, data compression, databases, information system, etc. [8,7]. Content-based retrieval poses special challenges, which call for new techniques allowing an easy development of video database applications [6]. The content-based retrieval capabilities are composed of two sets; appeared historically in this order: Query by Example (QBE) based on visual features (color, shape, motion, etc.) and semantic content-based query. Much research has been done in the area of indexing and accessing video based on its visual features; however relatively little research has been devoted to the problems of indexing sharing, querying and browsing the semantic content of video data [6].

Those cases where a database is queried by specifying semantic contents allows user to specify a subjective description of a query condition. This type of query specification is referred as query-by-subject in [2]. In query-by-subject a keyword representing a semantic content is specified. The semantic content implied in video data is extracted from raw data in order to evaluate a query. A simple way of managing the semantic of video data is to annotate a video with text. The semantic content annotation for a video data is assumed to be defined by a human where extraction of content through image processing from video is not possible.

In another approach query-by-subject is implemented by providing the system with a rule base or knowledge base. Here knowledge is used for feature extraction from raw data; content matching, query analysis and translation and so on. Despite the realization of the central role of video database in many areas of application, a little research work has been done on finding semantic foundations for representing and querying information.

The need of managing spatial and temporal information of video data is common to many application areas. Among them, we focus on echocardiogram video data management. In this paper we exploit the possibility of semantic content-based querying of an echocardiogram video database through object state transition data modeling and indexing scheme.

2. Background & Motivation

For diagnosis of cardiac problems, echocardiogram is a very important tool. The echocardiogram captures the images of heart chambers at various states of their functioning. The video stream generated by this technique provides valuable information regarding the types of cardiac abnormalities and their evolution over time. The radiologists often try to satisfy the following queries from echocardiogram video database.

- Find the rate of contraction/expansion of heart chamber of the patient.
- Find the volume of the ventricles and atria at peak of expansion/contraction.
- Retrieve details of patients whose rate of expansion/contraction of a particular heart chamber is more or less than a given standard rate.
- Display video clips where rate of contraction/expansion of a heart chamber or peak volume of fully contracted/expanded heart chamber is more, less or equal to the given value.

Depending on the results of these queries, it may be pertinent to go for further queries, for example, for a pathological heart the rates of expansion/contraction and/or duration of fully expanded/contracted state do not match with the normal values, and then the following queries can be made:

- Which of the six segments of LV (Left Ventricle) has maximum anomaly with respect to volume plot?
- What is the maximum pressure during systole (one state of functioning of heart)?

The results of first query can identify which cardiac artery is the cause of the problem of a possible muscular dystrophy of LV (Left Ventricle), whereas the result of the second query may indicate a case of mitral stenosis or regurgitation of blood into LA (Left Atrium).

Furthermore, from the volume data of LV, pressure can be calculated by simple approximation and this can, in turn, give hints to amount of blood flowing out of LV. These results are very much helpful for a cardiologist to diagnose heart diseases of a patient.

3. Video Data Modeling for SCBR

Our approach introduces a state-transition representation of a video-object in which the states of an object are described by the values of its associated attributes e.g., volume of a ventricle. All attributes, which are required by the potential semantic content-based queries in an application, are recognized as dynamic attributes.

We have proposed an Object State Transition (OST) data model [1] which supports standard features of object oriented data modeling and adopt some of the concepts from Video Semantics Directed Graph (VSDG) data model [4] and Object Composite Petri-Net (OCPN)[5] model.

The objects in the OST model can be primitive objects such as integer, float, string, etc. or composite objects. A composite object is formed by aggregation of other objects, which will be treated as its parts. The parts of an object which have primitive values is referred to as its attributes e.g., volume of the ventricle. The operations that can be performed by an object is specified by appropriate methods. The objects in the OST model can be static or dynamic. The temporal properties of a dynamic object is represented by its states and state transitions.

An object $\boldsymbol{X}$ at a particular time instance can be in a given state. The set of possible states of an object $\boldsymbol{X}$ will be denoted by $\boldsymbol{S_x} = \{\boldsymbol{s^x_1}, \boldsymbol{s^x_2}, \ldots, \boldsymbol{s^x_m}\}$ for some integer m. Once an object, say $\boldsymbol{X}$ reaches a particular state, say $\boldsymbol{s} \in \boldsymbol{S_x}$ it will remain in that state for specified time interval, say $\boldsymbol{T}$. This time interval, however, may depend on the state $\boldsymbol{s_i}$ of the object $\boldsymbol{X}$, where $\boldsymbol{i} \in [1 \ldots m]$. The state of the object $\boldsymbol{X}$ may change only at the end of this interval i.e. the object may move to a new state $\boldsymbol{s'} \in \boldsymbol{S_x}$ after the time interval $\boldsymbol{T}$. An object is said to execute periodic behavior if it repeats a sequence of states. Thus, an object $\boldsymbol{X}$ exhibits periodic behavior if it passes through the sequence of states say $\boldsymbol{s_1}, \boldsymbol{s_2}, \boldsymbol{s_3}$ repeatedly, with total time period being determined by the sum of the time required for the transitions $\boldsymbol{s_1} \rightarrow \boldsymbol{s_2} \rightarrow \boldsymbol{s_3} \rightarrow \boldsymbol{s_1}$. The objects having same structure and behavior will constitute a class and a particular object will be treated as an instance of its class.

As mentioned above, the most significant difference of OST objects with objects in standard object-oriented databases is the concept of state and state transitions. While defining the states and state transitions of a composite dynamic object, one has to look into the states of constituent objects, which have dynamic behavior. Thus the state of the heart will depend on the states of ventricle, aorta, atrium and valves. Such dependency of the state of composite dynamic object on the states of its dynamic parts will be expressed by suitable rules. A sample rule is,

- The **heart** is in the **systolic** state *if* the **ventricle** is in **Isovolumic-Contraction** *or* **Isovolumic-Relaxation** state *and* **atrium** is in **Atrial Systole** state. **… … …** (A)

An inherent assumption while specifying such rules is the concept of simultaneity. This means that all the constituent objets are simultaneously in their specified states. For example in (A), the ventricle and the atrium are simultaneously in the states **Isovolumic-Contraction** and **Atrial Systole**, or **Isovolumic-Relaxation** and **Atrial Systole,** respectively**.**

3.1. Feature Extraction

Extraction of a visual feature (color, shape, texture, etc.) of an individual frame from the video is one important requirement for analysis of contents. The strategy involves the extracting of shape attribute of a physical object present in the video frames with the help of bounding volumes that describes the spatial projection of an object in three dimensions. The bounding volume is computed with reference to a coordinate system with an origin at the lower left corner of each frame. For extraction of a single cavity (bounding volume of an object), we are using active contouring based relaxation technique [3].

Object extraction depends on the type of video clip. For example, apical view of echocardiogram may contain an image of one or two heart chambers. Accordingly, we classify the video clip, while performing the echocardiogram as A1 or A2 respectively. Based on these classifications, we can determine the spatial relationship among objects using a 2D-string based pre-defined rules.

3.2. Knowledge based segmentation of Echo-video

Temporal modeling of a video clip is crucial to describe segment and episode in the clip. A segment of video clip is a sequence of frames, which collectively represents one state of the system i.e. composite object. For example, sequence of frames which shows the systole or diastole state (two possible states of functioning of heart as a system) of heart in the echo video clip. Further, a segment is partitioned into episodes and each episode can express a state behavior of an object in a sequence of frames. For example: the occurrence of a maximum expansion of one heart chamber in an echo video clip can be considered as an episode.

We propose an object state change segmentation-process, which involves analysis of shape or size (features) of a video-object between two consecutive frames. Our approach of segmentation is based upon the following considerations:

1. Video is first segmented into shots using twin comparison algorithm. Each shot represents a sequence of frames in which an object of interest is present. For instance, an echo video is partitioned into shots in which heart chambers are captured in a particular view (e.g., apical view).
2. Shots are segmented based on the states of an object.
3. A state of an object is classified based on the value of shape metrics, using suitable rules:

    - If the absolute difference of shape of an object, e.g. difference of cavity area of heart chamber, between two consecutive frames is less than a threshold value, object is considered in static state between these two frames.
    - A sequence of frames in which the shape difference between any two consecutive frames satisfy the above criteria is a segment of video representing static state of the object.
    - If the absolute shape difference between consecutive frames is more than the threshold value, the object is in dynamic state during those frames and a segment of video representing dynamic state of the object is obtained.

The dynamic state of an object can be further specified depending on the nature of shape changes e.g., a dynamic state of a heart chamber may be contracting or expanding. For instance, if the cavity area of a heart chamber is increasing in the consecutive frames of segment of echo video, then the segment is representing expanding state of that heart chamber. The classification of static state of an object also depends on the occurrence of previous state. For example, in echo video a static state of a heart chamber is called fully expanded state if the previous occurred state is expanding (dynamic) state.

The state behavior of an object can be described by observing the total duration that object appears in a given video clip and their relative movement over all the frames. Temporal information of objects can be captured by specifying the changes in the spatial parameters associated with the bounding volume of the objects over the sequence of frames. At the finest level, these changes can be recorded at each frame. Although this fine-grained temporal specification may be desirable for frame based indexing of video data, it may not be required in most of applications. Alternatively, a coarse grained temporal specification can be maintained by analyzing frames.

4. Content based Querying

Based on the above mentioned knowledge based segmentation scheme, a Video Data Abstraction (VDA) tree is formed, which facilitates the storage and retrieval of video data for satisfying queries from the database.

4.1. Video Data Abstraction (VDA) Tree

The video data abstracting are based on perceiving a video at different levels of abstraction. At the lowest level the video comprises a series of frames, each of which is a still image. Frames are grouped into shots (shown as Echo-shots in figure -1), which are a continuous uninterrupted sequence of frames recorded at same capturing view. Further, an echo-shot is partitioned into segments and episodes depending on the states of objects in the video frames. Thus, a video can be represented at multiple levels of data abstraction. Figure – 1 shows its granularity at different levels of abstraction.

A video clip is partitioned into episodes with reference to an object. Here, an episode corresponds to a state of a heart chamber e.g., ventricle is in the expanding state. If a video clip contains only one object i.e., image of only one heart chamber is present in each frame of the clip. Thus, episodes correspond to the state of this chamber only.

However, if more than one objects are present in the frames of the video clip e.g., images of two, three or all four heart chambers are present in each frame, the clips may have partitioned into multiple episodes corresponding to the state of the objects in the frame. Thus the same frame may be included into number of episodes.

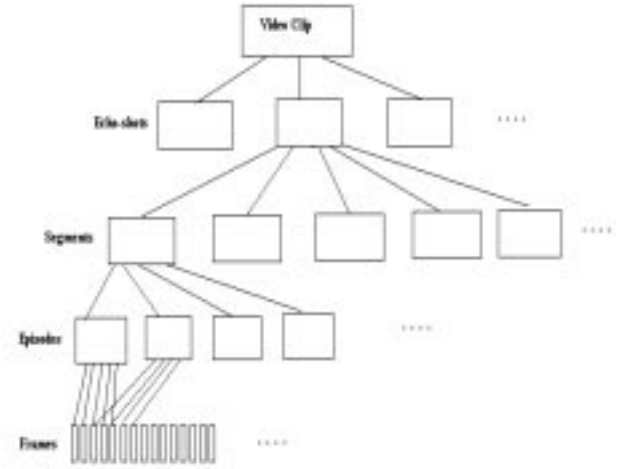


Figure 1: Video Data Abstraction Tree

4.2. Navigation of VDA Tree for Query Evaluation

The data abstraction tree depicts a hierarchy of representation of video data contents for content-based indexing and retrieval. The tree is navigated for satisfying various kinds of queries. Segment level video structuring can provide facility for content-based access of video parts. Here temporal video can be divided along the time axis where each part of video represents the occurrence of a defined state of the composite object e.g., heart is in systole state. This is a macro granularity of video data. The separation of video into such a meaningful part is a difficult task. The best way of to achieve such fragmentation is by first dividing it into atomic units called episodes. An episode provides a mini granularity of video data. The task of achieving such granularity is called video segmentation. Once the parts of the video representing episodes are identified they can be indexed for a future search mechanism. The last level of granularity in video structuring is an individual frame. No effort is needed to do frame separation in digital video. Such a separation is meaningful for indexing based on the presence of objects and their spatial relationship in the frame.

Queries related to a particular heart chamber can be dealt based on episode information. And heart related queries could be addressed by a set of segments. The segment of the echo clip represents the episode information of all objects occurs simultaneously in the video frames

5. Implementation

A prototype system has been developed under Window NT environment on a multimedia PC. The system was developed on top of a commercial object oriented database management system named, Jasmine v1.2. User Interface is developed using Java and Jasmine Java Proxies (JJP). Visual C++ was used, for compilation of query methods written in ODQL (a query language, like SQL, supported by Jasmine).

5.1. Logical Data Model

The logical data model of our system uses OST data model framework of video-objects, set of videos, and type abstraction hierarchies to represent the data generated by the application domain.

Schema is designed for capturing the semantics of our problem and fulfills a number of requirements, the most significant of which are the ability to track the functional behavior of heart chamber over time and the ability to capture the spatial features of the heart chamber.

For representation of standard patient the entity named, Patients have attributes such as a unique PatientId, PatientName, PatientHistory, and list of VideoId. VideoClips entity contains attributes like VideoId, VideoName, VideoSize, VideoLength, VideoCaptureRate, DateOfCapture, digital video data (AVI file), etc.

The functionality of heart is captured using HeartStates entity, which contains attributes like StateId, StateName, and StateDescription. Subparts of heart e.g. left ventricle, are mapped into HeartChambersPhases class, which have attributes like PhaseId, PhaseName, HeartChamberName, and PhaseDescription. The HeartStates entity and HeartChambersPhases entity contains definition of states through which heart and heart chambers undergo. VideoObjects entity is used to represent individual objects that are captured in the videos e.g., right atrium, mitral valve, left ventricle etc., which have attributes such as unique ObjectId, ObjectName, and ObjectDescription.

A video clip can be segmented into number of shots, each representing a particular view e.g., apical view of echo capturing. For representing echo shots, the entity EchoShots has attributes like ShotId, NameOfCaptureView, StartFrameNumbert, EndFrameNumber, and list of ObjectId. Shots are partitioned into segments depending on heart states, that is, a segment can represent each heart state. So video clip entity has 1-to-n relationship with EchoShots entity and each EchoShots has similar relationship with Segments entity. Segments have attributes SegmentId, StartFrameNumber, EndFrameNumber. A segment can be further classified according to the states of heart chambers, which is represented by an episode. Thus, a segment entity has 1-to-n relationship with Episode entity, which contains attributes like EpisodeId, StartFrameNumber, EndFrameNumber MinimumArea, and MaximumArea of heart chamber cavity in the frame sequence covered in the Episode.

HeartStates entity has 1-to-1 relationship with Segments and likewise HeartChambersPhase has also similar relationship with Episodes. All such relationships are shown in Figure-2. In the system each above-mentioned entity is mapped as a class of Jasmine database.

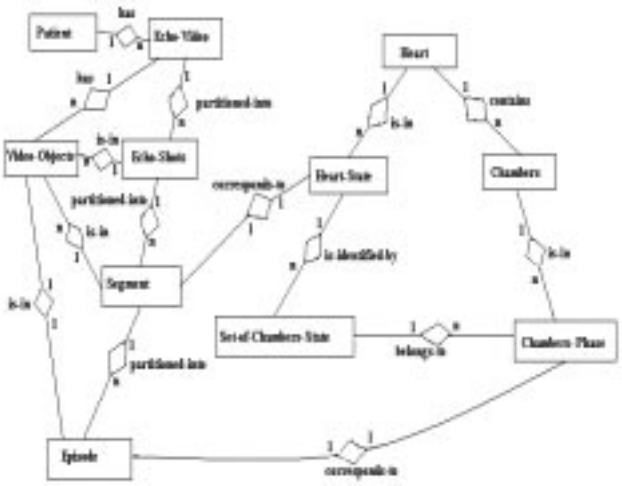


Figure – 2: Logical Data Model for Echo Video Database

5.2. Query Processing

The system responds to a query by retrieving the relevant frame and presenting the requested data to the user. The query results can be reused to place another query or to display more detailed information. The query processing algorithms for handling different types of queries efficiently on the developed prototype system based on our proposed data modeling approach, which facilitates the indexing of derived semantic content of video data for fast and effective retrieval, is given as under and the complexity of these procedures are explained.

5.2.1. Elementary Queries

Query type 1 of form: "Display a list of all the objects who are currently in a particular class in the database".

Example: "Display all objects of VideoClips class"
Algorithm: ObjectLists

/* find all the objects of the class of given name ClassName and store in a object bag */
Begin
ProcClassBrowse(string ClassName)
1. ObjectBag=ClassName from ClassName;
2. Display each element of ObjectBag in the "Result Display Window";
End.

Complexity: Finding and displaying all objects currently present in a class can be done in constant time, the time required is proportional to number of objects in the class.

Query type 2 of form: "Find and display all VideoClips of a given patient currently available in the database"

Example: "Find and Display all videos of patient named, Manoj"
Algorithm: PatientVideos

/* get all video id. numbers of videos of a given patient from Patients class */
/* for each video id. find and display video from VideoClips class */
Begin
ProcVideos(string PatientName)
1. Find the object of given patient from Patients class;
2. Get the list of all videos from the above object ;
3. For every element of video list
   - Find object from VideoClips class;
   - Fetch the digital video part from the object;
   - Display video in the result window;
End.

Complexity: The time required is proportional to the number of videos related to the given patient. The total cost includes finding time of object from Patients class and searching time of individual video from VideoClips class.

5.2.2. Object State Occurrence Queries

Query type 1 of the form: "Find and display segments or episodes of a given video where an object is in a given state of heart (a composite object)"

Example: "Find and Display video segments of HeartVideo1 where heart is in Systole state"
Algorithm: SegmentFind

/* construct a search string by clubing VideoId and StateId */
/* find all objects from Segments class where a part of SegmentId is equal to search string */
Begin
ProcSegmentsOfVideo(String VideoName, string StateName)
1. Find the VideoId of the given video object from VideoClips class;
2. Find the StateId of the given state object from States class ;
3. Construct search string by combining VideoId and StateId;
4. SegmentObjectsBag = Segments from Segments where part of SegmentId is equal to search string;

5. For each element of SegmentObjectsBag, display the part of the given video;

End.

Complexity: The cost is proportional to the number of segments in the video plus construction cost of search string.

Query type 2 of the form: "Find and display episodes of a video where heart states and one phase of a heart chamber is given"

Example: "Find and Display video episodes of HeartVideo1 where heart is in Systole and left ventricle is contracting"

Algorithm: EpisodeFind

/* construct a search string by clubing VideoId, StateId, ObjectId, and PhaseId */

/* find all objects from Episodes class where a part of EpisodeId is equal to search string */

Begin

ProcEpisodesOfVideoSegments(string VideoName, string StateName, string ObjectName, string PhaseName)

1. Find the VideoId of the given video from VideoClips class;
2. Find the StateId of the given state object from States class ;
3. Find the ObjectId of the given video object from VideoClipObjects class;
4. Find the PhaseId of the given state of object from Phases class;
5. Construct search string by combining VideoId, StateId, ObjectId, and PhaseId;
6. EpisodeObjectsBag = Episodes from Episodes where part of EpisodeId is equal to search string;
7. For each element of EpisodeObjectsBag, display the part of the given video;

End.

Complexity: The cost is proportional to the number of episodes in the video plus construction cost of search string.

5.2.3. Conjunctive Queries

Query type 1 of the form: "Find volume/rate of a chamber of heart of a given patient at a given state of heart chamber"

Example: "Find volume of LV at fully expanded state of patient named Raju"

Algorithm: VolumeComputation

/* find the videos of the given patient */

/* find episodes of expanded states of LV in individual video which related to patient */

/*compute average volume of LV at expanded state from each video*/

Begin

ProcVolume(String PatientName, string ObjectName, string PhaseName)

1. Retrieve object of given patient name from Patients class;
2. Get list of videos from the object obtained in step 1;
3. For each element of video list obtained in step 2, do
   - Find the video name;
   - Construct two search strings: one by clubing PatientId and VideoId and second by combining ObjectId, and PhaseId;
   - Find episodes from Episodes class using search string;
   - Compute volume from the data available in each object of Episodes class;
   - Compute average volume;
   - Display volume along with video name;

End.

Complexity: The time required in processing is equal to the result of multiplication between the number of videos (related to the given patient) and cost of average volume computation; whereas average volume computation cost is proportional to the number of episodes of interest in each video, plus construction cost of search strings.

Query type 2 of the form: "List all patients(videos) whose(where) volume/rate of given object is more than a given value"

Example: "Display list of all patients whose volume of LV is more than a given value at fully expanded state"

Algorithm: PatientsList

/* find the episodes having LV as an object at fully expanded state */

/*compare the average volume of the LV of one video and compare with given value*/

/* if condition satisfied, find patient object from Patients class using PatientId (chopped from EpisodeId) and display the name of patient */

Begin

ProcPatientsList(String ObjectName, string PhaseName, real Value)

1. Find ObjectId of the given object name from VideoClipObjects class;

2. Find PhaseId for the given phase name from HeartChambersPhases class;
3. Construct search string by combining ObjectId and PhaseId;
4. Retrieve objects of Episodes class where part of EpisodeId is equal to search string;
5. Group the episode objects based on VideoId (which is part of EpisodeId);
6. For each group obtained in step 5 Compute average volume and compare with given value
   - if given condition is satisfied then fetch the name of patient from Patients class using PatientId (which is part of EpisodeId);
   - display the name of patient and required details;

End.

Complexity: The time cost is equal to the result of multiplication between the number of episode group (obtained at step 5) and combined cost of average volume computation & finding the name of patient; whereas average volume computation cost is proportional to the number of episodes of interest in each group; plus construction cost of search strings and finding & grouping the episodes from Episodes class.

5.2.4. Compound Queries

These queries involve the relationship between different objects of the video segments. For example, what is state of atrium (a video object) when left ventricle is in expanding state. In general, answers to compound queries are found by combining the result of two separate queries or by directing the output of one query as an input to the another query.

Example: Find the objects of the video clip (under consideration) causing the abnormal behavior of heart at the systole state.

5.3. A Sample Query Processing

A graphical user interface specifically designed and developed for the echocardiogram database. This enables users to specify the query by clicking into respective button. Let us consider a typical query given below:

"What is average rate of expansion of heart chamber of (patient) Manoj"

We begin by opening the Query Table (a user interface for submitting the query to system), which provides format to specify the predicate in the database. The query to be answered needs to find the patient entity that have value of PatientName attribute as Manoj and the video clips related to this patient is accessed from the videos attribute which keeps the array of VideoId of the concerned video clips. PatientId number and StateId number of expansion state ids used for accessing the episode entity related to the query and computation of average rate of expansion is done for all video clips related to Manoj and the result is displayed in the result window.

Figure – 3: Query Result Table

5.4. Medical Significance of Semantic Content-Based Query Results

The proposed system can be used as a diagnostic decision support tool and as a tool for medical research and training. Particularly, interesting possibility in this context is to provide physicians and other personnel with the capability of content-based access to medical videos for the purpose of assisting diagnostic decision-making and management of further treatments. The proposed system's framework is very much useful for therapeutical management of patients who have undergone some kind of cardiac operations and/or periodically followed up, to prevent sufferance from new complications. Thus, database models temporal therapy, relative to previous and current therapies, and diagnosis, relative to previous and current pathologies, as heart rate, systolic and diastolic blood pressure, and others.

For instance, result of a query related to expanding rate of left ventricle (LV) as shown in the Figure - 3 is providing useful information to a physician about his/her medication pattern. Suppose patient named, Manoj is undergoing for treatment and his echo video is being prepared after a fix interval of 15 days, initially his LV was having expansion rate 117.6367 mm/sec (as shown, expansion rate of HeartVideo3) post medication it has improvement in the subsequent echo tests, this may give an indication that the treatment is in the right direction (positive effect) or wrong.

6. Summary

Our framework provides an appropriate mechanism for semantic video indexing. A semantic index provides keyword based searching and brings video clips at par with text database.

Once the video clips are automatically annotated in object state tansition model, video search reduces to text-search using the keywords. Use of proposed approach in conjuction with the query-by-example paradism, can prove to be a powerful tool for content-based multimedia access.

7. Conclusion

The proposed methodology of semantic content based retrieval of video data employs computer vision and image processing techniques to automate the construction of the video databases based on the object state transition model. The main contribution of this work is to integrate the techniques for capturing and indexing the derived semantic contents of video data for fast retrieval using various query processing algorithms. We show how to use this proposed technique through the implementation and experimentation on a synthetic, but realistic, database of cardiogram video. The proposed approach of modeling and content-based retrieval of video information is applicable to a wide variety of domains where objects state of transition are of repetitive in nature, such as medical video, engineering and scientific simulation etc.

Future development includes the detailed specification, implementation and prototype testing under more sophisticated visualization and annotation techniques. In addition, techniques for automated analysis of extracted objects of video stream and generation of segments and episodes are also being explored.

8. References

1. A.K. Majumdar, Jayanta Mukkarjee, B. Acharya and P.K. Singh, "An Object State Transition Model for Echocardiogram Video Data". Intl Conf on Multimedia Proc. and Systems, August 2000.
2. A. Yoshitaka and T. Ichikawa, “A Survey on Content-Based Retrieval for Multimedia Databases”, IEEE Trans. On Knowledge and Data Engineering, Vol. 11, No 1, Jan/Feb 1999, pp 81-93.
3. B. Acharya, J. Mukharjee and AK Majumdar, "Two-phase Relaxation Approach for Extracting Contours from Noisy Echocardiogram Images", ICAPRDT'99.
4. Day Y.F., Dagtas S., Iino M., Khokhar A. and Ghafoor A. “Object Oriented Conceptual Modeling of Video Data”, Proc. IEEE Conf. Data En*g.*, 1995, pp401-408.
5. Little T.D.C. and Ghafoor A. “ Interval-Based Conceptual Models for Time Dependent Multimedia Data”, IEEE Trans. Knowledge and Data Eng., Vol. 5, No. 4, August 1993, pp 551-563.
6. M.S. Hacid, C. Decleir and J Koulomdjian, “A Database Approach for Modelling and Querrying Video data”, IEEE Trans. on Knowledge and Data Eng., Vol. 12, No 5, September/October 2000, pp 729-750.
7. N Dimitrova, “The Myth of Semantic Video Retrieval”, ACM Computing Surveys, Vol. 27, No. 4, December 1995, pp 584-586.
8. “Special Issues in Video Information System”, ACM Transaction on Information System Vol. 13, No. 4, October 1995.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

An Interactive Environment for Kansei Data Mining

Nadia Bianchi-Berthouze
Database Systems Lab,University of Aizu
Tsuruga Ikki Machi, Aizu Wakamatsu
Japan

ABSTRACT

Kansei engineering is a relatively recent discipline aimed at understanding and modelling (1) how the user's brain process subjective information and (2) how this information can be manipulated by a computer. In this paper we address the modelling of visual impression from the point of view of multimedia data mining. Visual impressions are impressions experienced when observing images. They are highly subjective, complex and difficult to explicit. We propose a methodological approach that takes into account the large amount of information involved in the mapping between images and visual impressions arising in an observer and eventually the way the observer expresses such impressions. From a computational point of view, the modelling process integrates different techniques of multimedia data mining to learn associations between image characteristics and impression words. The user assumes an active role in directing the system's mining activity through mechanisms of externalisation. The externalisation process is supported by a conceptual space endowed with tools that allow the user to express his/her mental process and naive models into formal specification. A WEB based meta-search engine to retrieve images by impression words has been developed. It is used as a support in the close loop of creating and testing new modelling hypothesis.

KEY WORDS: image retrieving, user modeling, subjectivity, multimedia data mining, human machine interaction, adaptation.

1. INTRODUCTION

The rapid evolution of Internet services has led to a constantly increasing number of Web search-engines [1,2,3,4,5] that allow user to get large amount of multimedia information simply by entering some multimedia keywords. An important issue is therefore to facilitate users to get the right information among the thousands received. Many efforts are made in this direction by personalising the search engines. However still little attention is paid to most subjective aspects of user personality that evidently play an important role in our activities, goals and choices.

Web search engines use the typical information retrieval paradigm of a few words to characterise the information to be searched, and then match them against very large numbers of documents, in particular, web pages. The results are often not very precise, partially because of too poor user models. It is even truer when the requested information is related to the user subjectivity. When we are looking for images or song, mostly it is the impression conveyed by these media which should be the selection key. However how to express or to handle such kind of request, how to differentiate a same request from two people with a different personality and finally how to handle the variability intrinsic to our more subjective aspects of our personality, are very complex issues. This study, called Kansei engineering in Japanese, is a quite recent discipline aimed at understanding and modelling how the user's brain processes subjective information and how a computer can possibly handle such information.

A model of user subjectivity based on psychological user profile or on limited training sets is not sufficient. Our subjectivity is very complex because it is affected by our goals and by our past experience. Similar situations or similar information can be judged differently even by a same person according to the experience recalled by his/her brain [6]. Thus the system must be able to continuously learn and adapt to the user through its use. This requires the ability to collect and analyse large quantities of information in order to create a more comprehensive model of the user subjectivity.

In this paper, we address a particular aspect of kansei engineering, which is visual impression. Visual impressions are impressions experienced when observing images. They are highly subjective and difficult to explicit. We propose a methodology to dynamically create kansei user models in order to allow the retrieval of images from the web by using impressions as keywords. The modelling process aims to model the mapping between low-level features of the images, responsible for the visual impression, and the impression word used by the user to label such visual impression.

Kansei Modelling has been dealt with using different computational techniques, such as neural networks, fuzzy sets, genetic algorithms, etc. [7,8,9,10,11,12,13]. However these works are characterised by a static approach, where models are created from a fixed set of features and

relevance feedback from the user. The variability characteristic of any subjective information that is intimately related with our personal experience has not been taken into account.

We propose here a hybrid multimedia data mining approach that takes advantage of the large amount of multimedia data and user feedback gathered when using KDIME [14], a forking and evolving kansei retrieval system. From a computational point of view, our approach integrates different techniques [15] in use in multimedia data mining [16] to model different aspects of this mapping. Clustering techniques are used to identify nuances in the use of a same impression word, while classification techniques are used to learn associations between image characteristics and impression words. In order to identify the correct patterns among the many singled out by the system, co-operation of the user is crucial. The role of the user is to direct the system mining activity through mechanisms of externalisation. Using natural language and visual examples, the user attempts to explicit what should be the focus of attention in the image considered and sheds light on the inconsistencies detected by the system in the user's feedback.

The paper is organised as follows. After having defined "subjective visual impressions", the architecture of KDIME and its use are briefly introduced. Then, the kansei multimedia mining environment (K-MMminer), along with the integrated functions used to clean the data and to mine them, is described. Finally, we describe the externalisation process by which the user is involved in order to direct the mining process.

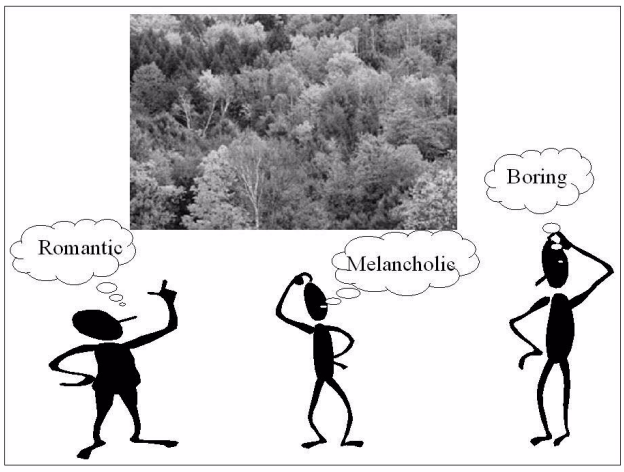


Fig. 1: Visual impression and verbal language as a communication mean

2. SUBJECTIVE VISUAL IMPRESSION

In this paper we propose an approach to model the complex mapping between still images and impression words used to express the visual impression that arises in the observer (see fig. 1). The complexity of this mapping derives both from the complexity of the information in the observed image and from the role that other factors such as personal life experience, state of mind, goal, etc. of the observer play in his/her visual subjective experience [6]. The impression that arises when observing a landscape derives not only from the characteristics of the landscape itself, with its colour, objects, etc. but also from the past experience that our brain recalls when triggered by those characteristics. This is due to two correlated factors:

(a) our state of mind and our life experience change over time and the experiences that our brain can recall are different and even generate opposite impressions;
(b) our visual system uses selection and attention mechanisms to filter the information from the observed object to the brain [17].

Continuing with the characterisation of visual impressions, we have to consider the way in which these are communicated and that is the main point of this research. We communicate our subjective experiences and our impressions using different channels such as facial expressions, body language, metaphors and in particular natural language. Because computers are mainly dis-embodied, we focus our work on natural language as the mean for communication. It will be interesting in the future to extend it to the others modalities [6,18].

Natural languages, as rich as they might be, do not have a one-to-one relation with our subjective impression. How many times do we feel that words do not really express what we want to say? or as in this case, what we feel? However language is a powerful tool that we exploit most of the time. Words are associated with a definition given in the dictionary but these definitions can be very fuzzy especially when aiming to define subjective states [19]. We can use a same word to convey different nuances of a subjective impression or we can use a word in a metaphorical way. Even the way we refer to our impression can change because our language ability, e.g. the richness of our language, will vary over time as well as our ability to discern among visual experience. At the early stages, a child will only be attracted by bright colours [20] but while growing up, s/he will learn to appreciate finer differences in her/his visual experience. This is true for other modalities of subjective experience. For example, evolving music listening skills can possibly allow us to detect new aspects of a melody and change our impression.

3. K-DIME

K-DIME, or Kansei Distributed Information Management Environment [14], is a software environment that enables users to retrieve images from the Web on the basis of textual keywords and to filter the results according to User Kansei Models (KUM) built on relations between impression words and image low-level features. K-DIME has been used as a co-operative electronic kansei postcard editor.

K-DIME (see fig. 2) relies on three essential components (a) a Dispatching component, (b) a Filtering Component and (c) a Feedback Processing Component. K-DIME acts as a meta-search engine for the Web and let users retrieve images on the basis of subjective and objective keywords. Users describe the images to be retrieved using Web forms. The objective parameters in users' requests (i.e. images of "Maui") are used to query existing search engines/databases, such as Alta-Vista, Lycos or Yahoo [3,4,5] and local Image Databases. The Dispatching component selects and tunes web search engines and analyses and integrates the results of the objective-query to prepare the data for the image filtering process. The Filtering component selects and configures Kansei User Models to assess them against the subjective criteria specified by the user (i.e. "fresh" images of "Maui").

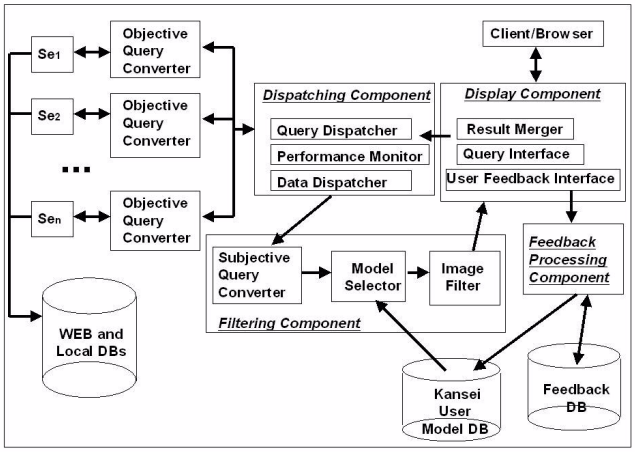


Fig. 2: Architecture of K-DIME: a meta-search engine for image retrieval and kansei filtering.

A Kansei User Model (KUM) is a computational model of the mapping function between impression words and images. The core idea behind our definition of KUM is that the "meaning" of impression words used by the users are grounded into the low-level characteristics of the images (or whichever media) at the origin of that impression. Hence a KUM is a set of mapping functions between images and impression words. In order to associate an impression word to an image, the KUM first computes a signature of the image that characterises it in term of its low-level features (colour, text, shape), than it computes the mapping between the signature and the impression word. For more details on the image processing, the reader should refer to [21].

The Feedback Processing component aims at storing feedback obtained from the user in the user's feedback database and used them to adapt the exiting KUM or to create a new one through the learning process. After visualising the resulting images, the user can enter relevance feedback, e.g. discarding images that do not fit his/her understanding of the queried subjective keyword. Reclassified images or discarded images are added to the training set of the corresponding word with the value (positive or negative) given by the user and the KUM is re-learnt (fig 3a). New impression words (i.e. words not yet modelled by any KUM) can be also used to judge the retrieved images. These judged images will be stored and used as training set for the new words.

4. NEW HYPOTHESIS GENERATION

A simple retraining of the KUM is generally not sufficient to really improve the precision of the user model. Typical problems that reduce the quality of the model are inconsistencies in the training set, limitations of the image processing and expansion of the training set. Hence, the learning process must be significantly modified, i.e. a new modelling hypothesis has to be created. The problem we have to face is how to modify the computational structure of a KUM to improve its performance, i.e.:

- How to define the correct image processing algorithm for each word, that is how to select and than combine the features of an image which will be relevant to a given impression word?
- How to detect real inconsistencies from nuances?
- How to limit the cardinality of the training set while maintaining meaningful and comprehensive samples.

We have defined and implemented an interactive computational environment (K-MMminer) endowed with multimedia mining tools, visualisation tools and externalisation tools.

The main idea behind this environment is dictated by the crucial role of the user in the mining process. One of the main functions of the environment is to trigger an externalisation process in the user in order to bring him/her to be conscious of his/her internal process (when possible). It is implemented here in the form of multimedia dialogs between the user and a Kansei Agent. A K-Agent is a computational agent in charge of creating KUM, i.e. modelling the user subjectivity. On the one hand, the K-Agent performs a complex analysis aimed first of all at detecting incosistencies, recurrent patterns or unusual patterns in the large set of user's feedback. On the other hand, the K-Agent uses such information to help the user in monitoring its activity and solve the inconsistencies.

To summarise, we adopt a dynamic/hybrid approach to continuously create and adapt KUM to users and to new users. We use a top-down approach (fig. 3a) to define the starting hypothesis. It is implemented through a bootstrapping mechanism based on user profile similarity. From the starting hypothesis, a first KUM is created for a new user. Using KDIME the KUM can be evaluated on the basis of user's feedback and automatically adapted. The result of the evaluation with the user's feedback can

trigger a bottom up approach (fig. 3b) in order to generate new hypothesis. The new hypothesis is derived by mining function and user externalisation process to direct the discovery of relevant patterns. In the next part of the paper, we describe those functions and the externalisation process.

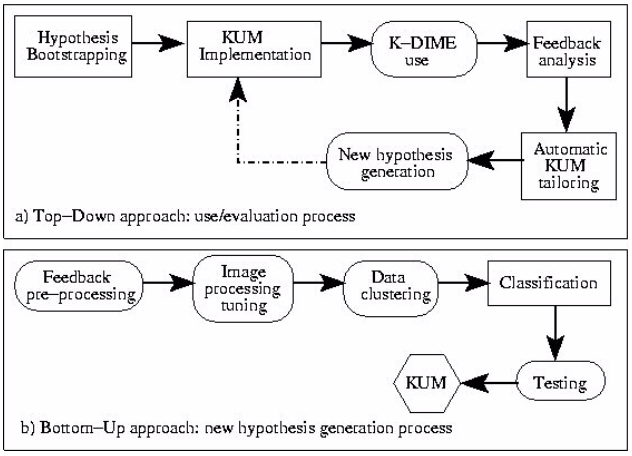


Fig. 3 Hybrid approach for kansei modeling. Oval boxes denotes interactive process.

4.1 MULTIDIMENSIONAL DATA MODELING

In order to analyse the user feedback and to generate new hypotheses, a multimedia database is created. The data stored are: the user profile, the user feedback and the user classified images. A brief description of each follows.

The user profile describes characteristics of the user that can affect his/her subjective experience. It contains user identifier, name, gender, age, nationality, hobbies, studies, job, language, favourite search engines, and reason for image retrieving activity. The user profile also contains the list of impression words that have been modelled by the user so far.

Each user feedback consists of a user identifier and a feedback. Feedback can be of different types:

- Relevance feedback is composed of an image identifier, the query for which the image has been retrieved, the agreement or disagreement of the user, the KUM that was used for the filtering and signature computed to filter the image;
- Classification feedback is composed of an image identifier and the impression words used by the user to classify the image;
- Annotation feedback is composed of an image, a region in the image, a set of low/high-level feature dimensions (e.g. dark, brigth, red, etc.) and an impression word.
- Hypothesis feedback consists of an hypothesis relating low/high level features to impression words, or relating impression words or objective words and impression words.

The thumbnail of an image addressed in a user feedback is also stored along with the following information: an image identifier, the search engine and the set of objective keywords used to fetched it, its URL and a multidimensional signature.

4.2 MULTIMEDIA DATA PRE-PROCESSING

Data in the feedback database are pre-processed before being mined. This pre-processing follows the flow depicted in fig. 4. It aims to detect inconsistencies. Inconsistencies are very common when related to subjective information such as visual impression, because of the richness of the data and the richness of personal factors that can affect visual perceptions. In such situations, it is not rare that a user classifies a same image with opposite kansei words over time. Another reason of inconsistency can be a limited/incorrect low-level characterisation of the image (its signature), i.e. an incorrect selection of features to be associated with a kansei word.

Because each image in the training set contains a large amount of information, the mining activity is done at different abstraction levels. At the lower level, a local analysis of each dimension is performed in order to remove invariant dimensions and conflicting information. Invariant dimensions can be discarded from further data analysis activity and used instead to generate association rules. For example, let's consider the dimension "cyan colour" in the modelling of the impression word "romantic". In the signature of an image taken as a positive example for the word "romantic", cyan dimension assumes generally a negative value (i.e. almost absence of this colour). It can therefore be translated into an association rule such as:

(Img, S) and (S, cyan>thr) → "not romantic"
with confidence Co = (C-PTS-romantic)/(Tot PTS-romantic).
Where:
- S is the signature of an image Img,
- C-PTS-romantic is the number of images in the positive training set for the word "romantic" that present a signature with the dimension cyan greater than threshold thr
- Tot-PTS-romantic is the cardinality of the positive training set for the word "romantic".

In such way, the number of dimensions can be significantly reduced and rules can be defined in order to simplify the learning process and the retrieving process. In this last case for example, an image showing a high quantity of cyan colour can be soon discarded in a query for "romantic" images with Co confidence. Fig. 5 shows a quantitative representation of the signature of each image of the positive training set for “romantic”. The rolling up

operation reduces of 30% the dimensionality of the colour category.

A more global analysis of the training set is performed to detect inconsistencies. There is conflicting information when a set of images presents high similarity, where similarity is computed after the multidimensional signature has been pruned, but with opposite judgement. Once those images have been detected, an interactive session with the user takes place to solve the inconsistencies. A more detailed description is given in section 5.

Ideally, the result of the data pre-processing is a new training set with a possibly reduced number of dimensions and no inconsistencies.

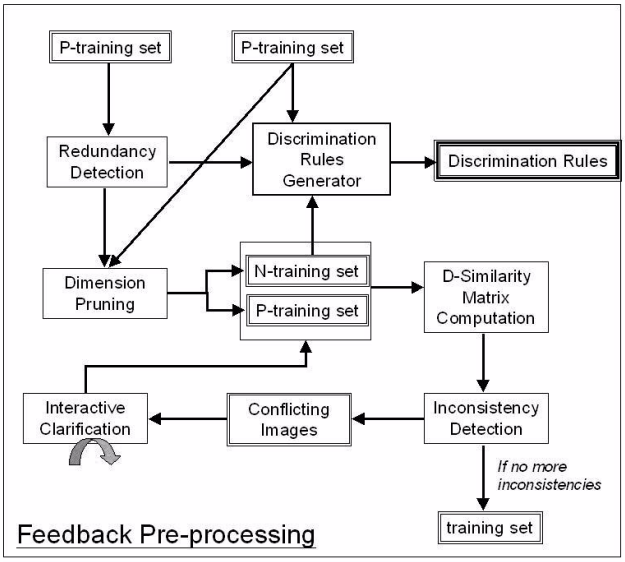


Fig. 4 Data Pre-processing flow. Circular arrows denotes interactive process.

4.3 IMAGE PROCESSING TUNING

Even though humans share a same visual apparatus, emotive responses over visual perceptions vary from an individual to the other and contextually because perceptions are affected by other cognitive processes. These mechanisms are known as attention and selection mechanisms. From our perspective, it implies that: (a) any low-level feature will not systematically be relevant to all impression words, (b) its relevance will vary from word to word and even within a word, it could vary over time and (c) the pre-processing of the visual features and their labelling will change from word to word. Hence, the image processing algorithm embedded into the KUM has to be adapted to the word. A library of image processing tools [21, 22, 23, 24] has been integrated in the environment to create multidimensional signatures of the images.

Each image is described in terms of colour characteristics, shape, texture, direction, etc. The image processing tuning consists in selecting the most relevant dimensions (e.g. black, red, dark, bright are dimensions for the category colour), and tune the algorithms to compute them according to each single impression word. In [21] we present in details the image processing used in the creation of the model for a new impression word (bootstrapping mechanism).

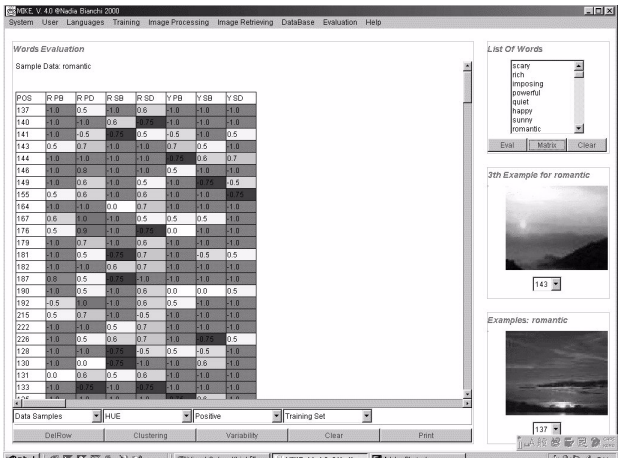


Fig. 5: This screenshot shows the listing of the positive training set for the word "romantic". For each image, a row of the table shows a subset of its signature. The dimensions which are negative over 80% of the training set have been removed from the signature of each image.

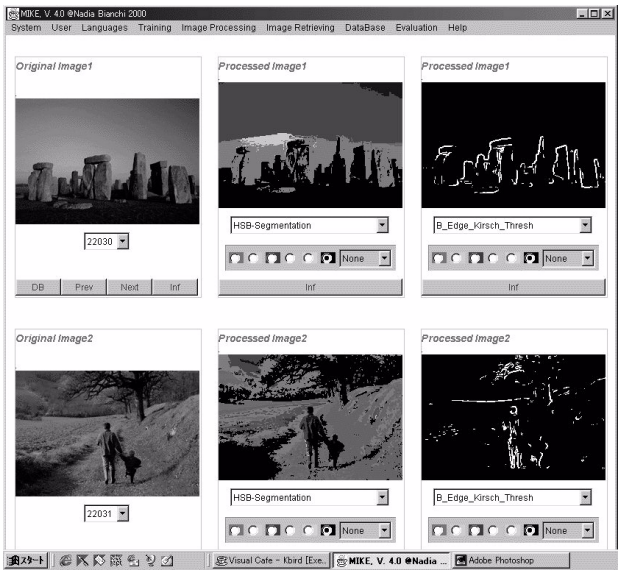


Fig. 6: This visualization tool allows the users to have a comparative view on the similarity criteria on images belonging to the training set for a chosen word, e.g. "sad". The user can validate the choice of relevant low-level features.

The similarity criteria, and hence the importance of low level features within a word can be evaluated by the K-Agent applying the similarity criteria for the different features on the training set and identify the most relevant. Figure 6 allows the user to visually evaluate the result on the selection and weight of relevant low-level features. The two images on the left of the screen are positive examples belonging to the training set for the word "sad". According to the similarity criterion on colour (green, blue, etc.) and on the detected edges, the two images result very distant, while on tonality values they shows high similarity. The user can easily confirm that the focus of attention for the "sad" impression should be shifted to the darkness of the image rather than the exact colour or edge distribution. The K-Agent can then verify the hypothesis validity and extent on other images of the training set.

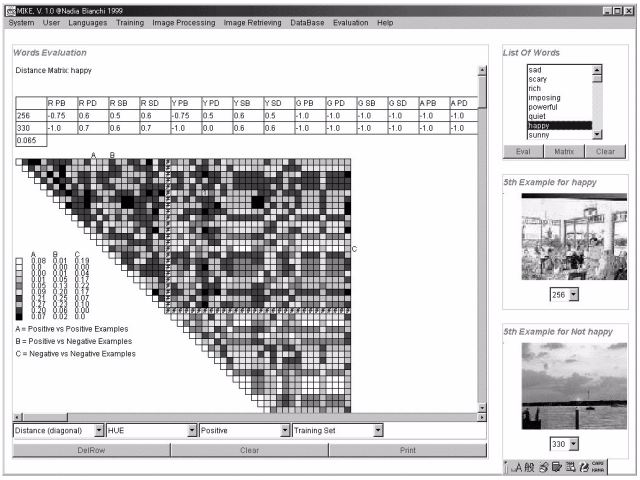


Fig 7: Quantitative and qualitative evaluation of the partition of the training set for the impression word "romantic". See details in the text.

4.4 CLUSTERING FUNCTIONS

In order to model the different aspects of a same impression word, such as different nuances, clustering techniques are applied. The cluster can be performed at two different levels: the category of images (e.g. landscape vs. cityscape) and the meaning of the kansei word (e.g. cold as emptiness or cold as a temperature). The clustering on image category creates a specialisation of the word directed to the field of application, while the clustering on low level features aims to detect different uses of the same word. When the K-Agent creates a partition of the training set of an impression word, the user is invited to associate an appropriate kansei word to each. A hierarchical network is then created among words. A same word can belong to one or more hierarchical network to reflect the fuzziness of our language.

Figure 7 shows an example of the interface through which the user can browse and evaluate the clusters proposed by the K-Agent on the training set for the word "romantic". The first table indicates the three identified clusters. The first column indicates the identifiers of images used as centroid of each cluster. Each row indicates the identifier of the images belonging to the cluster. The clusters have been computed with the k-means method [15]. In the two image windows on the right, the user can browse the images in the clusters. The diagonal matrix denotes the distance matrix computed on the element of the training set. Each cell of the matrix indicates, by darkness of its colour, the distance between two images in the training set. The set A shows the distance between the positive examples while the set B shows the distance between positive and negative examples. By clicking on a cell, the corresponding two images are visualised on the right-hand windows and their signatures are displayed in the table above the matrix for a quantitative comparative evaluation. By using this tool, the user can better understand the clusters created and eventually assign an impression word to each of them.

The availability of systems capable to classify images through their high-level content [25] supports our idea to specialise words by a category of images. The creation of independent models for "romantic city" and "romantic landscape" can reduce the complexity of the modelling process for the word "romantic".

4.5 CLASSIFICATION

The classification function learns the mapping function between low-level features and impression words (label) [21]. For each impression word the classification function is implemented by a set of neural networks trained by back-propagation with momentum on subpart of the signature of the images. The input nodes of each neural network correspond to low-level features computed by the image processing. Thus the input layer of each network and the number of network is determined mainly at the image processing tuning. Each network is associated with a weight that measures its relevance for the impression words. The output of the classification for an impression word is computed by a linear combination of the output of each neural network weighted with the respective relevance weights.

5. USER INTEGRATION TROUGH EXTERNALIZATION

Because of the complexity and of the intrinsic variability of the mapping function, the integration of the user in the mining process is fundamental. By integration we mean that the user has to acquire an active role not just limited to enter relevant feedback but extended to direct the knowledge discovery process. We suggest that the active role of the user takes the form of processes of *externalisation*. In daily life, such process can take various forms, from conversation, written texts, sketch and memos, to simply physical "records" of actions taken

in the world [26]. Figure 8 shows an example of dialog between the user and the K-Agent that aims to capture the model of "sad" impression. The dialog exploits visual examples, annotation and natural language.

One could argue that it is almost impossible for the user to determine the real factors/dimensions that give rise to an impression. We agree with the difficulties of this process and with the uncertainty of its results. However, while this process is difficult, it has been shown [27] through empirical results that subjects can be trained to become sensitive in perceiving dimensions. Secondly, and very important, the position we take in this work is to consider the externalisation process as a support to the mining activity. The contribution from the user should reduce the complexity of the data to be mined.

> *It is known from naive theory hypothesis and the mental models theory that humans, through generalisation and abstraction process, construct complex and high-level explanations of phenomena. These explanations are called "naive" theories because: (1) they are a set of not organised ontological assumptions, specific casual principles and general rules of inference; (2) they are expressed in user's common language, without making any attempt towards formal languages; (3) they are based on tacit and explicit knowledge. Ontological assumptions and casual links are not objectively established, but subjectively inferred; casual links emerge from the specific experience in the domain and are not conceived as a general causality principles. People derive mental models that guide their performances in the real worlds from their naïve theories. [28].*

Hence, we need a bridge to translate mental models into formal representations that allow implementing them in the computer and to objectively reason on these subjective beliefs. We propose a conceptual space [27] as a space where the user can externalise his/her internal mental models through a geometrical and topological representation that will translate the user's externalisation into a formal hypothesis.

5.1 A CONCEPTUAL SPACE FOR EXTERNALIZATION

The conceptual space is based on concepts (in this case, visual impression concepts). A concept is defined by a set of dimensions from different domains or categories. The description of a concept is not unique but context dependent. A domain (e.g. colour) has a geometrical and topological structure that allows the definition of relations, such as similarity, among its dimensions. Similarity is an important mechanism in cognitive processes. Generally, when making a subjective/objective experience we are brought to relate it to previous experiences and in particular to say how similar or different they are. The domains correspond to the perceptual dimensions by which the system and the user perceive the image (signature).

The set of concepts is not defined a priory. Low level dimensions (listed in table 1) are defined according to image processing algorithms and tuned upon user feedback. High level dimensions are defined as a combination of lower level dimensions. This hierarchical structure is built through mining activity and user's externalisation process. For example, the meaning of the low level categories "colour" and "tone" can be associated to the Munsell structure [29] with their dimensions formally specified as a segmentation of the Hue-Saturation-Brightness space. The high-level category "Sad" can be described in terms of "dark" or "greyish" dimensions defined respectively in the "tone" and in the "intensity" categories.

The geometrical structure of each domain supports the user understanding of the modelling process and mining activity and allows the user to express hypotheses in a formal way so that the system can interpret, implement and test them.

CATEGORY	DIMENSION-LABELS
Colour	Red, Orange, Yellow, Green, Emerald, Cyan, Sky, Blue, Violet, Purple
Tone	Dark, Dull, Strong, Pale, Vivid, Bright
Intensity	White, LightGrey, DarkGrey, Black
Morphology	Circular, Oval, Rectangular, Irregular
Position	Up, Down, Right, Left, Center, Far, Close Overlapped
Area	Small, Medium, Large
Length	Short, medium, long
Direction	Horizontal, Diagonal, Vertical,
Texture	Homogeneous, Contrasting, Fading
Importance	Background, foreground

Table 1: Low-level categories and dimension labels.

5.2 TRIGGERING PROCESSES OF EXTERNALISATION

Externalisation processes are not generally self triggered but a motivation should bring the user to engage him/herself in this difficult process. The role of the system is to trigger the user externalisation process. This is implemented here through the creation of objective representation of the externalisation steps combined with the evaluation of the user's hypotheses.

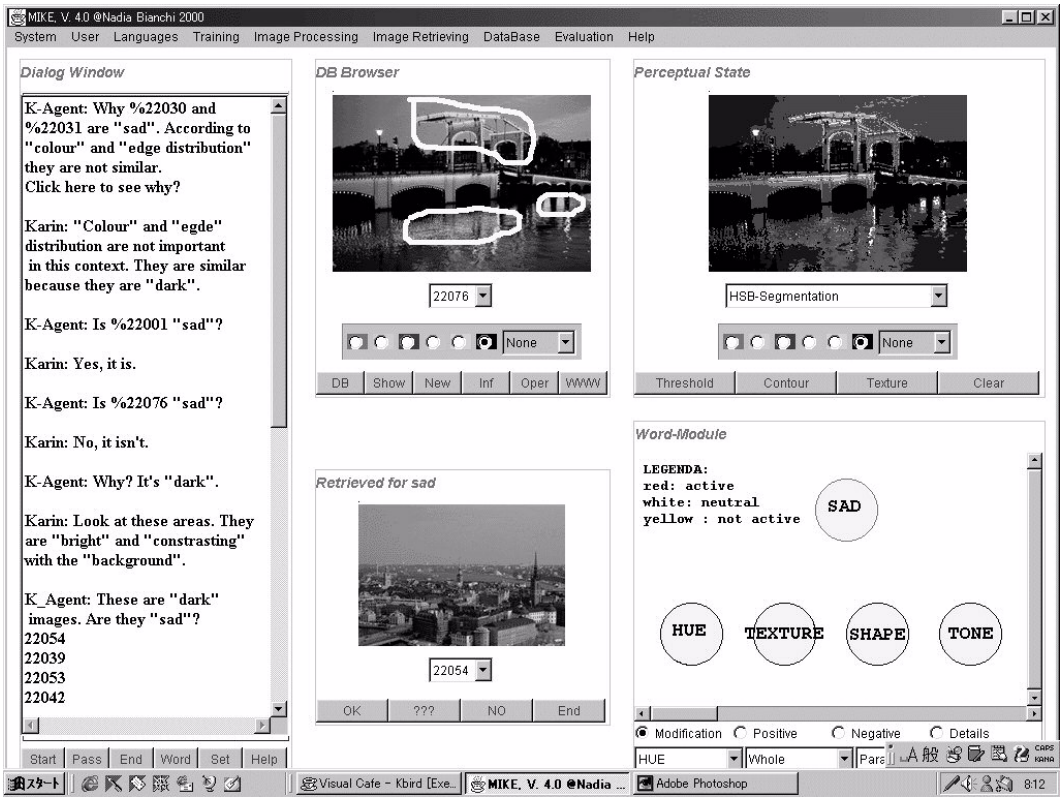


Fig 8: Dialog between a user and the K-Agent on the word "sad". The K-Agent is testing the user's hypothesis of "sad" impression associated with "dark" dimension

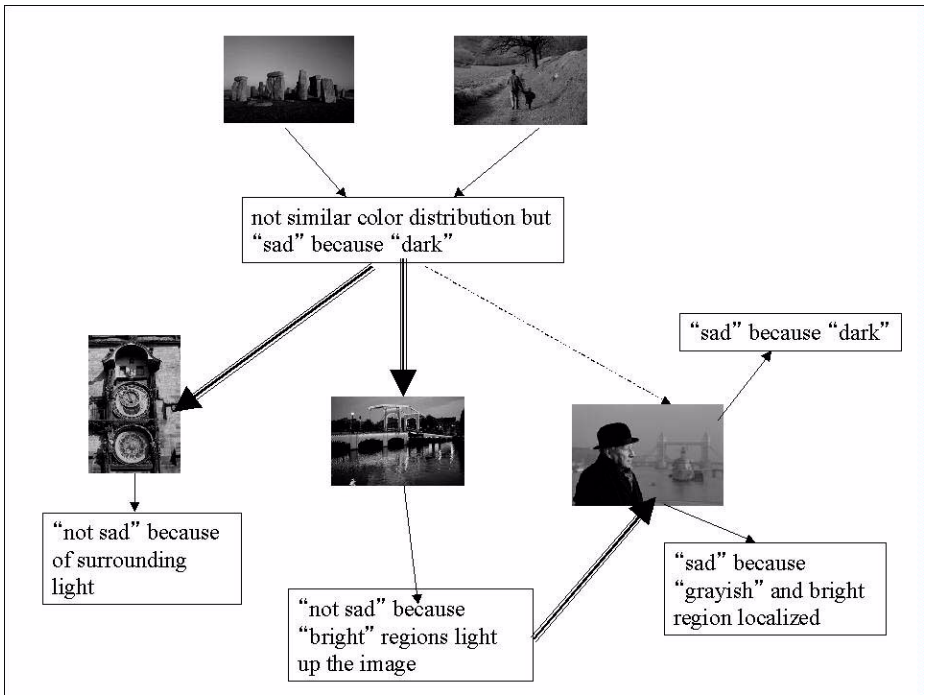


Fig 9 An example of cognitive map for the externalization process of figure 6 and figuer 8. Dotted arrows indicate consistency with the user hypotheses, double arrows indicate inconsistencies and single arrows indicate hypotheses confirmed or proposed by the user.

These representations are called multimedia cognitive maps [30]. Cognitive maps provide a simple and intuitive means of highlighting important strands of thought.

> *... Here, 'cognition' is used to refer to the mental models, or belief systems, that people use to interpret, frame, simplify, and make sense of otherwise complex problems. These mental models are referred to, variously, as cognitive maps, script, schema, and frames of reference. They are built from past experiences and comprise internally represented concepts and relationships among concepts that an individual can then use to interpret new events. This is important because decision-makers have a limited capacity for processing information so that, when dealing with complex problems like innovation, they could rarely process all the information that would be relevant. Their mental models help decision-makers to select information and to decide what actions are appropriate ...[31].*

Cognitive mapping techniques aim to provide a tool for revealing peoples' subjective beliefs in a meaningful way so that they can be examined not only by the individual for whom the map is constructed, but also by other individuals and groups [32]. The resultant cognitive map will not represent an entire belief system (this would be impossible) but hopes to portray those beliefs that are held to be most significant. The visual representation of these subjective beliefs creates an objective ground for reasoning upon them. Figure 9 shows an example of cognitive map for the externalisation process of figure 6 and figure 8. The multimedia cognitive map contains feedback from the user, hypothesis formulated by the user and inconsistencies identified by the K-Agent. Using multimedia cognitive maps the user can recollect easily the path of his/her externalisation process and correct or complete his/her hypothesis. The creation-evaluation of multimedia cognitive maps corresponds to a cyclic loop aimed at identifying inconsistencies that help the user to correct/complete his/her mental model. We consider here three different types of inconsistencies that can appear in the way an observer expresses subjective experiences. We define as:

(a) intrinsic, the inconsistencies derived from a different evaluation on the same selected information (temporal evolution in the observer).
(b) real, the inconsistencies derived from a misuse of a word in conveying a subjective experience;
(c) attentional, the inconsistencies derived from different selection mechanisms.

The detection of intrinsic inconsistencies will indicate that the training set is not (anymore) a correct and extensive sample for the associated impression word. Their detection can help the user to identify more meaningful examples for the learning process. Instead, real inconsistencies originate with a misuse of a word. Their detection can lead the user to discard them from the training set or to identify a more appropriate impression word or nuance for the classification of the images. Finally, the detection of attentional inconsistencies signals a limited or incorrect low-level characterisation of the image (its multi-dimensional signature), i.e. an incorrect selection of features to be associated with an impression word.

6. CONCLUSION

In this paper, we have presented a methodology to address the issue of having computers able to communicate over subjective visual impressions. We propose an interactive mining approach to addressed the issue of reducing the complexity and uncertainty of the mapping between patterns in images and impression words, i.e. two expression modalities for a same visual impression. New hypotheses generated by the co-operation between the user and the system are evaluated using a software system for electronic postcard edition. Images are retrieved on the basis of visual impressions the user wants to convey with the postcard.

A conceptual space is proposed as a formal space where user and system can derive formal hypotheses from the externalisation process of the user. The conceptual space is based on a set of predefined categories defined on low-level characteristics of images and a set of evolving categories (impression words being modelled). The dimensions in the low-level categories are geometrically and spatially specified, the higher level categories are built as a combination of the lower one. Three main mechanisms allow the user to construct his/her mental process and are identified as: inconsistency detection, similarity measurement and word specialisation. In order to support the user in this difficult activity, cognitive maps are used to visually and objectively represent the paths in the process.

Extensive experiments are now planned to understand the limits of the externalisation process, the variability of the user subjective responses and to identify interactive tools to facilitate the externalisation process. We believe that the approach proposed here can be extended to other modalities (music, body language, etc.) whenever subjective impressions should be conveyed or understood.

7. BIBLIOGRAPHY

[1] http://www.ctr.columbia.edu/webseek

[2] L. Taycher, M. La cascia, S. Scaroff, “Image digestion and Relevance Feedback in the ImageRover WWW Search Engines”, Proceedings of the Second International Conference on Visual Information Systems, pp. 85-91, San Diego, Dec. 15-17, 1997
[3] http://image.altavista.com/cgi-bin/avncgi
[4] http://ipix.yahoo.com
[5] http://multimedia.lycos.com
[6] N. Bianchi-Berthouze, C. Lisetti, "Modeling multimodal expression of users's affective subjective experience", International Journal on User Modeling and User-Adapted Interaction: Special Issue on User Modeling and Adaptation in Affective Computing, 2001, to appear
[7] S. Loman, H. Merman: "The KMP: A Tool for Dance/Movement Therapy," American Journal of Dance Therapy, vol. 18, No. 1, Spring/Summer, 1996.
[8] Kitajima, M. & Don-Han, K. (1998). Communicating kansei design concept via artifacts: A cognitive scientific approach. In Proceedings of the International Workshop on Robot and Human Communication, RoMan'98, Hakamatsu, Japan (pp. 321-326). IEEE Press.
[9] T. Shibata, T. Kato, “Kansei” Image Retrieval System for Street Landscape. Discrimination and Graphical Parameters based on correlation of Two Images”, IEEE International Conference on Systems Man and Cybernetics ’99, V. 6, pp. 247-252, Tokyo (Japan), October 1999
[10] K Yoshida, T. Kato, T. Yanoru, “A Study of Database Systems with Kansei Information”, IEEE International Conference on Systems Man and Cybernetics ’99, V. 6, pp. 253-256 , 1999
[11] Y. Isomoto, K. Yoshine, H. Yamasahi, N. Ishi, “Color, Shape and Impression keywords as Attributes of Painting for Information Retrieval”, IEEE International Conference on Systems Man and Cybernetics ’99, V. 6, pp. 257-262 , Tokyo (Japan), October 1999
[12] R. Hattori, M. Fujiyoshi, M. Iida, “An Education System on WWW for Study Color Impression of Art Paintings Applied NetCatalog”, IEEE International Conference on Systems Man and Cybernetics ’99, V.6, pp. 218-223 , Tokyo (Japan), October 1999
[13] T. Imai, K. Yamauchi, N.Ishi, “Color Coordination System on Case Based Reasoning System using Neural Networks”, IEEE International Conference on Systems Man and Cybernetics ’99, V. 6, pp. 224-229, Tokyo (Japan), October 1999
[14] R. Inder, N. Bianchi Berthouze, T. Kato "K-DIME: A Software Framework for Kansei Filtering of Internet Material" Proc. of IEEE International Conference of Systems, Man and Cybernetics, V.6, pp.358-363, Tokyo, Japan 1999
[15] J. Han, M.K, “Data Mining: concepts and techniques”, Academic Press, 2001
[16] O.R.Zaine, J.Han, Ze-Nian Li, S.H. Chee, J.Y. Chiang, “MultiMediaMiner: A system Prototype for Multimedia Data mining”, Proceedings ACM-SIGMOD Conference on Data Management, 1998.
[17] H. Pashler, Attention and Visual Perception: Analysing Divided Attention, Visual Cognition V.2, S.M Kosslyn, D.N.Osherson (eds), MIT Press , 1996, pp. 71-100
[18] T. Nakata, T. Sato and T. Morį "Expression of Emotion and Intention by Robot Body Movement", Proc. of Conference of International Autonomous Systems 5 (IAS-5), June 1998.
[19] H.Kobayashi, “The semantic network of Kansei words”, Inter. Conf. on Systems, Men and Cybernetics, Nashville (TE), 2000, pp 690-694
[20] S. Kobayashi, “Colorist: a practical handbook for personal and professional use”, Kodansha
[21] N. Bianchi-Berthouze, L. Berthouze, "Exploring Kansei in Multimedia Information", International Journal on Kansei Engineering, Volume 2, N.1, issue 0005, September 2001
[22] Merelli, P. Mussio, M. Padula, "An approach to the Definition, Description and Extraction of structures in Binary Digital Images". Computer Vision Graphics & Image Processing, vol. 31, pp. 19-49, 1985
[23] Haralick, R. M., Shanmugan, K. & Dinstein, I. (1973). Texture features for image classification. IEEE Transactions Systems, Man and Cybernetics, 3, 610-621.
[24] Kirsh, R. (1971). Computer determination of the constituent structure of biomedical images. Computers and Biomedical Research, 4(3), 315-328.
[25] A. Vailaya, M.A.T. Figueiredo, A.K.Jain, H.J.Zhang, “Image Classification for Content-Based Indexing, IEEE Transaction on Image Processing, Vo. 10, N.1, pp.117-130, 2001
[26] Miyake, N. (1997). Making internal process external for constructive collaboration. In Marsh, J., Nehaniv, C. & Gorayska, B. (Eds.), Cognitive Technology (pp. 119-123). IEEE Computer Society.
[27] P. Gardenfors, “Conceptual spaces: the geometry of thought”, MIT Press, 2000
[28] N. Bianchi, P. Bottoni, P. Mussio, G. Rezzonico, MG. Strepparava, “Participatory Interface Design: from Naïve Models to Systems”, Intern. Conf. on Human Computer Interaction, 1997, 24-29
[29] A.H. Munsell, A color Notation, Boston, 1905
[30] E. Tolman, “Cognitive maps in rats and men”, Psychological Review, 1948 N.55, pp.189-208
[31] A.Gopnik, C. Glymour, D. Sobel, “Casual mas and Bayes nets: a cognitive and computational account of theory formation”. International congress on Logic, Methodology and Phylosophy of Science, 1999
[32] C.Eden, “On the nature of cognitive maps”, Journal of Management Studies, 1992V. 29, pp. 261-265,

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

Data Mining for Typhoon Image Collection

KITAMOTO Asanobu
National Institute of Informatics
2–1–2, Hitotsubashi, Chiyoda-ku, Tokyo, 101–8430, JAPAN
email: kitamoto@nii.ac.jp

ABSTRACT
This paper introduces the application of data mining methods to the analysis and prediction of the typhoon. The testbed for this research is the typhoon image collection that we established, now archiving approximately 34,000 typhoon images created from satellite images of geostationary meteorological satellite GMS-5. We claim that this data collection is a medium-sized, well-controlled, and richly-variational scientific database that are suitable as a testbed for data mining research. The main challenges of this paper is twofold: the analysis and the prediction of the typhoon. For the analysis, we apply various methods such as principal component analysis and self-organizing map to characterize or visualize the statistical properties of typhoon cloud patterns. We then apply an instance-based learning method for analogy-based prediction using past similar patterns, and also established similarity-based image retrieval system of the typhoon image collection. However, we also point out that we should not overlook fundamental difficulty in typhoon prediction from past similar patterns due to the chaotic nature of the atmosphere.

KEY WORDS
Typhoon Analysis and Prediction, Meteorological Data Mining, Principal Component Analysis, Self-Organizing Map, Instance-based Learning, Similarity-based Image Retrieval

1. Introduction

In many countries, the typhoon, or the hurricane, is ranked at the top most significant threat among all natural disasters because a single typhoon attack may lead to significant casualties and property losses due to strong winds, heavy rainfall and so on. Therefore, the importance of typhoon research has been acknowledged for a long time, and the meteorology community has devoted considerable endeavor toward the analysis and the prediction of typhoons [1]. Although the meteorology community has established many typhoon models that can be effectively integrated into numerical weather prediction systems, there still remain many research issues to be challenged, including track prediction, intensity prediction, and cyclogenesis prediction of the typhoon.

Our challenge to these research issues is a little "non-traditional," since, in contrast to the traditional paradigm that has been pursued in the meteorology community, our approach is based on the informatics paradigm, by which we mean the utilization of many models and algorithms that have been developed in the informatics community, including such research fields as pattern recognition, computer vision, computer graphics, artificial intelligence, knowledge discovery / data mining, information retrieval, database systems. The basic idea in this paper is to apply data-intensive approaches to the analysis and prediction of the typhoon based on the large collection of satellite images that capture the cloud pattern of the typhoon.

The motivation of our paradigm stems from what is called the "Dvorak method" [2, 3]. In recent years most of the typhoon information are derived from satellite observations with the help of (manual or semi-automatic) pattern recognition of satellite observations along with analyst interpretation of empirically derived rules. The Dvorak method is specially designed for this purpose, and gained extensive popularity among hurricane analysis centers in the world. The formulation of Dvorak method, however, reminds us of the informatics paradigm as we mentioned earlier in the sense that pattern recognition and the past experience of experts are regarded as the indispensable part of the method. This is in contrast to mathematical models used in numerical weather prediction systems, where the model is deduced from partial differential equations that describes the dynamics of fluid in the gravity field. We are inspired by the Dvorak method, but our goal is not simply to mimic the current Dvorak method, but to characterize the essential of typhoon cloud patterns using data-intensive approaches, or in other words, data mining approaches based on the large collection of typhoon images. The goal is to discover new insights into the pattern and the dynamics of the typhoon through quantitative analysis of the typhoon image collection.

Past researches that could be related to our paradigm include the motion analysis of hurricane clouds from image sequences based on fluid dynamics [4, 5, 6], the interpretation of tropical cyclone patterns using dynamic link architecture and active contours [7], typhoon data mining by neural network [8], and typhoon track prediction by fuzzy modeling [9]. However it seems that those studies were not tested on the large collection of typhoon data, nor they gave significant impact on the meteorology community, which is undoubtedly an authority about typhoon research. Hence the goal of this paper is to discover new insights that even

meteorology experts find them interesting with solid evidence based on the large collection of data.

This paper is organized as follows. Firstly Section 2. describes data sources, the creation and current status of our typhoon image collection. Then the following sections discuss two main issues in this paper, namely, the analysis and prediction of the typhoon. Section 3. exemplifies a few results on the analysis of the typhoon using data mining methods such as principal component analysis, self-organizing map, and graph theoretic methods. On the other hand, Section 4. discusses the prediction of the typhoon using instance-based learning method based on k-NN similarity-based image retrieval, but we also compare the optimistic and pessimistic outlook on the analogy-based prediction of the typhoon. Finally Section 5. concludes the paper.

2. Typhoon Image Collection

2.1 Overview

In this section, we briefly describe relevant design issues in establishing our typhoon image collection, by which we mean a well-controlled typhoon image archives together with metadata related to the typhoon such as best track records and meteorological datasets. Some detail of this collection has already been introduced elsewhere [10, 11], but some of the important points regarding the typhoon image collection are summarized as follows:

Size The number of records (typhoon images) is moderate; it is approximately 34,000. However, the original satellite data scanned to create the collection amounts to about 600 gigabytes. Hence the creation of the typhoon image collection requires significant computation power.

Temporal data Since the frequency of the satellite observation is one hour, as in Table 1, the collection contains time series data that describes the evolution of typhoon cloud patterns over time.

Spatial data Needless to say, typhoon images are spatial data, and moreover these can be interpreted as volumetric data.

2.2 Data Sources

Best Track The first basic data source in this collection is the *best track*, which is a dataset officially compiled by national agencies in charge of meteorology such as Japan Meteorological Agency (JMA) and Australian Bureau of Meteorology (BOM). Best track records *metadata* related to the typhoon such as center location, central barometric pressure, and maximum wind speed for every three or six hours. Since this dataset is based on extensive study of the whole life cycle of the typhoon by a team of meteorology experts with the help of many kinds of external datasets, as far as off-line typhoon analysis is concerned, the typhoon center can be stably located on a satellite image with the help of best track records.

Satellite Images Most of the satellite images archived in our collection are taken by the Japanese meteorological satellite called GMS-5. Since this is a geostationary satellite that is located at 140 E above the equator, this satellite can observe both typhoons in the north-west Pacific ocean and tropical cyclones in the south Pacific and Indian ocean with relatively high frequency. The sensor of this satellite provides four bands, one in visible band and three in infrared bands. Because of the nature of electromagnetic wave, infrared bands can observe clouds even in nighttime, whereas visible bands cannot. We therefore archive infrared images as our main data source to have a uniform collection of images.

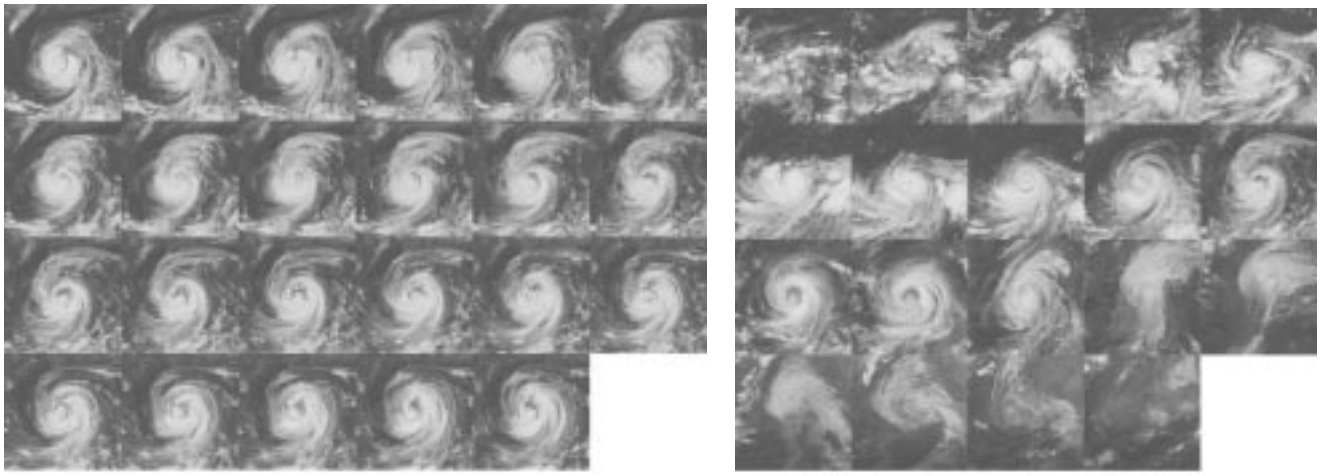
2.3 Data Preprocessing

Coordinate System The first design issue is the selection of coordinate system on which a typhoon image is created. There are two alternatives – the one is an Eulerian representation, in which the motion of typhoons is described on the fixed coordinate system, while the other one is a Lagrangian representation, in which the coordinate system moves along the motion of the typhoon. We employ the latter representation in order to focus on the time evolution of the typhoon over time, or in other words, in order to make separate the evolution of typhoon cloud patterns from the global motion of the typhoon cloud system, of which the former is our main concern. Under this coordinate system, we require accurate alignment between the image center and the typhoon center in order to create *well-framed* typhoon images.

Map Projection In addition to accurate alignment of centers, another design issue required for creating *well-framed* typhoon images is the selection of map projection. Because the typhoon is a meteorological phenomenon that occurs on the surface of a sphere, namely the earth, the careful choice of the map projection method is required not to introduce unnecessary shape distortion along with the representation of the typhoon on a two-dimensional image. The method appropriate for this purpose is azimuthal equivalent projection (Lambert azimuthal equal-area projection) [12] because of the following reasons. Firstly this projection is equal area, which means the area of clouds can be directly compared among multiple images irrespective of the geographical location of the typhoon. Secondly, shape distortion is proportional to distance from the image center, hence the effect of distortion is less harmful to circular objects such as the typhoon. The map projected well-framed image is then created so that the direction of north is always upward on the image.

Table 1. The best track and the current status of our typhoon image collection.

Basin	Northern Hemisphere	Southern Hemisphere
Best Track		
Name of agency	Japan Meteorological Agency (JMA)	Australian Bureau of Meteorology (BOM)
Latitude	$^{\circ}N$	$^{\circ}S$
Longitude	$^{\circ}$ $^{\circ}$	$^{\circ}$ $^{\circ}$
Typhoon Image Collection		
Typhoon seasons	6 Seasons (1995–2000)	5 Seasons (1995–2000)
Number of sequences	136	62
Number of images	24,800	9,400
Number of images per sequence	53 433	25 480
Observation frequency	1 hour	1 hour



(a) Hourly observation on August, 15, 1997 | (b) Daily observation for thew hole life cycle

Figure 2. The cloud patterns of Typhoon 9713 viewed in different time scales.

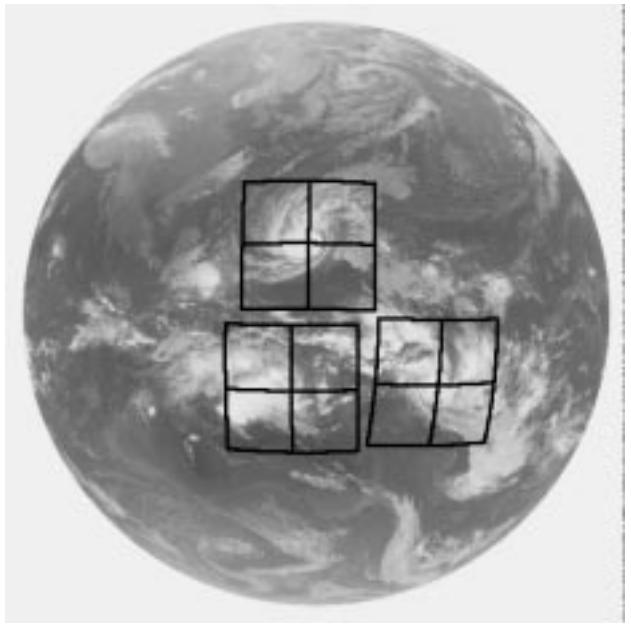
Image Classification Now a well-framed image is created as a gray-scale image, but here note that what we need is not a gray-scale pixel value itself but the classification of a pixel, or more specifically, whether *a pixel is cloudy or not.* Hence the next step is the pixel-based classification of the typhoon image utilizing three infrared bands. Because of the limited space, we omit the detail of the classification method [10, 11], but basically the method exploits the fact that such cloud parameters as air humidity and cloud top height can be estimated from the combination of pixel values of infrared bands. Then, if a pixel seems to be cloudy based on cloud parameters thus obtained, a specific cloud type such as cirrus and cumulonimbus is assigned to the pixel, or otherwise either a label "ocean" or "land" is assigned.

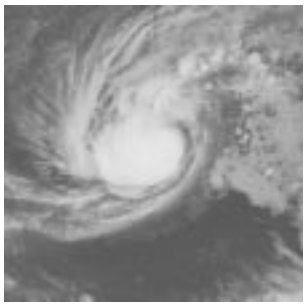
The *classified image* is further processed to create a *cloud amount image*, which represents the fraction of cloud amount within a small block of image pixels, where the cloud amount takes a real value between 0 (completely cloud-free) and 1 (completely cloud-covered). In this paper, a cloud amount image with the size of 64 64 pixels is created from a classified image with the size of 512 512 pixels, setting the size of the small block as 8 8 pixels. Thus, the final product of data preprocessing is a 64 64=4096 dimensional *cloud amount image* or *cloud amount vector*. These two terms are subsequently used interchangeably depending on the context.

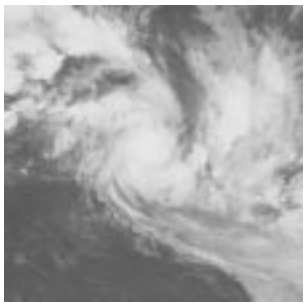
Data Cleaning Before closing this subsection, we address the issue of data cleaning. This is required when the satellite image is contaminated with burst noise or the satellite image is incomplete due to the failure of observation. The latter case is automatically removed but the former case is manually removed. However, manually removed cases are less than 0.3% among all the satellite images collected, hence the data cleaning requires little human intervention.

2.4 Current Status

Examples of typhoon images are illustrated in Figure 1, in which three typhoon images are created from the disk im-







9626(FERN)	0860 (SFERGUS)

Figure 1. The creation of *well-framed* typhoon images from the original satellite image of GMS-5. The geographical location of the typhoon center is determined from the best track, and a well-framed typhoon image is created using azimuthal equivalent map projection so that the image center always coincides with the typhoon center. The top image was taken on 1200 UTC Dec 26, 1996, in which three typhoons are simultaneously observed; that is, Typhoon 9626 (FERN) in the northern hemisphere, and Typhoon 0852 (SPHIL) in the left, and Typhoon 0860 (SFERGUS) in the right in the southern hemisphere. Two of the created typhoon images are shown below.

age of the earth originally received from the satellite. The position of the center of three typhoons are estimated from best track records published by JMA and BOM as summarized in Table 1. Then three *well-framed* typhoon images are created, which capture the core part of typhoon cloud patterns. Note that the direction of spiral bands is opposite between typhoons in the northern hemisphere and the southern hemisphere.

Table 1 also summarizes the current status of our typhoon image collection. Note that two collections, namely the northern hemisphere collection and the southern hemisphere collection represent independent typhoon sequences, because typhoons never stride over the equator due to the disappearance of Coriolis force on the equator[1]. The number of images are approximately 24,000 in the northern hemisphere collection and 10,000 in the southern hemisphere collection. We claim that these collections are medium-sized, well-controlled, and richly-variational scientific data collection which serves as an interesting testbed for data mining. For example, Figure 2 shows the cloud patterns of Typhoon 9713 viewed in different time scales. Notice the contrast between small variation in one day and significant variation in the whole life cycle. Hence these collections provide interesting challenges for pattern recognition, time series analysis, and so on. In addition, the presence of metadata such as (partly interpolated) central barometric pressure and maximum wind speed can be used as ground truth data for validation.

3. Typhoon Analysis

3.1 Exploratory Typhoon Analysis

The aim of this section is to gain some insights into typhoon cloud patterns. However, the application of various methods will be exploratory, since the effectiveness of existing data mining methods is still not evaluated for this particular target, namely the typhoon. For example, we will later apply clustering procedures, whose goal is to yield a data description in terms of clusters or groups of data points that posses strong internal similarities[13]. However, it is not intuitively clear whether the typhoon cloud patterns *really* have such underlying subclasses, despite the fact that we can at least always calculate clustering procedures. Hence, we conjecture that, to gain some insights into patterns whose characteristics are unknown, the best we can do is to apply available approaches to decide in which direction we should make advances. In the following, we introduce three data mining methods; namely variance maximization, clustering procedures, and time series analysis.

3.2 Variance Maximization – Principal Component Analysis

The first method is principal component analysis that reveals the maximum variance in the feature space, thereby gives us some insights about the variability of typhoon cloud patterns. Feature vectors to be analyzed are cloud amount vectors, which are the final product of data preprocessing. In addition, principal component analysis serves

[1] Based on the hypothesis that typhoons found in both hemispheres are fundamentally the same meteorological phenomena, there is a possibility of merging those two collections with vertically flipping southern hemisphere images, after which transformation the direction of wind circulation is the same in two collections. Although we do not further investigate this hypothesis, statistical analysis in Section 3. are related to this issue.

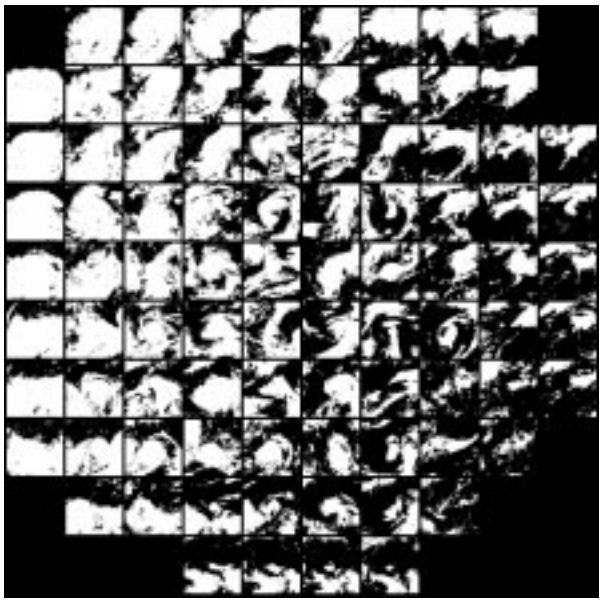


Figure 3. Principal components of typhoon cloud patterns for the northern hemisphere collection. Horizontal and vertical axis represents the first principal component and the second principal component respectively.

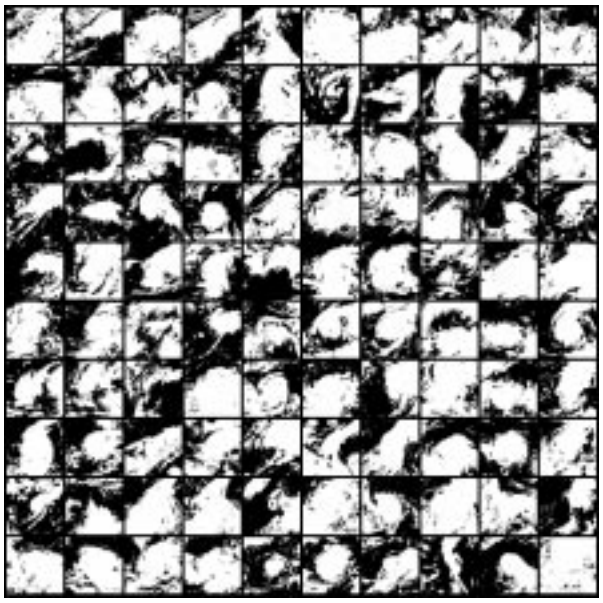


Figure 5. Clustering of typhoon cloud patterns using k-means clustering. Images shown are the ones nearest to the mean of clusters. Clusters are ordered in no particular order.

as a tool for dimensionality reduction by eliminating eigenvectors that correspond to small eigenvalues. Technically speaking, in the following experiments, each element of cloud amount vectors is normalized as mean and variance equals zero and one respectively[2], since, in general, the cloud amount of elements which are near the center of the typhoon takes values close to unity with small variance. Hence this normalization gives better representation power in terms of small-scale structure around central core clouds.

Figure 3 illustrates the distribution of typhoon cloud patterns on the space of first and second principal components. Each image that appears in Figure 3 is chosen based on proximity to regularly sampled points in the space of two principal components. This figure suggests that those two axes represent the slope of cloud amount between north and south. In other words, the maximum variance in typhoon cloud patterns is found in the north-south slope of the cloud amount. In addition, around the middle of extremes along two axes, we can observe cloud patterns with large curvature or circular shape with core cloud clusters. Thus these principal components visualize the distribution of typhoon cloud patterns by linearly projecting data vectors into a low dimensional space.

A principal component, or an eigenvector, can also be visualized as an *eigenpicture* by assigning a gray-scale value to each element of an eigenvector, as in face recognition [14] and in remote sensing image analysis [15]. Figure 4 illustrates eigenpictures for the northern and southern hemisphere. It is clear that the rotation of cloud bands is opposite in two hemispheres. It is also interesting to note that both sets of eigenpictures show similar tendency – from large-scale structures to small-scale structures, and the graphs of cumulative proportion also show similar tendency in two collections. These results suggest that typhoons in both hemispheres are in fact driven by the same atmospheric mechanism.

[2]This process is equivalent to using a correlation coefficient matrix instead of a variance-covariance matrix for calculating eigenvectors.

3.3 Clustering – Self-Organizing Map

As stated earlier, clustering procedures aims at yielding a data description in terms of clusters or groups of data points that posses strong internal similarities [13]. If, in some sense, the center of gravity of each cluster represents a "typical" pattern of the typhoon, a set of typical patterns may serve as a graphical representation of typhoon cloud patterns in a way that humans can easily understand. In fact, the Dvorak method, acknowledged as the standard method for typhoon analysis, also uses similar representation – assigning empirically derived rules for "typical" cloud patterns derived from the long experience of analysts. Hence it is expected that clustering procedures may produce the intuitive summarization of the typhoon cloud patterns that can be used as the *catalog* of collections. This issue is also related to the *visualization* of typhoon cloud patterns.

The basic (non-hierarchical) clustering procedure is the k-means clustering, whose result is shown in Figure 5. In this experiment, the number of clusters is fixed to 100, and clusters obtained through experiments are shown in Figure 5 in no particular order. In this representation, al-

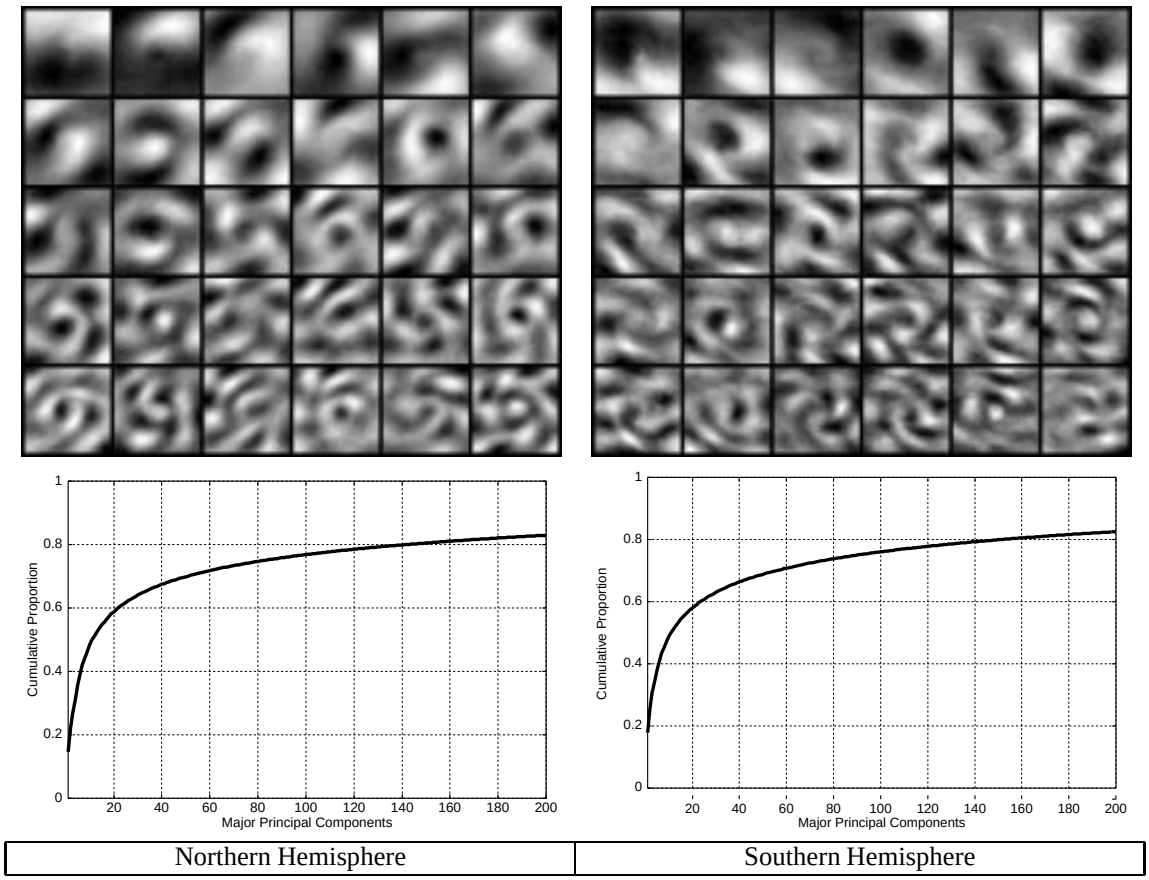


Figure 4. The eigenpictures of the typhoon in the northern and the southern hemisphere. These eigenpictures are ordered by corresponding eigenvalues from 1st (upper-left corner) to 30th (lower-right corner), and the cumulative proportion is illustrated in the graph below.

though those images can be regarded as representative images of subclasses, it is not easy to grasp the global distribution of typhoon cloud patterns because clusters are not ordered in any ways.

One method for having better representation in terms of the ordering of clusters is called the self-organizing map (SOM) [16], which summarizes high dimensional data vectors with a set of reference vectors having a spatial organization on a (usually) two-dimensional lattice. The detail of the algorithm can be found in many publications [16], so we only describe the settings we used for the basic SOM. The input vectors to the SOM are dimensionality reduced vectors derived from cloud amount vectors. That is, based on the result of principal component analysis, we transform originally 4096 dimensional cloud amount vectors into dimensionality reduced vectors computed by the linear combination of eigenvectors. The number of dimensions is determined by the cumulative proportion of eigenvalues; in the subsequent experiments, we use 83 dimensional vectors that corresponds to the cumulative proportion of 75% as illustrated in Figure 4.

The array of neurons are configured on a square lattice with the size of 10 10 neurons. We tested two types of topology; namely normal lattice topology and torus-type topology in which neurons on the left and the upper boundary are connected with ones on the right and lower boundary, respectively. Some of the other settings are described below.

1. Topological neighborhood is defined in reference to chess-board distance on a square lattice.
2. Learning rate factor is proportional to the inverse of the number of steps with some minimum limit.
3. Reference vectors are randomly initialized.

Although, in general, slightly different ordering of neurons are obtained after learning, depending on the initial condition of the SOM, we show a result in Figure 6 obtained from particular learning steps. Images shown on this map are the ones which are the nearest to reference vectors of neurons, hence they approximate the distribution of typhoon cloud patterns through the nonlinear projection of the SOM. From a visualization point of view, this representation provides a ”birds-eye-view” of typhoon cloud patterns, which clearly illustrates gradual shape transition from neuron to neuron. There appear to be clusters of large clouds, small clouds, elongated clouds, circular clouds, etc.

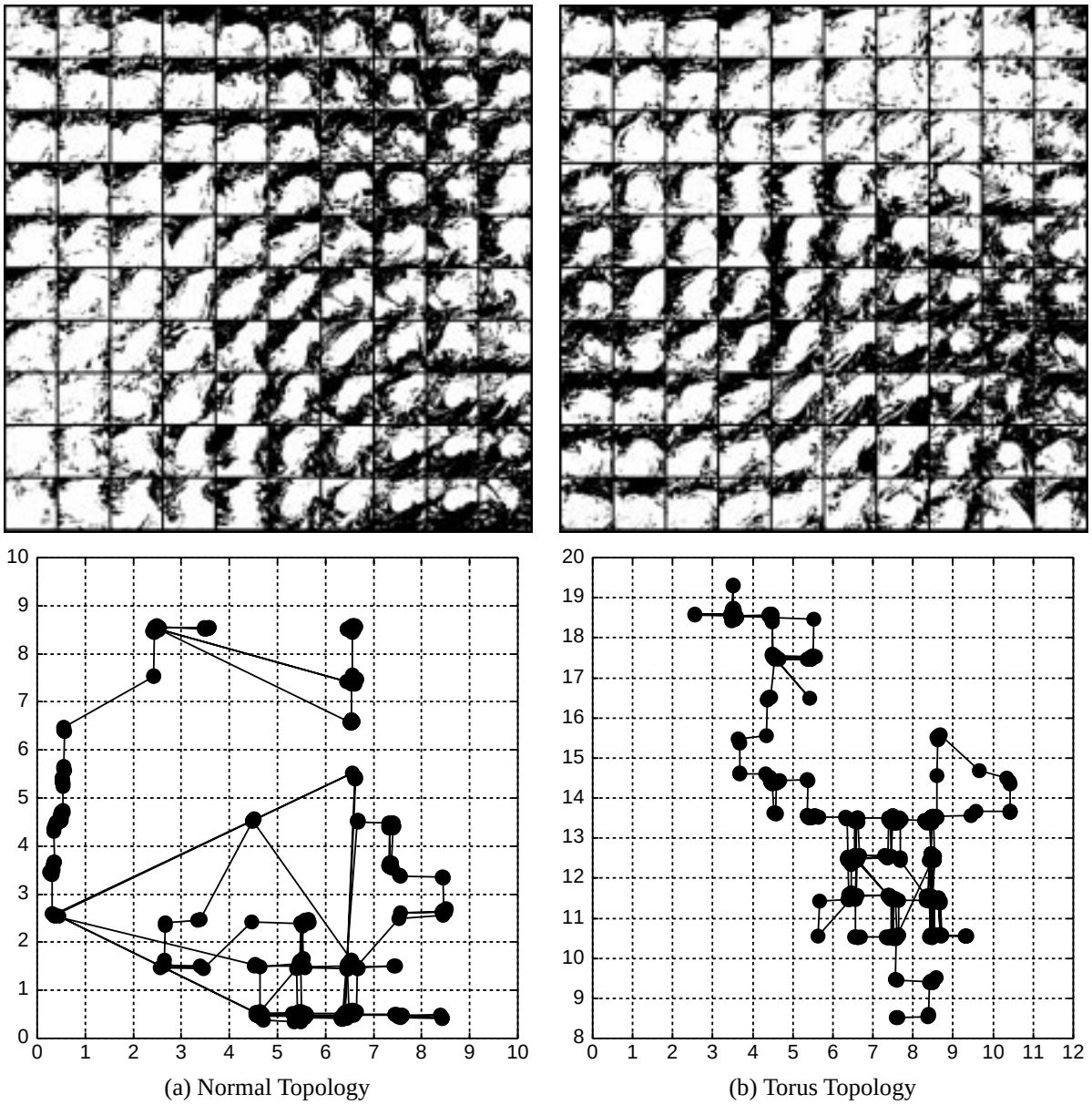


Figure 6. Clustering of typhoon cloud patterns using Self-Organizing Map, and the evolution of Typhoon 9713 on the SOM space. This typhoon is the same sequence as the one in Figure 2.

Thus the ordering of typhoon cloud patterns attained by the SOM gives a unique insight into the nature of typhoon cloud patterns.

3.4 Time Series Analysis – Graph Theoretic Analysis

Figure 6 also illustrates the evolution of typhoons on the obtained SOM space. Here hourly observation of typhoons are projected into the SOM space as a node, and sequential observations are then connected with edges, forming a (directed) graph structure as a whole. Numbers along the axis represent the column and row index of neurons, but in (b) numbers should be interpreted with mod 10 because torus topology is employed. If we regard a neuron as a *state*, and focus on the state transitions, then we can compare the properties of state transitions among typhoon sequences. Comparing both charts, (b) seems to be a more natural representation, because state transition is more ordered than (a), meaning that the ordering attained by the SOM may better corresponds to time series ordering of the typhoon cloud patterns. Assuming that the typhoon in the real world changes its state in a continuous manner, big leaps between distant states may be spurious due to the inappropriate ordering of neurons. However, some of the big leaps may not be spurious but may indicate the presence of rare phenomena. Hence future investigation into state

Figure 7. Similarity-based retrieval of typhoon imagery. The typhoon image in the upper-left corner is a query, and others are similar images to the query. Note that typhoon images in the same sequence to the query are excluded from the search.

transition rules may uncover interesting hidden knowledge in the evolutionary patterns of the typhoon. This is still an open question.

4. Typhoon Prediction

4.1 Instance-based Learning and Image Retrieval

Instance-based learning is conceptually straightforward approaches to approximating real-valued or discrete-valued target function. Learning in this algorithm consists of simply storing the presented training data. When a new query instance is encountered, a set of similar related instances is retrieved from memory and used to classify the new query instance. It is robust to noisy training data and quite effective when it is provided a sufficiently large set of training data [17]. A simple instance-based method called the k-nearest neighbor (k-NN) algorithm assumes that all instances correspond to points in the n-dimensional space, in which the nearest neighbors of an instance are defined in terms of a distance measure.

In order to perform nearest neighbor search on the typhoon image collection, we established a WWW (World Wide Web) interface to our prototype image database systems as illustrated in Figure 7 [18]. This system can perform k-NN similarity search using Euclidean distance between the vector of a query image and the vectors of archived images[3], and clicking on one of the retrieved images spawns another similarity-based retrieval of images. In addition, other search criteria such as temporal and spatial location are also provided, and the summarization of the whole life cycle of each typhoon gives an intuitive idea on the evolution of the typhoon. Extending this idea, we can imagine a typhoon prediction system based on similarity-based image retrieval system that should work as follows:

1. We first create the well-framed typhoon image of the current typhoon to be predicted, and use this image as a query.

2. Image database system performs a similarity-based retrieval and returns a list of similar typhoon images from past typhoon sequences, together with information on their subsequent evolution.

3. We then predict the evolution of the current typhoon based on the *ensemble* of the evolution of similar typhoon sequences.

In short, this scenario seems to be a natural way of predicting the typhoon using analogy; however, in fact, this scenario may be too optimistic because of the reasons we will discuss in the following.

4.2 Criticism

The optimistic outlook for the instance-based prediction sounds all very well; however, we should not blind ourselves to fundamental difficulty in predicting atmospheric events, namely the chaotic nature of the atmosphere. In fact, instance-based, nearest neighbor search for similar atmospheric situation was proposed and numerically evaluated more than 30 years ago by one meteorologist well known for his discovery of chaos in the atmosphere. In his pioneering work [19, 20], he tried to find a similar weather situations (analogues), in terms of the pressure pattern of the upper troposphere, from historical weather data in the hope of utilizing historical data for the future prediction of current weather. However, the result was disappointing. He found out that there were indeed no truly good analogues, so he claimed that in practice this procedure might be expected to fail. Another recent paper [21] also reported similar difficulty in finding truly similar patterns after they searched for more than 15 million combinations of barotoropic pattern of the atmosphere.

We suspect that these pessimistic results are caused by an effect called *the curse of dimensionality*; that is, distance between neighbors is dominated by the large number of irrelevant attributes [13, 17], hence even similar image pairs takes relatively large distance (dissimilarity). To

[3] Sometimes it is better to exclude from similarity-based retrieval typhoon images that belong to the same typhoon sequence as the query because we are more interested in retrieving similar images from *different* typhoon sequences.

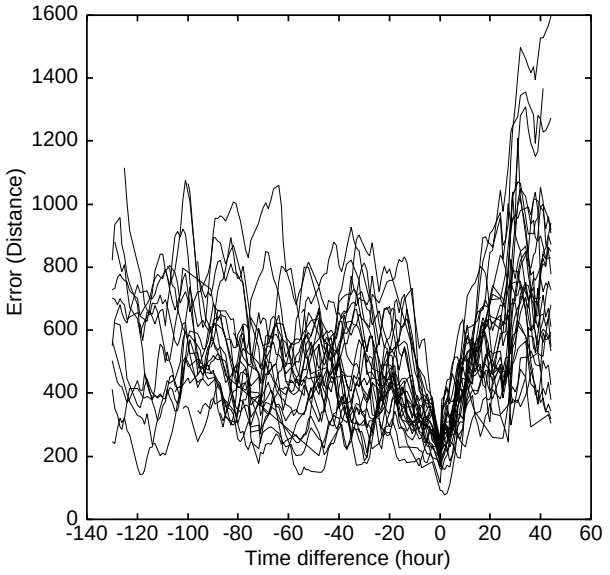


Figure 8. Predictability of the typhoon cloud patterns based on instance-based learning.

see this effect, we measured the evolution of distance[4] between similar patterns over time and illustrate the result in Figure 8. In this experiment, the most similar images to the query are found in sequence basis, and those images are aligned at time zero. Then at time t, we compare distance between the image t hours after the query image and the image t hours after the aligned similar image. The implication of Figure 8 is that even if we can find relatively similar images from other typhoon sequence, dissimilarity between typhoon sequences increases rapidly and soon similarity becomes insignificant compared to that between randomly chosen samples. In short, it is very difficult to find past similar patterns that are expected to keep similar time evolution over long period of time.

Another point related to the above argument is that what we should predict in terms of typhoons. Our final interest may be in predicting the intensity, track, and evolution of the typhoon. If we notice the fact that Dvorak method relies on the pattern recognition of cloud patterns, we can assume that the prediction of cloud patterns gives good estimate for the intensity that we want to predict. However, the cloud pattern of the typhoon has a very high degree of freedom, and direct prediction of cloud patterns may be too hard for predicting just some parameters which has only a lower degree of freedom. The curse of dimensionality indicates that the selection of good features and good distance metric to calculate similarity is important because otherwise distance does not give meaningful value. In future works, these points will be an important factor in building effective analysis and prediction algorithms of the typhoon.

[4] As before, the Euclidean distance between dimensionality reduced cloud amount vectors is used.

5. Conclusions and Open Problems

This paper described our project on informatics-based typhoon analysis and prediction. The infrastructure of this study is the collection of typhoon images that consists of approximately 34,000 typhoon images from the northern and southern hemisphere, created from the geostationary meteorological satellite GMS-5. This data collection is a medium-sized, well-controlled, and richly-variational scientific database that are suitable for a data mining testbed.

We discussed two main challenges of our study; namely the analysis and prediction of the typhoon. Firstly, we introduced various data mining methods for the analysis of the typhoon. Principal component analysis revealed that the maximum variance of the typhoon cloud pattern lies in the north-south slope of the cloud amount, and this method was also used for dimensionality reduction of cloud amount vectors. Next the clustering methods such as k-means and the self-organizing map (SOM) was applied for visualizing the distribution of typhoon cloud patterns on a two dimensional space. As a result, the SOM provided better visualization that gives intuitive notion on the entirety of our typhoon image collection. We also made a preliminary study about visualizing the evolution of typhoons on the SOM space, and found out that the torus-type topology of the SOM may be better suited for the representation of time evolution.

Another challenge in this paper is prediction, and we showed a prototype similarity-based image retrieval system for our typhoon image collection based on instance-based learning with k-NN search. In the optimistic scenario, this system should serve as an effective tool for analogy-based prediction of the typhoon. However, the pessimistic criticism fundamentally linked to the chaotic nature of the atmosphere also poses a formidable problem that should not be overlooked. Hence our mission is to develop data models and data mining methods that work effectively for the analysis and prediction of the typhoon against the curse of dimensionality.

Acknowledgments

We would like to thank Prof. M. Kitsuregawa and Dr. T. Nemoto in Institute of Industrial Science, University of Tokyo, and also Prof. R. Shibasaki in Center for Spatial Information Science, University of Tokyo, for the receiving and archiving of GMS-5 satellite data, which were used in this study.

References

[1] R.A. Pielke, Jr. and R.A. Pielke, Sr. Hurricanes : Their Nature and Imacts on Society. John Wiley & Sons, 1997.

[2] Dvorak, V.F. Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery. *Monthly Weather Review*, Vol. 103, pp. 420–430, 1975.

[3] Dvorak, V.F. Tropical Cyclone Intensity Analysis Using Satellite Data. *NOAA Technical Report NESDIS*, Vol. 11, pp. 1–47, 1984.

[4] Palaniappan, K., Kambhamettu, C., Hasler, A.F., and Goldgof, D.B. Structure and Semi-fluid Motion Analysis of Stereoscopic Satellite Images for Cloud Tracking. In *Proc. of International Conference on Computer Vision*, pp. 659–665. IEEE, 1995.

[5] Zhou, L., Kambhamettu, C., and Goldgof, D.B. Extracting Nonrigid Motion and 3D Structure of Hurricanes from Satellite Image Sequences without Correspondences. In *Proc. of Conference on Computer Vision and Pattern Recognition*. IEEE, 1999.

[6] Zhou, L., Kambhamettu, C., and Goldgof, D.B. Fluid Structure and Motion Analysis from Multi-spectrum 2D Cloud Image Sequences. In *Proc. of Conference on Computer Vision and Pattern Recognition*. IEEE, 2000.

[7] Lee, R.S.T. and Liu, J.N.K. An Automatic Satellite Interpretation of Tropical Cyclone Patterns Using Elastic Graph Dynamic Link Model. *Pattern Recognition and Artificial Intelligence*, Vol. 13, No. 8, pp. 1251–1270, 1999.

[8] Zhou, Z., Chen, S., and Chen, Z. Mining Typhoon Knowledge with Neural Networks. In *Proc. 11th Int. Conf. on Tools with Artificial Intelligence*, pp. 325–326. IEEE, 1999.

[9] Hiraoka, T., Maeda, H., and Ikoma, N. Two-stage Prediction of Typhoon Position by Fuzzy Modeling. In *Proc. of Int. Conf. on Systems, Man and Cybernetics*, pp. 581–585. IEEE, 1999.

[10] Kitamoto, A. The Development of Typhoon Image Database with Content-Based Search. In *Proceedings of the 1st International Symposium on Advanced Informatics*, pp. 163–170, 2000.

[11] Kitamoto, A. and Ono, K. The Collection of Typhoon Image Data and the Establishment of Typhoon Information Databases Under International Research Collaboration between Japan and Thailand. *NII Journal*, No. 2, pp. 15–26, 2001.

[12] Yang, Q.H., Snyder, J.P., and Tobler, W.R. Map Projection Transformation. Taylor & Francis, 2000.

[13] Duda, R.O. and Hart, P.E. Pattern Classification and Scene Analysis. John Wiley & Sons, 1973.

[14] Turk, M. and Pentland, A. Eigenfaces for Recognition. *J. of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71–86, 1991.

[15] Fayyad, U.M., Smyth, P., Weir, N., and Djorgovski, S. Automated Analysis and Exploration of Image Databases: Results, Progress, and Challenges. *Journal of Intelligent Information Systems*, Vol. 4, pp. 7–25, 1995.

[16] Kohonen, T. Self-Organizing Maps. Springer, second edition, 1997.

[17] Mitchell, T.M. Machine Learning. McGraw-Hill, 1997.

[18] http://www.digital-typhoon.org/.

[19] Lorenz, E.N. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *Journal of the Atmospheric Sciences*, Vol. 26, pp. 636–646, 1969.

[20] Lorenz, E.N. Three Approaches to Atmospheric Predictability. *Bulletin of Americal Meteorological Society*, Vol. 50, No. 5, pp. 345–349, 1969.

[21] Nohara, D. and Tanaka, H.L. Logarithmic Relation between the Initial Error and Predictability for the Barotoropic Component of the Atmosphere. *Journal of the Meteorological Society of Japan*, Vol. 79, No. 1, pp. 161–171, 2001.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

Multimedia Data Mining for Traffic Video Sequences

Shu-Ching Chen[1], Mei-Ling Shyu[2], Chengcui Zhang[1], Jeff Strickrott[1]
[1]Distributed Multimedia Information System Laboratory
School of Computer Science, Florida International University, Miami, FL 33199
[2]Department of Electrical and Computer Engineering, University of Miami,
Coral Gables, FL 33124

ABSTRACT

In this paper, a multimedia data mining framework for discovering important but previously unknown knowledge such as vehicle identification, traffic flow, and the spatio-temporal relations of the vehicles at the intersections from traffic video sequences is proposed. The proposed multimedia data mining framework analyzes the traffic video sequences by using background subtraction, image/video segmentation, object tracking, and modeling with multimedia augmented transition network (MATN) model and multimedia input strings, in the domain of traffic monitoring over an intersection. The spatio-temporal relationships of the vehicle objects in each frame are discovered and accurately captured and modeled. Such an additional level of sophistication enabled by the proposed multimedia data-mining framework in terms of spatio-temporal tracking generates a capability for automation. This capability alone can significantly influence and enhance current data processing and implementation strategies for several problems vis-à-vis traffic operations. A real-life traffic video sequence is used to illustrate the effectiveness of the proposed multimedia data mining framework.

KEY WORDS: Multimedia data mining, spatio-temporal relationships, multimedia augmented transition network (MATN), object tracking.

1. INTRODUCTION

As computers have become more powerful, their role in everyday life has become more pervasive. Recent efforts [8,23,26] have begun to shift the traditional focus from user centric applications (i.e., word processors, browsers, etc.) to that of a ubiquitous tool that facilitates everyday activities. Projects like EasyLiving [23,26] and HAL [8] aim to develop smart spaces that can monitor, predict, and assist the activities of its occupants. These efforts at developing smart environments are not confined to homes or offices, but extend to that of the world around us. Municipalities [1,24] are installing video camera systems to monitor and extract traffic control information from their highways in real time. Issues associated with extracting traffic movement and recognizing accident information from real time video sequences are discussed in [10,11,20,21,22]. Two common themes exist in these works. First, the video information must be segmented and turned into objects. Second, the behavior of those objects is monitored (they are tracked) for immediate decision making purposes. What is missing in these efforts is to model and index the data for on-line analysis, storage or later pattern mining.

The analysis and mining of traffic video sequences to discover information, such as vehicle identification, traffic flow, and the spatio-temporal relations of the vehicles at intersections, provides an economic approach for daily traffic operations. In order to identify and track the temporal and relative spatial positions of vehicle objects in video sequences, it is necessary to have object-based representation of video data. For this purpose, attention has been devoted to segmenting video frames into regions such that each region, or a group of regions, corresponds to an object that is meaningful to human viewers [9,13,14]. While most of the previous works are based on low-level global features, such as color histogram and texture, our video segmentation method focuses on obtaining object level segmentation; obtaining objects in each frame and their traces across the frames. In [3-5] we have addressed the issues of unsupervised image segmentation; object modeling with multimedia input strings to capture the spatial–temporal behavior of the object, and the application of these techniques to the domain of traffic monitoring.

Similar approaches to our segmentation technique are discussed in [12,25]. In [25] the authors consider a Bayesian technique to segment images based on feature distributions. The histogram of features around a pixel neighborhood is considered as an estimate of the conditional probability distribution P(c|Y) versus the parametric equation in our approach (see Section 2.2). This technique models the texture in a neighborhood. DeMenthon et al. [12] utilize a Hidden Markov Model approach for low level image segmentation. Associated with each pixel are an observation vector and a hidden state. The observation vector is the set of parameters (of interest) associated with each pixel, such as color, or the average intensity of the image region centered on that

pixel. The hidden state is a label for that pixel. Computational time is $O(ns^3)$, where n is the number of pixels in the image and s is the number of states (regions) to segment the image. Segmentation in an image can also be modeled as a pixel-labeling problem, in which we must decide from which of M number of classes the pixel belongs. The membership in each class is formulated as a Bayesian conditional probability decision, where class membership is estimated from the intensity distributions of neighboring pixels. When the image segmentation problem is considered for a fixed camera domain, a classic technique to resolve the foreground objects is background subtraction [16]. This involves the creation of a background model that is subtracted from the input image to create a difference image. The new difference image only contains objects not in the background or new features that have not yet been incorporated into the background.

Various approaches to background subtraction and modeling techniques have been discussed in the literature [11,17,19,28], ranging from modeling the intensity variations of a pixel via a mixture of Gaussian distributions to simple differencing of successive images. In [29] the authors provide some simple guidelines and evaluation of the various techniques for background modeling. We are in the beginning phases of evaluating the performance benefits of background subtraction methods for the various domains of our image segmentation applications. To that aim we have evaluated the effectiveness of simple image averaging techniques over stationary (non-changing) portions of the image data set.

In this paper, a multimedia data mining framework for traffic video sequences is proposed. The proposed framework considers image/video segmentation with initial background subtraction, object tracking, and modeling with multimedia augmented transition network (MATN) model and multimedia input strings [2,7], in the domain of traffic monitoring over an intersection. The multimedia input strings are used to capture the spatio-temporal relationships of vehicle objects thereafter. The video segmentation method mentioned here is unsupervised. Another advantage is that it uses the segmentation result of the previous video frame to speed up the segmentation process of the current video frame. Experiments were conducted to illustrate the effectiveness of the proposed framework using a real-life traffic video sequence. The traffic video sequence was captured with a *Sony Handycam CCD TR64* and digitized with an inexpensive *Brooktree Bt848* based capture card on a Windows NT 2000 Celeron based platform. The original images are 640x480, 24 bit color and the video sequence was sampled at 5 frames per second.

The organization of this paper is as follows. In next section, the knowledge discovery process that includes background subtraction, the unsupervised segmentation algorithm, object tracking techniques, MATN model, and multimedia input strings are introduced. Experiment results and analysis of the proposed multimedia data mining framework are discussed in Section 3. Along with the discussion, an example real-life traffic video sequence is used. Conclusions are presented in Section 4.

2. MINING INFORMATION FROM TRAFFIC VIDEO SEQUENCES

Traffic video analysis can discover and provide useful information, such as queue detection, vehicle classification, traffic flow, and incident detection at the intersections. To the best of our knowledge, the current transportation applications and research work either do not connect to databases or have limited capabilities to index and store the collected data (such as traffic videos) in their databases. Therefore, those applications cannot provide organized, unsupervised, conveniently accessible and easy-to-use multimedia information to traffic planners. In order to discover and provide some important but previously unknown knowledge from the traffic video sequences to the traffic planners, multimedia data mining techniques need to be employed. The proposed multimedia data-mining framework includes background subtraction, vehicle object identification and tracking, multimedia augmented transition network (MATN) model and multimedia input strings. The additional level of sophistication enabled by the proposed framework, in terms of spatio-temporal tracking, generates a capability for automation. This capability alone can significantly influence and enhance current data processing and implementation strategies for several problems vis-à-vis traffic operations.

MATNs and multimedia input strings are used to model the temporal and relative spatial relations of the vehicle objects. An unsupervised video segmentation method, i.e., the SPCPE algorithm (see Section 2.2), can identify vehicle objects. In our framework, we introduce the technique of background subtraction to enhance the basic SPCPE algorithm to get better segmentation results, so that the more accurate spatio-temporal relationships of objects can be obtained. In the following subsections, we will first introduce the background subtraction technique, then give an overview of the SPCPE algorithm and the object tracking techniques, after that we will briefly describe how to use MATNs and multimedia input strings to model key video frames. A portion of the traffic video clips are used to demonstrate how video indexing is modeled by the MATNs and multimedia input strings.

2.1 Background Subtraction

Background subtraction is a technique to remove non-moving components from a video sequence. The main

assumption for its application is that the camera remains stationary. The basic principle is to create a reference frame of the stationary components in the image. Once created, the reference frame is subtracted from any subsequent images. Those pixels resulting from new (moving) objects will generate a difference not equal to zero (i.e., difference ≠ 0).

In this work, those video sequences containing non-moving objects were manually selected from the video data and then averaged together. The image sequence used consists of about 16 minutes of video from a traffic intersection with approximately constant lighting conditions (the sun was out and in Miami that means a bright day). Our approach is similar to that of [18] or [21], where a reference frame is constructed by accumulating and averaging images of the target area (the intersection in our case) for some time interval. As mentioned above, this is not a robust technique as it is sensitive to intensity variations [19]. That is, it can generate false positives since the detection of moving objects solely due to lighting changes. It can also generate false negatives due to the addition of stationary objects to the scene that are not part of the reference frame. [29] provides a good summary of the problems associated with background modeling. We use a simple averaging technique for this work as it allows us to quickly evaluate an upper limit on the performance improvement with our unsupervised segmentation algorithm.

(a) Original frame 32

(b) Background reference frame

(c) Difference image

(d) Binary image

Figure 1: Example result of background subtraction

The difference image (as shown in Figure 1(c)) is created by subtracting the reference frame (as shown in Figure 1(b)) from the current image (as shown in Figure 1(a)). The results are scaled by $s = c\log(1+|d_{ij}|)$, where d_{ij} is the value for the difference at pixel ij. The scaling results in nonlinearly boosting the differences away from zero and towards 255 (the value of c will determine where saturation will occur). The results of the differencing step are fed to our unsupervised segmentation algorithm as the input images. Binary thresholding of the difference image can be used as an initial partition to improve the speed of converging (see Section 2.2) in our segmentation algorithm. Figure 1 gives an example result of background subtraction for frame 32.

2.2 Unsupervised Video Segmentation Method (SPCPE)

The SPCPE (Simultaneous Partition and Class Parameter Estimation) algorithm is an unsupervised video segmentation method to partition video frames. A given class description determines a partition. Similarly, a given partition gives rise to a class description, so the partition and the class parameter have to be estimated simultaneously. In practice, the class descriptions and their parameters are not readily available. An additional difficulty arises when images have to be partitioned automatically without the intervention of the user. Thus, we do not know a priori which pixels belong to which class. In the SPCPE algorithm, the partition and the class parameters are treated as random variables. The method for partitioning a video frame starts with an arbitrary partition and employs an iterative algorithm to estimate the partition and the class parameters jointly [6,27]. Since the successive frames in a video do not differ by much, the partitions of adjacent frames do not differ significantly. Each frame is partitioned by using the partition of the previous frame as an initial condition to speed up the convergence rate of the algorithm. A randomly generated initial partition, a learned partition for the domain or a binary image derived from the background difference is used for the first frame.

The mathematical description of a class specifies the pixel values as functions of the spatial coordinates of the pixel. The parameters of each class can be computed directly by using a least square technique. Suppose we have two classes. Let the partition variable be $c = \{c_1, c_2\}$ and the classes be parameterized by $\theta = \{\theta_1, \theta_2\}$. Also, suppose all the pixel values y_{ij} (in the image data Y) belonging to class k $(k=1,2)$ are put into a vector $\boldsymbol{Y_k}$. Each row of the matrix Φ is given by $(1, i, j, ij)$ and a_k is the vector of parameters $(a_{k0}, \ldots, a_{k3})^T$.

$$y_{ij} = a_{k0} + a_{k1} i + a_{k2} j + a_{k3} ij, \ \forall (i,j) \ y_{ij} \in \boldsymbol{c_k}$$
$$\boldsymbol{Y_k} = \Phi \, a_k$$
$$\hat{a}_k = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{Y_k}$$

We estimate the best partition as that which maximizes the a posteriori probability (MAP) of the partition variable given the image data Y. Now, the MAP estimates of $c = \{c_1, c_2\}$ and $\theta = \{\theta_1, \theta_2\}$ are given by

$$(\hat{c}, \hat{\theta}) = Arg \max_{(c,\theta)} P(c, \theta \mid Y)$$
$$= Arg \max_{(c,\theta)} P(Y \mid c, \theta) P(c, \theta)$$

We assume that the pixel values and parameters are independent and that the parameters are uniformly distributed. We also assume that the error function[1] of y_{ij} is represented by a Gaussian with mean 0 and variance 1. Let $J(c,\theta)$ be the functional to be minimized. With these assumptions the joint estimation can be simplified to the following form:

$$(\hat{c},\hat{\theta}) = Arg \min_{(c,\theta)} J(c_1,c_2,\theta_1,\theta_2)$$

$$J(c_1,c_2,\theta_1,\theta_2) = \sum_{y_{ij}\in c_1} -\ln p_1(y_{ij};\theta_1) + \sum_{y_{ij}\in c_2} -\ln p_2(y_{ij};\theta_2)$$

The minimization of *J* can be carried out alternately on *c* and θ in an iterative manner. Let $\hat{\theta}(c)$ represent the least squares estimates of the class parameters for a given partition *c*. The final expression for $J(c,\hat{\theta}(c))$ can be derived easily and is given by

$$J(c,\hat{\theta}(c)) = Arg \min_{(c_1,c_2)} \left\{ \frac{N_1}{2} \ln \hat{\rho}_1 + \frac{N_2}{2} \ln \hat{\rho}_2 \right\}$$

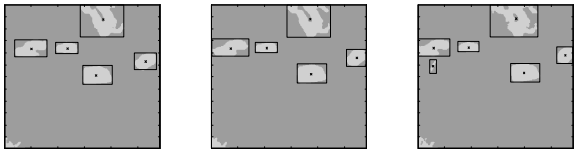
where $\hat{\rho}_1$ and $\hat{\rho}_2$ are the estimated model error variances of the two classes and N_1, N_2 are the number of pixels in each class. The algorithm starts with an arbitrary partition of the data and computes the corresponding class parameters. With these class parameters and the data, a new partition is estimated. Both the partition and the class parameters are iteratively refined until there is no further change in them.

2.3 Object Tracking

The first step for object tracking is to extract the segments in each class from each frame. Then the bounding box and the centroid point for each segment are obtained. For example, Figure 2(b) shows the segmentation results of the video sequence in Figure 2(a), where the vehicle objects belong to class 2 and the ground belongs to class 1. As shown in Figure 2(b), those segments corresponding to the vehicle objects are bounded by their minimal bounding boxes and represented by their centroid points.

The next step for object tracking is to connect the related segments in successive frames. The idea is to connect two segments that are spatially the closest in the adjacent frames [27]. In other words, the Euclidean distances between the centroids of the segments in the adjacent frames are used as the criteria to track the related segments. In addition, size restrictions are employed to determine the related segments in successive frames. A more sophisticated object tracking algorithm integrated into our framework is described in [5], which handles the situation of two objects overlapping under certain assumptions (e.g., the overlapped objects should have similar sizes). As shown in Figure 2 (case 1), there are two overlapped cars being identified as one segment because they are too close. In the algorithm in [5], if the two car objects have ever been separated from each other in the video sequence, then they can be split and identified as two objects, with their bounding boxes being fully recovered, since they have similar sizes.

(a) Example video sequence (frames 40, 41 and 42 from left to right).



(b) Segmentation mask maps and bounding boxes for (a)

Figure 2: Object tracking (case 1)

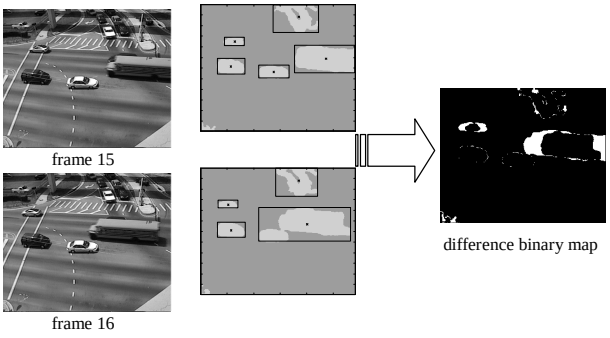


Figure 3: Object tracking (case 2)

On the other hand, in the situation that the overlapped objects have dissimilar sizes (case 2), for example the school bus and car in Figure 3, our existing algorithm [5] cannot find the school bus and car objects corresponding segments in the following frame (frame 16). In this example a large school bus and a small car that were detected as two objects in one frame (frame 15), were merged into a new big segment in the following frame (frame 16). However, from the new detected big segment in frame 16, we can reason that this is an '*overlapping*' segment that includes more than one vehicle object. A difference binary map knowledge discovery method is proposed to discover which objects the '*overlapping*' segment may include.

The idea is to obtain the difference binary map by subtracting the segment result of frame 16 from that of frame 15 and to compare the amount of differences between the two segmentation results of the consecutive frames. As shown in the difference binary map in Figure

[1] The model error is $e_{ij} = y_{ij} - (a_{k0} + a_{k1}i + a_{k2}j + a_{k3}ij)$.

3, the white areas in the difference binary map indicate the amount of differences between the segmentation results of the two consecutive frames. The car and school bus objects in frame 15 can be roughly mapped into the area of the big segment in frame 16 with relatively small differences. Hence, we can discover the vehicle objects in the big segment in frame 16 by reasoning that it is most probably related to the car and school bus objects from frame 15. In such a case, for the big segment (the '*overlapping*' segment) in frame 16, the corresponding links to the car and bus objects in frame 15 will be created.

2.4 Using MATNs and Multimedia Input Strings to Model Video Key Frames

A multimedia augmented transition network (MATN) model can be represented diagrammatically by a labeled directed graph, called a *transition graph*. A multimedia input string is accepted by the grammar if there is a path of transitions which corresponds to the sequence of symbols in the string and which leads from a specified initial state to one of a set of specified final states.

A MATN can build up a video hierarchy [7]. A video clip can be divided into *scenes*, a *scene* contains a sequential collection of *shots*, and each shot contains some contiguous frames that are at the lowest level in the video hierarchy [30]. It is advantageous to use several key frames to represent a shot instead of showing all these frames. Key frames play as the indices for a shot. The key frame selection approach proposed in [7] is based on the number, temporal, and spatial changes of the semantic objects in the video frames. Other features may also be possible for the key frame selections, but we focus on the number, temporal, and spatial relations of semantic objects. Therefore, these key frames can represent spatio-temporal changes in each shot. For example, in each shot of a traffic video sequence, the vehicles may change their positions in subsequent frames and the number of vehicles appearing may change at the time duration of the shot.

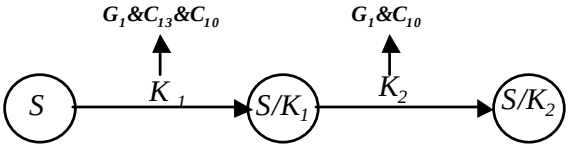
As introduced in [2], one semantic object is chosen as the target semantic object in each video frame and the minimal bounding rectangle (MBR) concept is used. In order to distinguish the 3-D relative positions, twenty-seven numbers are used [2]. In this paper, each frame is divided into nine sub-regions with the corresponding subscript numbers shown in Figure 4(a). Each key frame is represented by an input symbol in a multimedia input string and the "&" symbol between two vehicle objects is used to denote that the vehicle objects appear in the same frame. The subscripted numbers are used to distinguish the relative spatial positions of the vehicle objects relative to the target object "ground" (Figure 4(a)). For simplicity, two consecutive key frames are used to explain how to construct the multimedia input string and the MATN. The multimedia input string that represents these two key frames is as follows:

$$\underbrace{(G_1 \,\&\, C_{13} \,\&\, C_{10})}_{K_1}\ \underbrace{(G_1 \,\&\, C_{10})}_{K_2}$$

There are two input symbols, K_1 and K_2. The order of the vehicle objects in an input symbol is based on the relative spatial locations of the vehicle objects in the traffic video frame (from left to right and top to bottom). For example, the first key frame is represented by input symbol K_1. G_1 indicates that G is the target object. C_{13} *means* the first car object is on the left of and above G, and C_{10} means the second car object is on the left of G. For the next key frame, its multimedia input string is almost the same as that of frame 4 except that the car C_{13} that appeared in the first key frame has already left the road intersection in the next key frame. Hence, the number of vehicle objects decreases from two to one. This is an example to show how a multimedia input string can represent the change of the number of semantic (vehicle) objects.

13	4	22
10	1	19
16	7	25

(a) the nine sub-regions and their corresponding subscript numbers



(b) an example MATN model

Figure 4: MATN and multimedia input strings for modeling the key frames of traffic video shot *S*.

Figure 4(b) is the MATN for the above two key frames of the example traffic video sequence. The starting state name for this MATN is ***S***/. As shown in Figure 4(b), there are two arcs with arc labels the same as the two input symbols (K_1 and K_2). The different state nodes in the MATN model the temporal relations of the selected key frames. The multimedia input strings model the relative spatial relations of the vehicle objects.

3. EXPERIMENT RESULTS AND DISCUSSIONS

A real life traffic video sequence is used to demonstrate the knowledge discovery process, i.e., spatio-temporal vehicle tracking, from the traffic video sequence using the proposed framework.

3.1 Experiment Setup

The traffic video sequence was captured with a *Sony Handycam CCD TR64* and digitized with a simple

Brooktree Bt848 based capture card on a *Windows NT 2000 Celeron* based platform. The video sequence consists of about 16 minutes of video from a traffic intersection with approximately constant lighting conditions. The original video frames were of size 480 rows×640 columns, 24 bit color and frame rate sampled at 5 frames per second. For simplicity and real-time processing purpose, we transform the color video frames to grayscale images and resize them to half of the original size (240 rows×320 columns). The traffic video sequence shows the traffic flow of an intersection on *US 1*, one of the busiest state roads in Miami, FL, USA.

A small portion of the traffic video is used to illustrate how the proposed framework can be applied to traffic applications to answer spatio-temporal queries like "Estimate the traffic flow of this road intersection from 8:00 AM to 8:30 AM." This query requires the use of multimedia data mining techniques to discover information such as the number of vehicles passing through the corresponding road intersection in a given time duration as well as the types of the vehicles (e.g., "car", "bus", etc.). This process can be done in real-time or off-line.

3.2 Experiment Results

The enhanced video segmentation method is applied to the video sequences by considering two classes. The first frame is partitioned into two classes using an initial random partition. After obtaining the final partition of the first frame (via SPCPE), we compute the partitions of the subsequent frames using the previous partitions as the initial partition parameter for the subsequent segmentation steps (since there is little significant difference between consecutive video frames). The convergence speed of the SPCPE algorithm is increased by using the previous partition results and thus provides support for real-time processing. The segmentation results for a few frames – 4, 9, 15, 16 and 35 – are shown in Figure 5 (end of paper) along with the original frames adjacent to them. These frames are the key frames after applying the key frame selection method introduced in [7]. As can be seen, the background of the traffic video sequence is complex. Related work has been done on the base of highway traffic videos [15,20] that have relatively simple backgrounds. Our framework, however can deal with more complex situations such as the traffic video for intersection monitoring.

In Figure 5, the frames in the leftmost column (Figure 5(a)) are the original frames. The second column (Figure 5(b)) shows the difference images after background subtraction. The final segmentation results are shown in the third column (Figure 5(c)). As can be seen from Figure 5(c), almost all of the vehicle objects are captured as separate segments (objects) except for those vehicles in the two lanes located in the upper part of the video frame (which has been captured as one segment because they appear too close together due to the shooting angle of the camera). From Figure 5(c), one can observe that the two-class partitioning schema can capture most of the relevant scene information (in regard to traffic applications). One class captures relevant vehicle information and the second class captures most of the ground information (the background non-vehicle information). Some of the vehicles have been combined with other objects into a single segment when they are closely located, for example, in frame 16, the school bus is overlapped with the car that was waiting in the middle of the intersection, while the school bus was moving westbound. Other cars in the main area of the intersection are successfully identified in all of these frames.

As only the vehicles are important for our application, we use the rightmost column in Figure 5(d) to show the relative spatial relationships of the vehicle segments for each frame. For the simplified segmentation results (Figure 5(d)), we use symbolic representations (multimedia input strings) to represent the spatial relationships of the vehicle objects in each frame. As shown in Figure 5(d), the ground (***G***) is selected as the target object and the segments are denoted by ***C*** for cars or ***B*** for buses. For those cars combined together into a single segment (in the upper part of video frame), we use domain knowledge that there are two lanes located in the upper part of the scene where the vehicles are waiting before they enter the intersection. The use the symbol ***W*** for this special segment indicating that this is a '*waiting*' segment that may include more than one vehicle waiting to enter the intersection. Our data also contains vehicle objects in the main area of intersection that are combined into one segment. For example, the car object and the school bus were combined into one segment in frame 16, while they were separate segments in the preceding frame (frame 15). As discussed in Section 2, this occlusion situation can be detected by the proposed difference binary map knowledge discovery method. We use symbol ***O*** to denote an '*overlapping*' segment which has corresponding links to the related segments in the preceding frame.

As can be seen from Figure 5, the '*waiting*' segment always remains at the same location in the scene. In order to answer the query for traffic flow estimation, these '*waiting*' segments will not be counted. In the proposed symbolic representation, each vehicle segment is indexed in a multimedia input string based on the spatial relation of its centroid. The subscript numbers are used to denote the relative spatial relations of the vehicle objects with respect to the target object from the viewer's perspective. As mentioned earlier, $\boldsymbol{G_1}$ indicates that the ground (***G***) is the target object and the subscript numbers have the same relative spatial meanings. In frames 4 and 9, two cars in the middle of the intersection ($\boldsymbol{C_{10}}$ and $\boldsymbol{C_1}$) were waiting to pass while another car ($\boldsymbol{C_4}$) was driving slowly through the upper part of the intersection westbound. In addition car ($\boldsymbol{C_{13}}$ in frame 4) was leaving the intersection

westbound. In frame 15, a school bus appeared as B_{19} from the east side; while in frame 16, the school bus and the white car (C_1 in frame 15) were combined into one *overlapping* segment (O_{19}). In frame 35, the school bus (B_{10}) was separated from the other cars and left the intersection on the west side, while the two cars (C_{10} and C_1 in frames 4, 9 and 15) made the left turn and moved towards the northeast bound so that their relative spatial locations changed to C_1 and C_{19} in frame 35.

As described above, it can be seen that the multimedia input strings can model not only the number of objects, but also the relative spatial relations. In this case, in order to estimate the intersection traffic flow, we can choose the east or west side of the intersection as a 'judge line' in the frame to determine the traffic flow of the specified direction (east↔west), and any vehicles passing through that line will be recorded. Using the information of centroid's position of each object, the traffic flow of a specified direction in the intersection area can be determined. Moreover, since the types of vehicles are also important for estimating the traffic flow, the sizes of the bounding boxes can be utilized to determine the vehicle types (such as 'car' and 'bus'). For those '*overlapping*' segments, since they have links to specific vehicle segments, the corresponding number and types of vehicles in an overlapping segment can be obtained in order to count the traffic flow. Besides answering the traffic flow query, the proposed framework also has the potential to answer other spatio-temporal related database queries.

4. CONCLUSION

Traffic video analysis can discover and provide useful information such as queue detection, vehicle classification, traffic flow, and incident detection at the intersections. Multimedia data mining techniques need to be employed in order to discover and provide important but previously unknown knowledge from the traffic video sequences to the traffic planners. In this paper, a multimedia data-mining framework that discovers the spatio-temporal relationships of the vehicle objects in the traffic video sequences is presented. The spatio-temporal relationships of the vehicle objects are discovered and captured via the unsupervised image/video segmentation method and the proposed object-tracking algorithm. The discovered spatio-temporal relationships of the vehicle objects are modeled by the multimedia augmented transition network (MATN) model and multimedia input strings. In order to eliminate the complex background information in the traffic video frames, background subtraction techniques are employed. Using the background subtraction technique, both the efficiency of the segmentation process and the accuracy of the segmentation results are improved achieving more accurate video indexing and annotation. This paper uses a real-life traffic video sequence on a state road intersection in Miami, FL, USA as the example video source. As shown in the results, the proposed framework can model complex situations such as the traffic video for intersection monitoring.

5. ACKNOWLEDGEMENT

For Shu-Ching Chen, this research was supported in part by NSF CDA-9711582.

REFERENCES

[1] Caltrans. Caltrans Live Traffic Cameras, http://video.dot.ca.gov/.

[2] Chen, S.-C. and Kashyap, R. L., "A Spatio-Temporal Semantic Model for Multimedia Database Systems and Multimedia Information Systems," *IEEE Transactions on Knowledge and Data Engineering, to appear*.

[3] Chen, S.-C., Shyu, M.-L., and Zhang, C., "An Unsupervised Segmentation Framework For Texture Image Queries," *The 25th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC)*, Chicago, Illinois, USA, Oct. 2000.

[4] Chen, S.-C., Shyu, M.-L., and Zhang, C., "An Intelligent Framework for Spatio-Temporal Vehicle Tracking," *4th International IEEE Conference on Intelligent Transportation Systems*, Oakland, California, USA, Aug. 2001.

[5] Chen, S.-C., Shyu, M.-L., Zhang, C., and Kashyap, R., "Object Tracking and Augmented Transition Network for Video Indexing and Modeling," *12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2000)*, Vancouver, British Columbia, Canada, pp. 428-435.

[6] Chen, S.-C., Sista, S., Shyu, M.-L., and Kashyap, R., "An Indexing and Searching Structure for Multimedia Database Systems," *IS&T/SPIE conference on Storage and Retrieval for Media Databases 2000*, San Jose, CA, U.S.A, pp. 262-270.

[7] Chen, S.-C., Sista, S., Shyu, M.-L., and Kashyap, R., "Augmented Transition Networks as Video Browsing Models for Multimedia Databases and Multimedia Information Systems," *11th IEEE International Conference on Tools With Artificial Intelligence (ICTAI'99)*, Chicago, IL, U.S.A., Nov. 1999.

[8] Coen M, "The Future of Human-Computer Interaction or How I Learned to Stop Worrying and Love my Intelligent Room," *IEEE Intelligent Systems*, vol. 14, no. 2, pp. 8-10, Mar, 1999.

[9] Courtney, J. D., "Automatic Video Indexing via Object Motion Analysis," *Pattern Recognition*, vol. 30, no. 4, pp. 607-625, 1997.

[10] Cucchiara, R., Piccardi, M., and Mello, P., "Image Analysis and Rule-based Reasoning for a Traffic

Monitoring System," *IEEE Trans. on Intelligent Transportation Systems*, vol. 1, no. 2, pp. 119-130, June 2000.

[11] Dailey, D. J., Cathey, F., and Pumrin, S., "An Algorithm to Estimate Mean Traffic Speed Using Uncalibrated Cameras," *IEEE Transactions on Intelligent Transportations Systems*, vol. 1, no. 2, pp. 98-107, Jun, 2000.

[12] DeMenthon, D., Stuckelberg, M., and Doermann, D., "Image Distance using Hidden Markov Models," *International Conference Pattern Recognition (ICPR 2000): Image, Speech and Signal Processing*, Barcelona, Spain, pp. 147-150, Sept. 2000.

[13] Fan, L. and Sung, K. K., "Model-Based Varying Pose Face Detection and Facial Feature Registration in Video Images," *8th ACM International Conference on Multimedia*, Los Angeles, CA, pp. 295-302, Oct. 2000.

[14] Ferman, A. M., Guensel, B., and Tekalp, A. M., "Object-based Indexing of MPEG-4 Compressed Video," *in Proceedings of SPIE: Visual Communications and Image Processing*, San Jose; CA, pp. 953-963, Feb. 1997.

[15] Friedman, N. and Russell, S., "Image Segmentation in Video Sequences: A Probabilistic Approach," *Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence (UAI '97)*, Providence; RI.

[16] Gonzalez, R. C. and Woods, R. E. *Digital image processing*, Reading, Mass: Addison-Wesley, 1993.

[17] Grimson, W. E. L., Stauffer, C., Romano, R., and Lee, L., "Using Adaptive Tracking to Classify and Monitor Activities in a Site," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Preceding*, pp. 22-31, 1998.

[18] Haritaoglu, I., Harwood, D., and Davis, L., "W 4 - Who, Where, When, What: A Real-Time System for Detecting and Tracking People," *IEEE Third International Conference on Face and Gesture Recognition*, Nara, Japan, pp. 222-227, 1998.

[19] Haritaoglu, I., Harwood, D., and Davis, L., "A Fast Background Scene Modeling and Maintenance for Outdoor Surveillance," *15th IEEE International Conference on Pattern Recognition: Applications, Robotics Systems and Architectures*, Barcelona, Spain, pp. 179-183, Sept. 2000.

[20] Huang, T., Koller, D., Malik, J., and Ogasawara, G., "Automatic Symbolic Traffic Scene Analysis Using Belief Networks," *Proceedings of the AAAI, 12th National Conference on Artificial Intelligence (AAAI '94)*, Seattle, WA, pp. 966-972, July 1994.

[21] Kamijo, S., Matsushita, Y., Ikeuchi, K., and Sakauchi, M., "Traffic Monitoring and Accident Detection at Intersections," *IEEE International Conference on Intelligent Transportation Systems*, Tokyo Japan, pp. 703-708, Oct. 1999.

[22] Koller, D., Weber, J., and Malik, J., "Robust Multiple Car Tracking with Occlusion Reasoning," *3rd European Conference on Computer Vision, Eccv '94*, Stockholm Sweden, pp. 189-196, May 1994.

[23] Krumm, J., Harris, S., Meyers, B., Brumitt, B., Hale, M., and Shafer, S., "Multi-Camera Multi-Person Tracking for EasyLiving," *3rd IEEE International Workshop on Visual Surveillance*, Dublin, Ireland, pp. 3-10, July 2000.

[24] Montgomery. Co. Department of Public Works Transportation. ATMS Video Monitoring System Live Traffic Camera Pictures, http://www.dpwt.com/jpgcap/camintro.html.

[25] Puzicha, J., Hofmann, T., and Buhmann, J. M., "Histogram Clustering for Unsupervised Image Segmentation," *IEEE Computer Society Conference Computer Vision and Pattern Recognition*, Fort Collins; CO, pp. 602-608, June 1999.

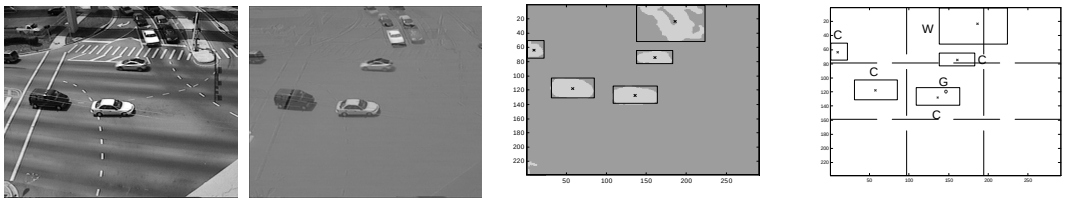
[26] Shafer, S., Krumm, J., Brumitt, B., Meyers, B., Czerwinski, M., and Robbins, D., "The New EasyLiving Project at Microsoft Research," *DARPA/NIST Workshop on Smart Spaces*, pp. 127-130, July 1998.

[27] *Sista*, S. and *Kashyap*, R. L., "Unsupervised Video Segmentation and Object Tracking," *Computers in Industry*, vol. 42, no. 2-3, pp. 127-146, June 2000.

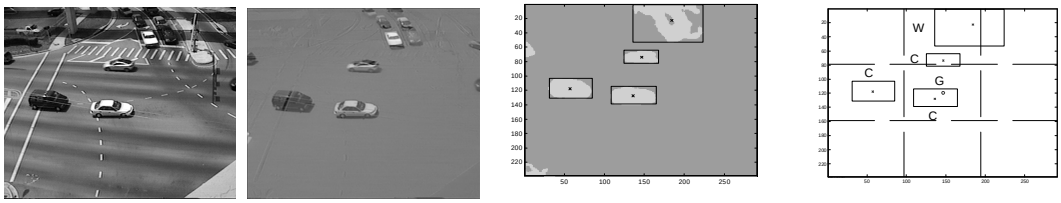
[28] Stauffer, C. and Grimson, W. E. L., "Adaptive Background Mixture Models for Real-Time Tracking," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 246-252, 1999.

[29] Toyama, K., Krumm, J., Brumitt, B., and Meyers, B., "Wallflower: Principles and Practice of Background Maintenance," *7th International Conference on Computer Vision (ICCV'99)*, Held on the Island of Crete, pp. 255-261, Sept. 1999.

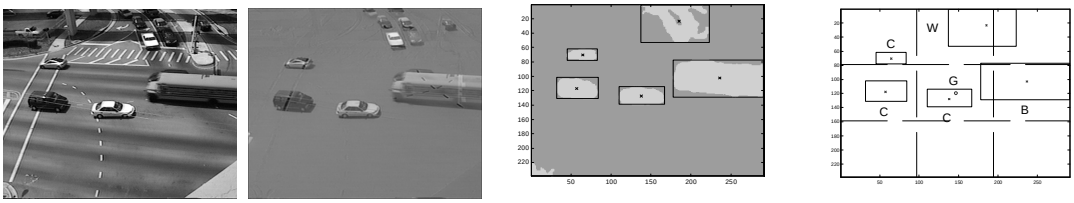
[30] Yeo, B.-L. and Yeung, M. M., "Retrieving and Visualizing Video," *Communications of the ACM*, vol. 40, no. 12, pp. 43-52, Dec. 1997.



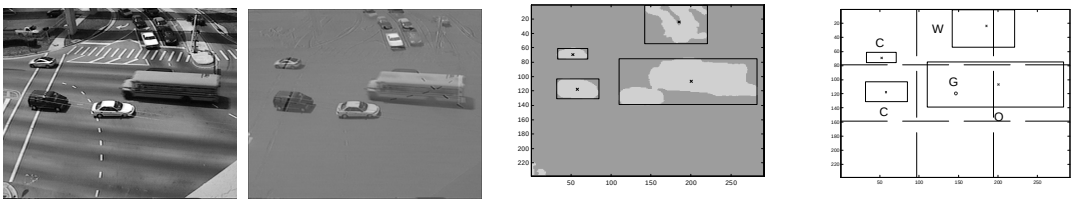
Frame 4 *Multimedia Input String:* $G_1\&C_{13}\&C_{10}\&C_1\&C_4\&W_4$



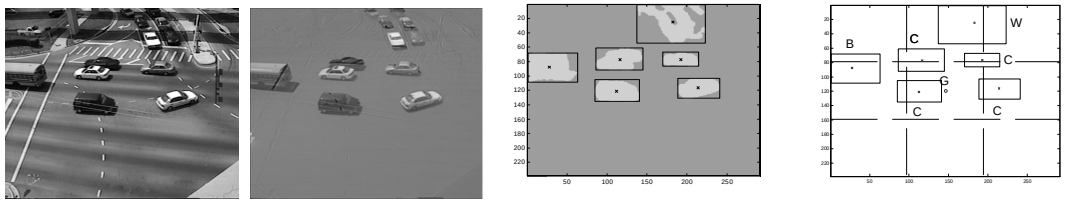
Frame 9 *Multimedia Input String:* $G_1\&C_{10}\&C_1\&C_4\&W_4$



Frame 15 *Multimedia Input String:* $G_1\&C_{13}\&C_{10}\&C_1\&W_4\&B_{19}$



Frame 16 *Multimedia Input String:* $G_1\&C_{13}\&C_{10}\&W_4\&O_{19}$



Frame 35 *Multimedia Input String:* $G_1\&B_{10}\&C_1\&C_4\&W_4\&C_4\&C_{19}$

(a) Original frames. (b) Difference frames. (c) Segmentation results. (d)Bounding boxes.

Figure 5: Segmentation results as well as the multimedia input strings for frames 4, 9, 15, 16 and 35. The leftmost column gives the original video frames; the second column shows difference images obtained by subtracting the background reference frame from the original frames; the third column shows the vehicle segments extracted from the video frames, and the rightmost column shows the bounding boxes of the vehicle objects.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

A Bayesian Learning Algorithm of Discrete Variables for Automatically Mining Irregular Features of Pattern Images

Hanchuan Peng
Center for Biomedical Image Computing,
Department of Radiology,
Johns Hopkins University, School of Medicine
601 North Caroline St, JHOC#3230,
Baltimore, MD, 21287, USA.
Email: phc@cbmv.jhu.edu

Fuhui Long
Department of Neurobiology, Box 3209,
Duke University Medical Center,
Durham, NC, 27710, USA.
Email: long@neuro.duke.edu

Abstract

Multimedia data mining often involves difficult problems where knowledge about patterns should be extracted automatically from datasets. It is also a critical task in pattern recognition systems to automatically extract highly irregular features of pattern images[#1], especially when these features (e.g. structural features of unconstrained handwritten characters, video objects, etc.) can hardly be described in a quantitative way. In this paper we propose an image mining algorithm for irregular feature extraction. This algorithm is based on learning belief networks of pattern image pixels, each of which is regarded as a discrete variable with a limited number of states. The probability of belief network, i.e. Bayesian metric, is chosen to measure the associations between image pixels and the pattern image category. A heuristic algorithm is designated to learn the ψ-structure of belief network, where clusters of "equivalent" pixels are regarded as the irregular features. We test the algorithm on both simulated data and real data (i.e. unconstrained handwritten characters). For simulated data, the algorithm can successfully discover the probabilistic associations implied by the ground truth mask. In the character feature extraction experiments, a hierarchy of statistically optimal feature vectors is obtained by averaging the pixel clusters over many independent experiments. The irregular features' soundness is verified with neural classifiers. Our results show that the heuristic Bayesian learning algorithm can produce significantly better features than the multi-layer perceptron, learning vector quantization network, sparse trace neural network, and simple template matching method.

Key Words

Image Mining, Belief Network, Bayesian Learning, Feature Extraction

1. Introduction

Data mining is a process by which previously unknown information and patterns are extracted from large quantities of data [16]. Three factors, i.e. statistical reasoning, machine learning techniques, and well-organized data, are important to the good outcome [16]. However, as a part of multimedia mining techniques [17], image mining, which seeks associations among different images from large image databases, is still very difficult [16]. A very similar problem is spatial mining [6] (i.e. knowledge discovery in spatial databases), which is termed as extracting implicit knowledge, spatial relations, or other implicitly stored patterns from spatial databases. Notably image mining has great resemblance with the traditional pattern feature extraction problem, especially when the highly irregular features of patterns (e.g. unconstrained handwritten characters, video objects, etc.) can hardly be described in a quantitative manner.

Feature extraction is one of the essential concepts of pattern analysis and recognition. A typical pattern recognition system can be divided into two basic cascade parts, i.e. feature extractor and feature classifier [3]. A weakly designed feature extractor will immediately lead to poor classification. Unfortunately, generally saying, extracting good features from an arbitrarily given pattern dataset is never easy because the definition of feature varies from one area to another. Consequently, pattern feature extraction techniques are usually restricted to specific engineering fields.

Based on some statistical criteria, data mining techniques can be expected to extract good pattern features. In this paper we discuss how to make use of the belief network to represent the knowledge of the irregular features of pattern images, and how to learn such belief network representation with a heuristic Bayesian model selection method. With a neural classifier, we also compare the proposed method with Multi-Layer Perceptron (MLP) with the Back-Propagation (BP) algorithm [15], a simple version of Learning Vector Quantization (LVQ) network [10], Sparse Trace Neural

Network (STNN) [13,14], and template matching (TM) method.

This paper is organized as follows. Section 2 presents our image mining framework with the belief network. Section 3 shows the heuristic learning algorithm. Section 4 gives results on the simulated data. Section 5 gives results on the unconstrained handwritten characters. Section 6 gives some brief discussions and the conclusion.

2. Belief Network and Image Mining

Belief network, which also has names such as Bayesian network, causal network, etc, is the Directed Acyclic Graph (DAG) model that describes the probabilistic relationships among events/variables [9]. Each node in the network represents a variable or event. Each directed arc, or arrow, between nodes is the conditional probability of the child node (the arrow's ending point) given the parent node (the arrow's starting point). Once a correct network structure is established and all parameters (i.e. values of arcs) are obtained, the knowledge about this set of events/variables is captured. Obviously, belief network will be a powerful tool to represent features of pattern images, – the feature extraction problem turns to be constructing a belief network of some low-level attributes (e.g. pixels) and categories of the pattern images.

There are two main approaches to constructing belief network from data [5, 7]. The constraint-based method uses the category information about the conditional independence constraints to construct a belief network. Differently, the Bayesian learning method weights the degree that the conditional independence constraints hold. The Bayesian learning method has three advantages over the constraint-based method [7]: (1) Bayesian approach is not susceptible to the incorrect categorical information of the data; (2) with Bayesian approach, finer distinctions among the belief network structures – both quantitative and qualitative – can be made; (3) information about several network structures can be combined and utilized for a better inference.

Bayesian learning belief network structures involves two issues: the measurement of network structures (i.e. standard for model selection) and the heuristic searching/learning strategy. Some existing heuristic learning algorithms include K2 [2, 8], K3 [1], etc. K2 uses the Bayesian metric, i.e. the conditional probability of each belief network structure given the data, as the measurement. K3 uses Minimum Description Length (MDL) to measure the network. It has been shown that the MDL metric and Bayesian metric in some situations can only differ in a constant independent of the case number of data [1].

We choose the Bayesian metric as the measurement of belief network structures because it offers a clear picture of which network structure is going to be selected – the algorithm will learn the network structure with the larger conditional probability given the data. Particularly, denote network structure as B and data as D, we have

$$M(B) = p(B \mid D) = \frac{p(B)p(D \mid B)}{p(D)} \propto p(D \mid B) \tag{1}$$

where $M(B)$ is the metric function, $p(B)$ and $p(D)$ are the priors of network structure and data, $p(D|B)$ is the likelihood function. Suppose $p(B)$ is uniformly distributed [2, 8], we see that $M(B)$ is proportional to the likelihood function, which takes the following form for discrete state variables [2, 8] when D is complete (i.e. every variable is observed in every case of D):

$$p(D \mid B) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{2}$$

where n is the variable number of the belief network, r_i is the number of state of the ith variable, q_i is the number of instantiations of the parent variables (denoted as π_i) of the ith variable, N_{ijk} is the number of cases where the ith variable takes the kth state while the set of its parents takes the jth instantiation, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ is the total case number of the ith variable's parents taking the jth instantiation.

Suppose we have the complete data D, some observable attributes of the pattern images can be defined as variables of the belief network. Simply, in this paper we regard each image pixel as a variable, i.e. an image I is described as a set of pixel variables $\{V_i, i=1,2,\ldots,n\}$, where V_i is the ith pixel and n is the total pixel number of each pattern image. We also define the pattern image category as a variable C. Further, we notice that pattern features consist of pixel clusters, thus we define a set of cluster variables, $R=\{U_j, j=1,2,\ldots\}$, where each cluster is the weighted combination of some pixels.

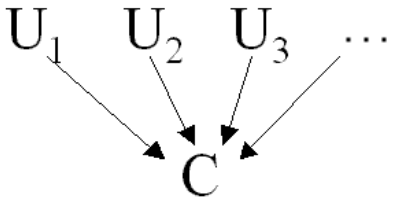


Fig.1 The ψ-structure of belief network

With the above notations, the irregular feature extraction problem can be described as constructing a belief network of cluster variables and the category variable. Particularly, in this paper we consider the "ψ-structure"[#2] depicted in Fig.1, where the parents U_j, $j=1,2,\ldots$, have converging arrows to the child node C. Obviously with ψ-structure we assume that cluster variables are marginally independent of each other, and

the pixel variables in a cluster possess strong associations between each other (we call such strong association as the "equivalence" of pixel variables.)

Notice that the ψ-structure is different from the traditional naive Bayesian model for pattern classification. In this paper we emphasize the pattern feature extraction, but not direct classification with Bayesian model, although the comparison is interesting.

3. Bayesian Learning Algorithm

A network growing method is designated to learn the ψ-structure from data. In each step when we decide if a new cluster variable should be added to the network, we compare the model with an arrow from the new variable to the category variable and the model without such an arrow. The model with the highest positive difference of metric values will be selected (it can be shown that if the association is stronger, the difference of Bayesian metric values is larger). The pixel variable for such model is selected as the cluster variable. Then another model selection procedure is employed to decide the equivalent pixel variables among all pixel variables that are conditional independent of C given the current cluster variable. The above procedures repeat until all pixel variables have been evaluated. Details of the algorithm can be seen in Table 1.

With the pair-wise comparison of the Bayesian metric values, we always select model where the data favor an arrow from the current cluster variable to category variable. Such an association between that cluster and the pattern category is further required to be the strongest one among all the currently possible associations. Additionally, by finding S_{j-1} ($j\geq 2$) we obtain a group of pixel variables that are conditional independent of C given U_{j-1}, although variables in S_{j-1} possess associations to C without U_{j-1}. Typically this conditional independence implies strong association between U_{j-1} and each variable in S_{j-1}. Therefore it is reasonable to locate the equivalent pixels using the strong constraint of probabilistic equivalence within S_{j-1} and put together U_j and S_j as the jth group of features.

Notably each pair of U_i,U_j being marginally independent does not mean every pixel variable in the ith cluster is marginally independent of every pixel variable in the jth cluster. However, although in this paper we do not search for more complex belief network structures, the experiments show for the data in next two sections there is little correlation between such variables. Further, according to the ordering of such clusters (recall that they are sequentially obtained according to the strength of associations), we can readily get a hierarchy of features for pattern classification.

In Step 7 the equivalent variables are determined by thresholding the conditional probabilities, i.e. for every state the conditional probability is demanded to be larger than a preset threshold. There are some methods to obtain such threshold adaptively. For example for each S_j the Expectation Maximization (EM) algorithm can be used to find the best threshold. Besides the conditional probability criterion, pairs of Bayesian metric values can also be used to determine the equivalent variables (actually it is not difficult to prove that conditional probabilistic equivalence is a sufficient condition for metric equivalence). Other alternative methods for clustering the pixel variables are discussed in another paper on medical image mining [12].

Table 1. The heuristic learning algorithm

Step 1	Initialization: let $I=\{V_i, i=1,2,\ldots,n\}$, $R=\{\}$ (i.e. set the cluster variable set as empty), $B=\{C\}$ (i.e. a belief network with only the category variable), $j=1$.
Step 2	For each $V_i \in I$, add V_i to B and compare the pair of metrics of belief networks with arrow $V_i \rightarrow C$ and without such arrow. Find i^* such that $i^*=\text{argmax}(M\{V_i \rightarrow C\}-M\{V_i,C\})$. If $M\{V_{i^*} \rightarrow C\} \leq M\{V_{i^*},C\}$, then go to Step 10.
Step 3	Let $U_j=V_{i^*}$ (i.e. use V_{i^*} to represent the new cluster variable) and $R=R\cup U_j$.
Step 4	Find pixel variable set $S_{j-1}=\{V_i$, where $M\{V_i \rightarrow C\} \leq M\{V_i,C\}\}$.
Step 5	$I=I \setminus \{V_{i^*} \cup S_{j-1}\}$ (i.e. exclude V_{i^*} and S_{j-1} from I).
Step 6	If $j=1$, then go to Step 9.
Step 7	In S_{j-1} find subset E_{j-1}, where each variable V_i satisfies the conditional probability $p(U_{j-1}=k\|V_i=k)\approx 1$ for every state k.
Step 8	The $(j-1)$th cluster is found as $\{U_{j-1}\} \cup E_{j-1}$, denote E_{j-1} as the equivalent pixel variable set of U_{j-1}.
Step 9	If I is empty, then go to Step 10; otherwise $j=j+1$ and go to Step 2.
Step 10	Terminate the algorithm and output B, R, and E_j, $j=1,2,\ldots$.

4. Mining Simulated Pattern Images

Here we generate simulated data with an underlying probabilistic structure and then test if the algorithm can discover the interesting probabilistic associations between image pixel variables and pattern image categories.

We design a 4-node ψ-structure belief network, where each of the three nodes n_1,n_2,n_3 has an arrow pointing to the fourth node C. Each of these 4 nodes has 2 states, i.e. "1" and "2". The conditional probabilistic structure is listed as Table 2, where for example, we see the conditional probability of $p(C=1|n_1=1,n_2=1,n_3=1)=0.8000$ and $p(C=2|n_1=1,n_2=2,n_3=1)=0.2300$.

Without loss of generality, we arbitrarily draw a 16×16 ground truth mask as shown in Fig.2(a), where there are 8 regions belonging to 3 groups (in different colors), respectively. In each group we randomly choose one pixel as the variable n_i, $i=1,2,3$. Then we randomly sample these three variables and the state of C according to the

conditional probabilistic structure in Table 2. For other pixels in each of these three region groups, we sample their cases according to another conditional probability T=0.8, e.g. $p(n_{1a}=1|n_1=1)=T$ and $p(n_{1a}=2|n_1=2)=T$, where n_{1a} stands for any pixel variable (except n_1) in group 1. For all the pixels outside of these three region groups, we set them as noise variables, which have a marginal probability 0.5. We generate a binary pattern image dataset of 1000 images (each image corresponds to a case of C). Two such cases are shown in Fig.2(b) and (c), where no apparent feature can be seen in common sense.

Table 2. The ground truth probabilistic structure

States of n_1,n_2,n_3			Conditional Probability of C's states given n_1,n_2,n_3	
n_1	n_2	n_3	1	2
1	1	1	0.8000	0.2000
1	1	2	0.7800	0.2200
1	2	1	0.7700	0.2300
1	2	2	0.7500	0.2500
2	1	1	0.7600	0.2400
2	1	2	0.7000	0.3000
2	2	1	0.7100	0.2900
2	2	2	0.1000	0.9000

(a) (b) (c)

Fig.2 (a) Ground truth mask (b,c) two sample images

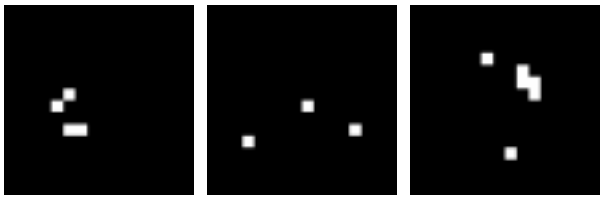


Fig.3 Three identified groups of pixel variables

We apply the proposed algorithm to the simulated data (a threshold 0.8 is used to decide the equivalent variables). As shown in Fig.3, the three region groups are well separated and the noise variables are excluded. Clearly our algorithm successfully locates the pixels that have interesting probabilistic associations with the pattern image category variable.

5. Mining Unconstrained Handwritten Digits

5.1 Design

Unconstrained handwritten character recognition is a difficult pattern classification problem and has been used as the test bed for a lot of pattern analysis and recognition algorithms [4]. There is a long history of debate on how to define the "features" of unconstrained handwritten characters because such features can hardly be described in an accurate way without any exceptions. The template-based matching method, statistical methods, and neural networks have been extensively applied to such problem. Here we apply the Bayesian learning algorithm to mine the unknown features, and compare the features extracted by a couple of existing methods.

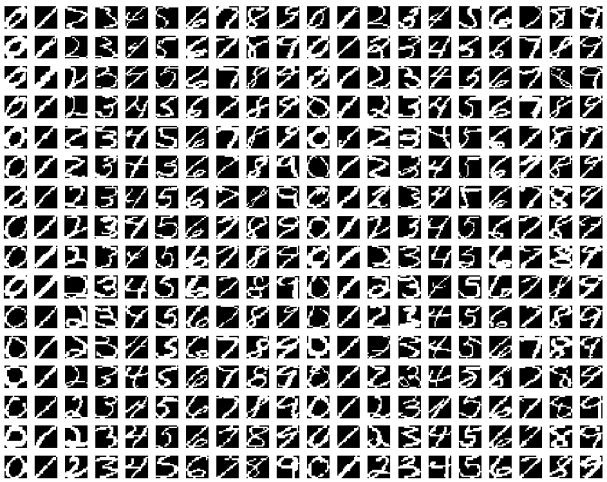


Fig.4 Some examples of unconstrained handwritten digits

Here we use a dataset from the CENPARMI lab in Condordia University in Canada. This data set contains 6000 unconstrained handwritten digits of 0~9, which are normalized to be 16×16 pixels. Each pixel takes two states, i.e. 1 and 0. For the purpose of evaluation, we randomly separate this dataset as a training set with 4800 samples and a testing set with 1200 samples. 320 samples of the dataset are shown in Fig.4.

Instead of mining the features of the 10 types of digits simultaneously, we try to identify the features of these pattern images separately. For each of the 10 digits, we produce its dataset separately. For example, for digit 0, we use all character images of digit 0 as one part of the training data and all other 9 types of character images as the complementary part of the training data. Accordingly, the category variable C for pattern samples of digit 0 takes state 'yes' and for all other pattern samples takes state 'no'. This paradigm is depicted in Fig.5: from the Dataset I of all 4800 training samples, we separate the 480 images of digit 0 as Dataset II and all other 4320 images as Dataset III. Then from Dataset III we random sample 480 images as Dataset IV. Then we combine Dataset II and Dataset IV (totally 960 cases) to produce the training set for digit 0, i.e. Dataset V. The category

variable *C* has state 'yes' for Dataset II and state 'no' for Dataset IV. For all other digits, the training set is generated in the same way.

Since that the Dataset IV is randomly sampled from Dataset III, the features mined from Dataset V vary from trial to trial, although these variations are predictably small. We perform *N* times of such sampling and mining procedures and then average the extracted features (clustered pixels) to obtain grayscale feature vectors for each type of digits. In our experiments, *N* is chosen as 10.

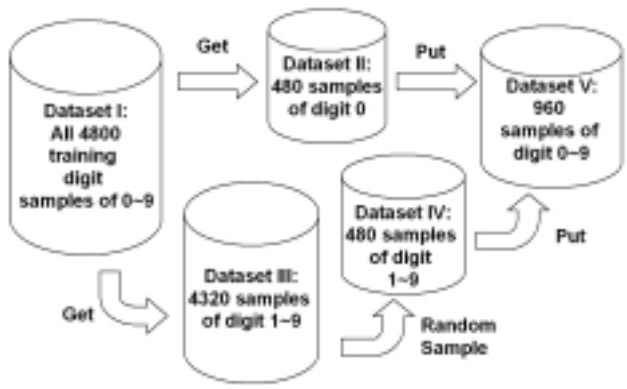


Fig.5 Procedure to form the input dataset (Dataset V) of digit 0

The soundness of the extracted features can be compared to what extracted by existing methods. We utilize a 3-layer feedforward neural network, whose input layer has 16×16 neurons, the hidden layer has 16×16 neurons (each neuron has an *r*×*r*-sized Receptive Field (RF)) for feature extraction, and the output layer has 10 neurons for feature classification. This neural network has been used in [13, 14] to compare the MLP, LVQ-NN and STNN. We apply the feature vectors produced by the Bayesian learning algorithm (referred as BN method hereafter) as the fixed trace of STNN, and then force the hidden layer neurons to learn such fixed features. (In STNN's scheme, the "trace" is the "feature" of patterns. Further, in comparison with such fixed features, MLP doesn't has any constraint of such feature vectors, STNN updates the feature vectors adaptively, and unlike STNN, LVQ-NN doesn't possess the sparse term although it also updates the feature vectors adaptively). To compare with the simple template matching method, we generate the average pattern images as the classification templates (referred as AV method hereafter). To make a fair comparison, in each trial of experiments we set the initial parameters (weights) of neural network exactly the same and then train each neural network for 100 iterations. Then we examine its classification accuracies on both training set (4800 samples) and testing set (1200 samples). If a comparing method has better features, it should have faster convergence and higher classification accuracy in training and better generalization (higher classification accuracy) in testing.

5.2 Experimental Results

With the BN method, we can identify (about) 8 feature clusters for each type of digit. To investigate if these features can serve as digit templates, we directly use these feature vectors to classify both the training set and the testing set. (The method is: for each character image, compute the L_2 norm as the distance and categorize it to the template that has the minimum distance to the current character image.) Results (percentage of correct classification) in Fig.6 show that the optimal number of feature clusters is 6. The reason for worse classification accuracies of more feature clusters might be over-fitting.

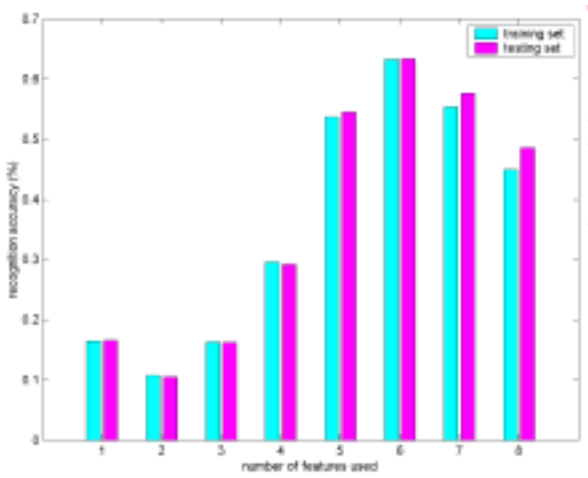


Fig.6 Classification accuracy of different feature hierarchies

We compare the classification of the BN method and AV method using (1) the above simple template matching method and (2) STNN (RF size *r*=3; feature vectors of BN and AV are used as the fixed "traces"). Results in Table 3 show that BN method performs better. Notably although BN method uses more time in obtaining the features than the simple AV, once BN method gets the better features, the training time for BN-NN is predictably significantly shorter than AV-NN (we do not compare speeds here because we fix training iteration number for a fair comparison of classification accuracies).

Table 3. Classification accuracies of template methods

Methods	Training set (%)	Testing set (%)
BN Template	63.40	63.30
AV Template	59.20	56.60
BN-NN	91.06	89.17
AV-NN	85.48	83.17

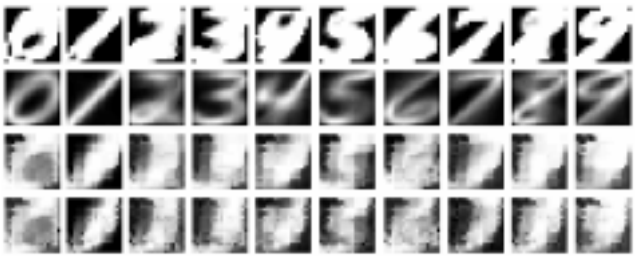
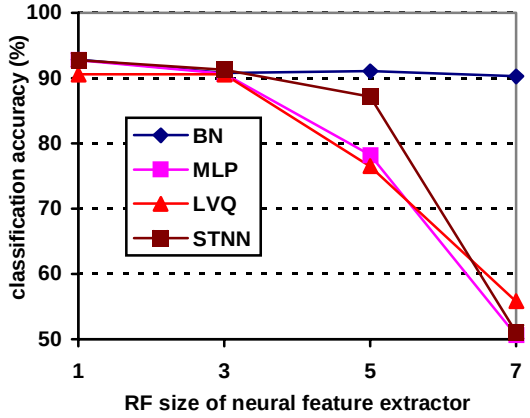
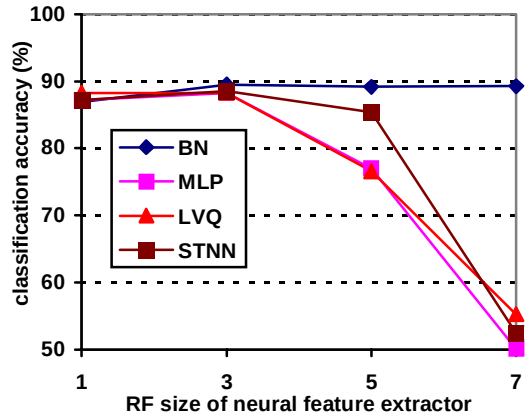


Fig.7 Comparison of the feature vectors. The first row: feature vectors with the first 6 feature clusters of BN method. The second row: feature vectors of AV method. The third row: the feature codebook vectors produced by the LVQ-NN. The fourth row: the traces (feature codebook vectors) produce by the STNN.



(a) For training set



(b) For testing set

Fig.8 Comparison of classification accuracies for both training set and testing set

In Fig.7 the extracted feature vectors are compared visually. We see BN method can produce perceptually more meaningful features than AV, LVQ and STNN. Particularly, the feature vectors of AV lose many key points that may be important to distinguish the patterns (for example, AV does not emphasize the lower part of digit 4 and the lower cave of digit 5). LVQ almost cannot produce perceptually meaningful features (although this is not necessary to indicate such features are meaningless).

In Fig.8 we compare the classification accuracies of several methods. We see that with different receptive field size r (1,3,5,7) of hidden layer neuron, the neural classifier with BN feature vectors always achieves higher classification accuracies than other methods. Therefore we can conclude that Bayesian learning method does produce better features.

6. Discussions and Conclusion

It is shown that the Bayesian learning algorithm can be used to extract a hierarchy of irregular features of pattern images. These features are optimal in the sense of Bayesian model selection. In the comparing experiments, the proposed method produces better results than MLP, LVQ, STNN and TM, although the Bayesian learning module is an additional part to the neural classification systems.

Since that the proposed algorithm learns the ψ-structure belief network, the clusters are assumed to be marginally independent. This point is not necessary to be true in all situations. The algorithm can be refined to search for more elegant structures, to decide which clusters are mutually associated, and to achieve the optimal clusters automatically.

In this paper the algorithm is only applied to the bi-state variables. It can also be applied to multi-state variables in more general multimedia mining and retrieval applications, e.g. semantic video object extraction [11]. For example, we can set the category variable in Section 5 taking 10 states and then the algorithm can directly construct a belief network from Dataset I. One drawback of such approach is that we have to look at the conditional probabilities (i.e. parameters of the belief network) after the belief network has been obtained to decide which feature corresponds to which category. This is one reason why we only discuss bi-state variables in this paper to illustrate the idea in a clearer way.

Besides, although the variable number in our experiments is only 257, while the case num is around 1000, our method has the power to deal with more challenging applications, e.g. medical image mining, where there are millions of variables while each variable only has tens of cases [12]. For example, the algorithm in Table 1 has been extended and applied for medical image mining to detect both linear and nonlinear associations between registered MRI image voxels and subjects' functional deficits. Information about such applications can be found at our web site http://cbic1.rad.jhu.edu/~phc/demo.htm.

It is also an interesting issue to compare the classification accuracy of the proposed method with the naive Bayesian classifier and a latent variable Bayesian classifier. The related results will be given elsewhere.

Acknowledgement

We are grateful to Dr. Edward Herskovits and Dr. Christos Davatzikos for so many enlightening discussions on the Bayesian learning method. Thanks also go to the anonymous referees for their suggestive comments to improve this paper.

References

[1] Bouckaert, R.R., "Probabilistic network construction using the minimum description length principle," Technical Report, RUU-CS-94-27, Dept of Computer Science, Utrecht University, 1994.

[2] Cooper, G., and Herskovits, E., "A Bayesian method for the induction of probabilistic networks from data," Machine Learning, Vol.9, pp.309-347, 1992.

[3] Fu, K.S., Applications of Pattern Recognition, CRC Press, 1982.

[4] Gan, Q. and Suen, C.Y, "Neural networks for handwritten character recognition," in Fuzzy Logic and Neural Network Handbook, McGraw-Hill Book Company, pp.16.1-16.6, 1996.

[5] Glymour, C., and Cooper, G., (Ed.), Computation, Causation, & Discovery, AAAI Press & MIT Press, 1999.

[6] Han, J., "Mining spatial datasets: a new frontier for data mining," Spatial Data Mining Panel Talk, 2000 Int. Workshop on Mining Scientific Databases, Univ. of Minnesota, Minneapolis, Minnesota, July 2000. (For more information see http://www.cs.sfu.ca/~han/ and http://www.cs.sfu.ca/people/GradStudents/koperski/personal/research/research.html)

[7] Heckerman, D., "Bayesian networks for data mining," Data Mining and Knowledge Discovery, Vol.1, No.1, pp.79-119, 1997.

[8] Herskovits, E.H., "Computer-based probabilistic-network construction," Doctoral Dissertation, Medical Informatics, Stanford University, 1991.

[9] Jensen, F.V., An Introduction to Bayesian Networks, UCL Press, 1996.

[10] Kohonen, T., Self-Organizing Maps, Springer-Verlag Press, 1995.

[11] Long, F.H., Feng, D., Peng, H.C., and Siu, W., "Semantic video object extraction," IEEE Computer Graphics & Applications, Vol.21, No.1, pp.48-55, 2001.

[12] Peng, H.C., Herskovits, E.H., Davatzikos, C., "Bayesian methods for detecting human brain atrophy from MRI images," submitted to IEEE Trans. Medical Imaging, 2001.

[13] Peng, H.C., Sha L, Gan Q., and Wei Y., "Combining sigmoid packet and trace neural network for fast invariance learning," Electronics Letters, Vol.34, No.9, pp.898-900, 1998.

[14] Peng, H.C., Sha L, Gan Q., and Wei Y., "New Energy function for learning invariances in multilayer perceptron," Electronics Letters, Vol.34, No.3, pp.292-294, 1998.

[15] Rumelhart, D.E., Hinton, G.E., and Williams, R.J., "Learning representations by back-propagating errors," Nature, Vol.323, pp.533-536, 1986

[16] Thuraisingham, B., Data Mining: Technologies, Techniques, Tools, and Trends, CRC Press, 1999.

[17] Zaiane, O.R., Han, J., Li, Z.N., Chiang, J.Y., and Chee, S., "Multimedia-miner: a system prototype for multimedia data mining," Proc. 1998 ACM-SIGMOD Conf. on Management of Data, (system demo), Seattle, Washington, June 1998.

[#1] In this paper, "pattern image" is termed as an image containing perceptually meaningful patterns. Examples include images of unconstrained handwritten characters, video frames of semantic objects, etc. Usually, it is difficult to describe the key features of such patterns in an accurate and quantitative way, however, such descriptions, if obtained, will play an important role in identifying these patterns.

[#2] The so-called ψ-structure resembles the often-mentioned v-structure of belief network, where two arrows converge to a common child node.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

Application of Data Mining Techniques for Medical Image Classification

Maria-Luiza Antonie
Database Laboratory
Department of Computing Science
University of Alberta
Canada
email: luiza@cs.ualberta.ca

Osmar R. Zaïane
Database Laboratory
Department of Computing Science
University of Alberta
Canada
email: zaiane@cs.ualberta.ca

Alexandru Coman
Database Laboratory
Department of Computing Science
University of Alberta
Canada
email: acoman@cs.ualberta.ca

ABSTRACT
Breast cancer represents the second leading cause of cancer deaths in women today and it is the most common type of cancer in women. This paper presents some experiments for tumour detection in digital mammography. We investigate the use of different data mining techniques, neural networks and association rule mining, for anomaly detection and classification. The results show that the two approaches performed well, obtaining a classification accuracy reaching over 70% percent for both techniques. Moreover, the experiments we conducted demonstrate the use and effectiveness of association rule mining in image categorization.

KEYWORDS
classification, medical imaging, association rule mining, neural networks, image categorization, image mining.

1. Introduction

The high incidence of breast cancer in women, especially in developed countries, has increased significantly in the last years. Though much less common, breast cancer also occurs in men [1][15, 14]. The etiologies of this disease are not clear and neither are the reasons for the increased number of cases. Currently there are no methods to prevent breast cancer, which is why early detection represents a very important factor in cancer treatment and allows reaching a high survival rate. Mammography is considered the most reliable method in early detection of breast cancer. Due to the high volume of mammograms to be read by physicians, the accuracy rate tends to decrease, and automatic reading of digital mammograms becomes highly desirable. It has been proven that double reading of mammograms (consecutive reading by two physicians or radiologists) increased the accuracy, but at high costs. That is why the computer aided diagnosis systems are necessary to assist the medical staff to achieve high efficiency and effectiveness.

[1]In the United States, for example, male breast cancer accounts for 1 of every 100 cases of breast cancers [15]

The methods proposed in this paper classify the digital mammograms in two categories: normal and abnormal. The normal ones are those characterizing a healthy patient. The abnormal ones include both benign cases, representing mammograms showing a tumour that is not formed by cancerous cells, and malign cases, those mammograms taken from patients with cancerous tumours. Digital mammograms are among the most difficult medical images to be read due to their low contrast and differences in the types of tissues. Important visual clues of breast cancer include preliminary signs of masses and calcification clusters. Unfortunately, in the early stages of breast cancer, these signs are very subtle and varied in appearance, making diagnosis difficult, challenging even for specialists. This is the main reason for the development of classification systems to assist specialists in medical institutions. Due to the significance of an automated image categorization to help physicians and radiologists, much research in the field of medical images classification has been done recently [16, 20, 9]. With all this effort, there is still no widely used method to classifying medical images. This is due to the fact that the medical domain requires high accuracy and especially the rate of false negatives to be very low. In addition, another important factor that influences the success of classification methods is working in a team with medical specialists, which is desirable but often not achievable. The consequences of errors in detection or classification are costly. Mammography alone cannot prove that a suspicious area is malignant or benign. To decide that, the tissue has to be removed for examination using breast biopsy techniques. A false positive detection may cause an unnecessary biopsy. Statistics show that only 20-30 percentages of breast biopsy cases are proved cancerous. In a false negative detection, an actual tumour remains undetected that could lead to higher costs or even to the cost of a human life. Here is the trade-off that appears in developing a classification system that could directly affect human life. In addition, the tumours existing are of different types. Tumours are of different shapes and some of them have the characteristics of the normal tissue. All these reasons make the decisions that

are made on such images even more difficult.

Different methods have been used to classify and/or detect anomalies in medical images, such as wavelets [3, 20], fractal theory [8], statistical methods [6] and most of them used features extracted using image-processing techniques [16]. In addition, some other methods were presented in the literature based on fuzzy set theory [2], Markov models [7] and neural networks [9, 5]. Most of the computer-aided methods proved to be powerful tools that could assist medical staff in hospitals and lead to better results in diagnosing a patient.

In this paper, we use a common classification method, namely neural networks, but significantly improve the accuracy rate of the classifier compared to other published results using the same data set. In addition, we investigate the use of association rules, typically used in market basket analysis, in the problem of image categorization and demonstrate with encouraging results that association rule mining is a promising alternative in medical image classification and certainly deserves further attention. To the best of our knowledge, association rules have never been used for image categorization. Some research work was published showing the use of FP-growth algorithm [11] for building classifiers [17]. We have also studied text categorization with association rules [21].

The rest of the paper is organized as follows: Section 2 depicts the general classification process, presents the data collection used for benchmarking and describes the image pre-processing phase. Feature extraction is also presented in Section 2. Classification of images using neural networks is presented in Section 3 and classification of images with association rules is introduced in Section 4. In Section 5, we discuss our experiments and the results. Conclusions are presented in Section 6.

2. Data Collection and Preprocessing

To automatically categorize medical images, we have experimented on real mammograms with two data mining techniques, association rule mining and neural networks. In both cases, the problem consists of building a mammography classification model using attributes extracted from and attached to mammograms, then evaluating the effectiveness of the model using new images. The process of building the classification model (classifier) includes preprocessing and extraction of visual features from already labelled images (i.e. training set).

Figure 1 shows an overview of the categorization process adopted for both systems. The first step is represented by the image acquisition and image enhancement, followed by feature extraction. The last one is the classification part where the technique for supervised learning is different. All these parts of the classification systems are discussed in more detail later.

2.1 Mammography Data Collection

To have access to real medical images for experimentation is a very difficult undertaking due to privacy issues and heavy bureaucratic hurdles. The data collection that was used in our experiments was taken from the Mammographic Image Analysis Society (MIAS) [18]. This same collection has been used in other studies of automatic mammography classification. Its corpus consists of 322 images, which belong to three big categories: normal, benign and malign. There are 208 normal images, 63 benign and 51 malign, which are considered abnormal. In addition, the abnormal cases are further divided in six categories: microcalcification, circumscribed masses, spiculated masses, ill-defined masses, architectural distortion and asymmetry. All the images also include the locations of any abnormalities that may be present. The existing data in the collection consists of the location of the abnormality (like the centre of a circle surrounding the tumour), its radius, breast position (left or right), type of breast tissues (fatty, fatty-glandular and dense) and tumour type if exists (benign or malign). All the mammograms are medio-lateral oblique view.

2.2 Pre-processing Phase

Mammograms are images difficult to interpret, and a preprocessing phase of the images is necessary to improve the quality of the images and make the feature extraction phase more reliable. Pre-processing is always a necessity whenever the data to be mined in noisy, inconsistent or incomplete and pre-processing significantly improves the effectiveness of the data mining techniques [12]. This section introduces the pre-processing techniques applied to the images before the feature extraction phase. In the digitization process, noise could be introduced that needs to be reduced by applying some image processing techniques. In addition, at the time that the mammograms were taken, the conditions of illumination are generally different.

We applied to the images two techniques: a cropping operation and an image enhancement one. The first one was employed in order to cut the black parts of the image as well as the existing artefacts such as written labels etc. For most of the images in our dataset, almost 50% of the whole image comprised of a black background with significant noise. Cropping removed the unwanted parts of the image usually periferal to the area of interest. An example of cropping that eliminates the artefacts and the black background is given in Figure 4.

The cropping to eliminate noise was done first before the image enhansement to avoid enhancing noise and hindering the cleaning phase. The cropping operation was done automatically by sweeping through the image and cutting horizontally and vertically the image those parts that had the mean less than a certain threshold.

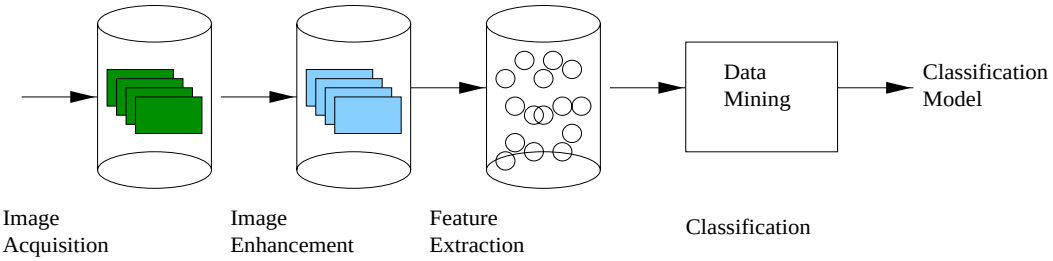


Figure 1. Image categorization process.

Image enhancement helps in qualitative improvement of the image with respect to a specific application [10]. In order to diminish the effect of over brightness or over darkness in the images and accentuate the image features, we applied a widely used technique in image processing to improve visual appearance of images known as Histogram Equalization. Histogram equalization increases the contrast range in an image by increasing the dynamic range of grey levels (or colours) [10]. This improves the distinction of features in the image. The method proceeds by widening the peaks in the image histogram and compressing the valleys. This process equalizes the illumination of the image and accentuates the features to be extracted. That is how the different illumination conditions at the scanning phase are reduced. Figure 4 shows the result of histogram equalization on the cut image.

2.3 Feature Extraction

After cropping and enhancing the images, which represents the data cleaning phase, features relevant to the classification are extracted from the cleaned images. The extracted features are organized in a database in the form of transactions, which in turn constitute the input for both classification algorithms used. The transactions are of the form $\{$ImageID, Class Label, $F_1, F_2, ..., F_n\}$ where $F_1 ... F_n$ are ne n features extracted for a given image. This database is constructed by merging some already existing features in the original database with some new visual content features that we extracted from the medical images using image-processing techniques. The existing features are:

- The type of the tissue (dense, fatty and fatty-glandular);
- The position of the breast: left or right.

The type of tissue is an important feature to be added to the feature database, being well known the fact that for some types of tissue the recognition is more difficult than for others. Training the classification systems with these features incorporated could increase the accuracy rate. The extracted features are four statistical parameters:

1. mean;
2. variance;
3. skewness and
4. kurtosis.

The general formula for the statistical parameters computed is the following:

$$M_n = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^n}{N}$$

where N is the number of data points, and n is the order of the moment.

The skewness can be defined as:

$$Sk = \frac{1}{N} \left(\frac{(x - \bar{x})}{\sigma} \right)^3$$

and the kurtosis as :

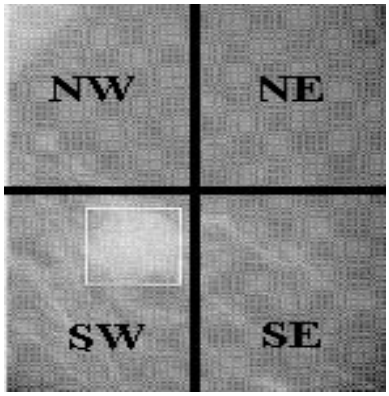
$$Kurt = \frac{1}{N} \left(\frac{(x - \bar{x})}{\sigma} \right)^4 - 3$$

All these extracted features are computed over smaller windows of the original image. The original image is first split in four parts as is shown in Figure 2. For a more accurate extraction of the features and for a further investigation of the localisation we split each of these four regions in other four parts. The statistical parameters were computed for each of the sixteen sub-parts of the original image.

3. Neural Networks

3.1 Theoretical Background

Artificial neural network models have been studied for many years in the hope of achieving human-like performance in several fields such as speech and image understanding. The networks are composed of many nonlinear computational elements operating in parallel and arranged in patterns reminiscent of biological neural networks. Computational elements or nodes are connected in



NW: North West
NE: North East
SW: South West
SE: South East

Figure 2. The four regions of the first division, and then, for each of the areas is further divided in four.

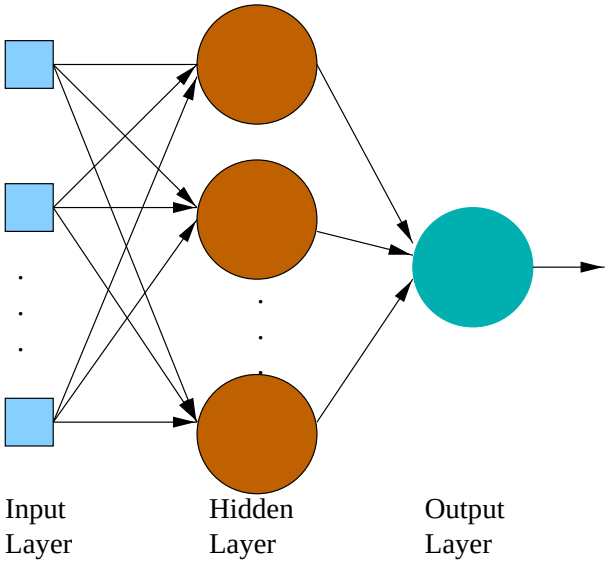


Figure 3. A 2-layer neural network.

several layers (input, hidden and output) via weights that are typically adapted during the training phase to achieve high performance. Instead of performing a set of instructions sequentially as in a Von Neumann computer, neural network models explore simultaneously many hypotheses using parallel networks composed of many computational elements connected by links with variable weights.

The back-propagation algorithm is an extension of the least mean square (LMS) algorithm that can be used to train multi-layer networks. Both LMS and back-propagation are approximate steepest descent algorithms that minimize squared error. The only difference between them is in the way in which the gradient is calculated. The back-propagation algorithm uses the chain rule in order to compute the derivatives of the squared error with respect to the weights and biases in the hidden layers. It is called back-propagation because the derivatives are computed first at the last layer of the network, and then propagated backward through the network, using the chain rule, to compute the derivatives in the hidden layers. For a multi-layer network, the output of one layer becomes the input of the following layer. A typical 2-layer neural network is depicted in Figure 3. It schematizes the neural network we used with one node in the output layer since we aimed at two class labels only.

In the following sections, we shall describe the details of our architecture.

3.2 The architecture of the neural network based system

The architecture of the neural network consists of three layers: an input layer, a hidden one and an output layer. The number of nodes in the input layer is equal to the number of elements existing in one transaction in the database. In our case, the input layer had 69 nodes. For the hidden layer, we chose 10 nodes, while the output layer was consisting of one node. The node of the output layer is the one that gives the classification for the image. It classifies it as normal or abnormal.

In the training phase, the internal weights of the neural network are adjusted according to the transactions used in the learning process. For each training transaction the neural network receives in addition the expected output. This allows the modification of the weights. In the next step, the trained neural network is used to classify new images.

4. Association Rule Mining

4.1 Theoretical Background

Association rule mining has been extensively investigated in the data mining literature. Many efficient algorithms have been proposed, the most popular being apriori [1] and FP-Tree growth [11]. Association rule mining typically aims at discovering associations between items in a transactional database. Given a set of transactions $D = \{T_1, .., T_n\}$ and a set of items $I = \{i_1, .., i_m\}$ such that any transaction T in D is a set of items in I, an association rule is an implication $A \Rightarrow B$ where the antecedent A and the consequent B are subsets of a transaction T in D, and A and B have no common items. For the association rule to be acceptable, the conditional probability of B given A has to be higher than a threshold called minimum confidence. Association rules mining is normally a two-step process, wherein the first step frequent item-sets are discovered (i.e. item-sets whose support is no less than a minimum support) and in the second step association rules are derived

from the frequent item-sets.

In our approach, we used the apriori algorithm in order to discover association rules among the features extracted from the mammography database and the category to which each mammogram belongs. We constrained the association rules to be discovered such that the antecedent of the rules is composed of a conjunction of features from the mammogram while the consequent of the rule is always the category to which the mammogram belongs. In other words, a rule would describe frequent sets of features per category normal and abnormal (benign and malign) based on the apriori association rule discovery algorithm.

After all the features are merged and put in the transactional database, the next step is applying the apriori algorithm for finding the association rules in the database constrained as described above with the antecedent being the features and the consequent being the category. Once the association rules aree found, they are used to construct a classification system that categorizes the mammograms as normal, malign or benign. The most delicate part of the classification with association rule mining is the construction of the classifier itself. Although we have the knowledge extracted from the database by finding the existing association rules, the main question is how to build a powerful classifier from these associations. The association rules that have been generated from the database in such a manner that they have as consequent a category from the classification classes. The association rules could imply either normal or abnormal. When a new image has to be classified, the categorization system returns the association rules that applies to that image. The first intuition in building the classification system is to categorize the image in the class that has the most rules that apply. This classification would work when the number of rules extracted for each class is balanced. In other cases, a further tuning of the classification system is required. The tuning of the classifier is mainly represented by finding some optimal intervals of the confidence such as both the overall recognition rate and the recognition rate of abnormal cases are at its maximum value. In dealing with medical images it is very important that the false negative rate be as low as possible. It is better to misclassify a normal image than an abnormal one. That is why in our tuning phase we take into consideration the recognition rate of abnormal images. It is not only important to recognize some images, but to be able to recognize those that are abnormal.

By applying the apriori algorithm with additional constraints on the form of the rules to be discovered we generate a relatively small set of association rules associating sets of features with class labels. These association rules constitute our classification model. The discovery of association rules in the mammogram feature database represents the training phase of our classifier. Generating the constrained association rules is very fast by comparison with training a neural network. To classify a new mammogram, it suffises to extract the features from the image as was done for the training set, and applying the association rules on the extracted features to identify the class the new mammogram falls into.

5. Experimental Results

In our experiments, we considered the 322 images from the database for both classification systems. From these set of images we considered 90 percent for training the systems and 10 percent for testing them. We considered ten splits of the data collection and computed the results for all of them in order to obtain a more accurate result of the systems' potential.

5.1 Neural Networks

The results obtained using the neural network as classifier are presented in Table 1.

On average, the classifier performed extremely well compared to other methods presented in the literature. However, the classification success ratio was not consistent among the different splits and ranged from 65.6% for split 7 to 93.7% for split 10. This inconsistency makes the method nonviable in real life applications. Nevertheless, even the lowest success rate of 65.6% can be a significant helper for a physician as an initial categorization.

Database split	Success ration (percentage)
1	96.870
2	90.620
3	90.620
4	78.125
5	81.250
6	84.375
7	65.625
8	75.000
9	56.250
10	93.750
	Average: 81.248

Table 1. Success ratios for the 10 splits with the neural network based classifier.

5.2 Association Rule Mining

As in any learning process for building a classifier, the classification performed with association rule mining comprised two steps. The first one is represented by the training of the system, while the second one deals with the classification of the new images.

In the training phase, the apriori algorithm was applied on the training data and the association rules were

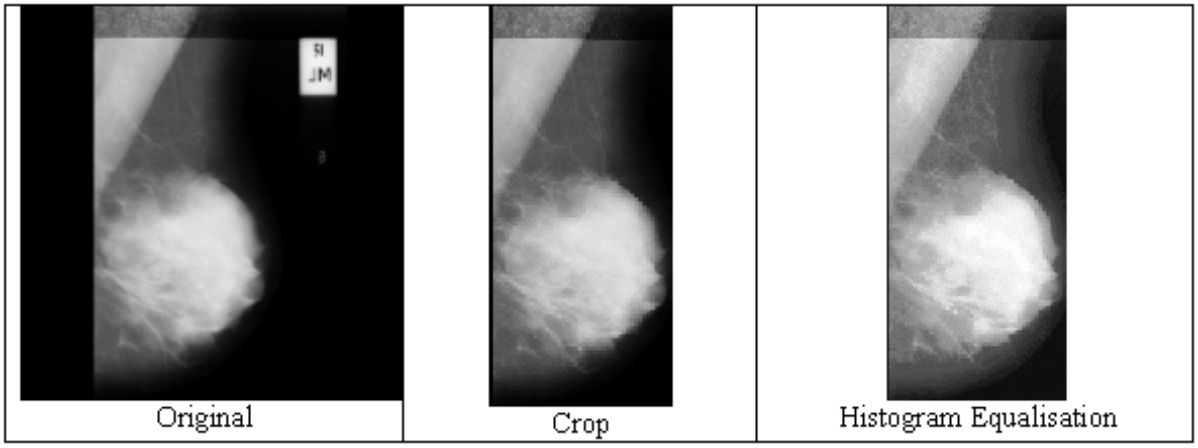


Figure 4. Pre-processing phase on an example image.

extracted. The support was set to 10% and the confidence to 0%. The reason for choosing the 0% percent for the confidence is motivated by the fact that the database has more normal cases (about 70%). The 0% confidence threshold allows us to use the confidence of the rule in the tuning phase of the classifier. In the classification phase, the low and high thresholds of confidence are set such as the maximum recognition rate is reached.

The success rate for association rule classifier was 69.11% on average. The results for the ten splits of the database are presented in Table 2. One noticeable advantage of the association rule-based classifier is the time required for training, which is very low compared to other methods such as neural networks.

Database split	Success ration (percentage)
1	67.647
2	79.412
3	67.647
4	61.765
5	64.706
6	64.706
7	64.706
8	64.706
9	67.647
10	88.235
	Average: 69.11

Table 2. Success ratios for the 10 splits with the association rule based classifier.

The recognition rate obtained using association rule mining is close to some other results reported in the literature. Another interested fact to be noticed is that the classifier proves to perform well on all the splits of the database, being more compact and consistent than the neural network classifier.

We noticed that the association rule classifier was sensitive to the unbalanced data collection that contained about 70 percent normal cases and only 30 percent abnormal, this being further divided into benign and malign. This is why we decided to build another classifier using an equilibrated distribution of normal and abnormal cases. For comparison reasons, we used a split that was also chosen in [9]. The same split is not the only reason for choosing [9] as comparison. In addition, the feature extraction phase is similar and a radial basis function network represents the classifier. We considered the 22 mammograms containing circumscribed lesions existing in the database. From these 22 mammograms, there are 18 benign and 4 malign. The abnormal mammograms are further split according to tissue type in fatty (11 cases), fatty-glandular (8 cases) and dense (3 cases). For the training procedure, we have selected 22 abnormal images and 22 normal images selected at random. For the evaluation of the results, we have used all the abnormal mammograms from MIAS database containing circumscribed masses and another 22 normal mammograms randomly selected. For this split the success rate was better (78.69%) than the previous splits. A noticeable fact is, that due to the imbalance between benign and malign images, the number of rules generated for the malign one was extremely reduced thus all the malign images being misclassified. Three out of four malign images were classified as abnormal (benign) which means that the classification in just normal and abnormal categories was actually higher (84.09%) which is a significant improvement over the results presented in [9] (75.2%). The classification results in normal and abnormal categories are presented in Figure 5.

As compared to the results presented in [9] we obtained a lower recognition rate for the fatty abnormal mammograms, but higher for all the normal cases.

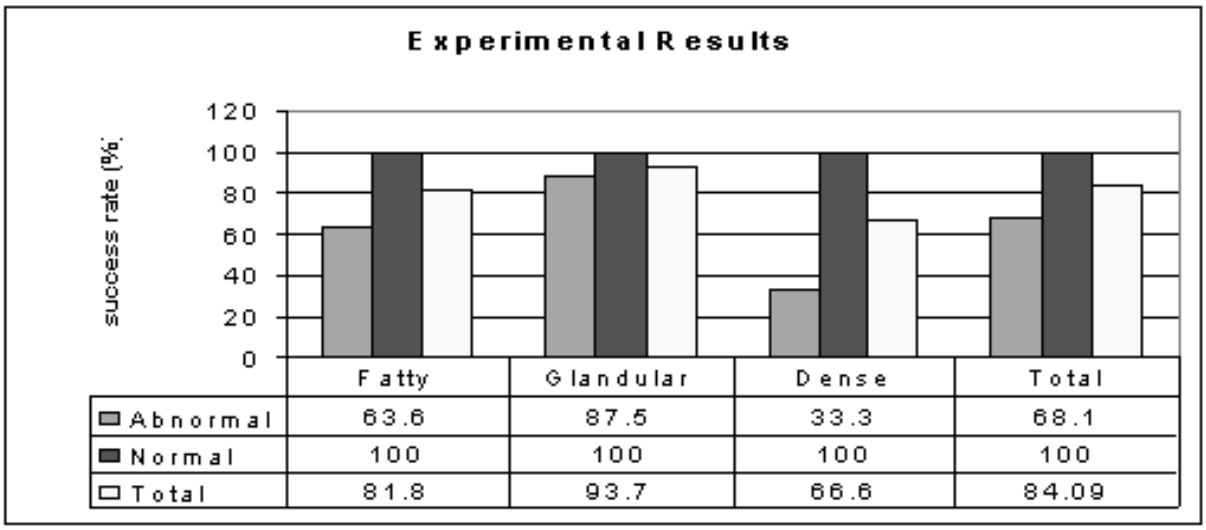


Figure 5. Success rates of association rule mining classifier.

6. Conclusions and Future Work

Mammography is one of the best methods in breast cancer detection, but in some cases, radiologists cannot detect tumours despite their experience. Such computer-aided methods like those presented in this paper could assist medical staff and improve the accuracy of detection.

In this paper, we presented two methods for tumour classification in mammograms. One system exploited the use of neural networks using back-propagation and the second one was built employing association rule mining with constraint form. The first method proved to be less sensitive to the database imbalance at a cost of high training times. The second one, with a much more rapid training phase, obtained better results than reported in literature on a well balance dataset. Both methods performed well which proves that association rules mining employed in classification process is worth further investigation.

It is well known that data mining techniques are more suitable to larger databases than the one used for these preliminary tests. We intend to use a larger mammographic database and to extract more features from the images. In particular, a classification model based on association rules becomes more accurate with a larger dataset than in the order of 300 images. In addition, more features from the database, in particular non-visual features attached to the images such as age, with/without children etc., could be interesting and relevant as additional attributes for classification. We intend to study the influence on the performance of those added features. In the case of the association rule mining approach, image split in more windows than we used could improve the detection by better localization of the cancerous tumour, thus more specific rules being extracted. For the neural network, once we use a larger database we intend to use more sophisticated neural networks in order to reduce the training times and improve accuracy. It has also been observed that the techniques employed for pre-processing the images can significantly improve or worsen the accuracy of the classifier. This was the reason our neural network performed better on the same dataset than other published research also using neural networks. We have also investigated techniques for segmenting the mammograms to determine regions of interest (not reported in this paper). Such segmentations can isolate specific regions that may be of interest to physicians [19]. We have used single link region growing algorithm [13, 4] to segment the image in regions of interest and used the features of eash region as attributes of the image. Unfortunately, this technique while reaching very encouraging accuracy with the association rule mining approach (better than the results reported in this paper) didn't perform as we hoped with the neural network approach. This, yet again, emphasized the importance of image pre-processing and the techniques used for visual feature extraction in the process of multimedia data mining. The pre-processing of mammography and the extraction of features should be dictated by rules that make sense medically. This is one of our future goals to validate the feature extraction by radiologists.

7. Acknowledgement

The authors would like to thank Veena Sridhar for her collaboration in the initial experiments of image classification using neural networks and for her contributions in the image pre-processing by segmentation (not presented in this final paper).

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, pages 207–216, Washington, D.C., May 1993.

[2] D. Brazokovic and M. Neskovic. Mammogram screening using multiresolution-based image segmentation. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(6):1437–1460, 1993.

[3] C. Chen and G. Lee. Image segmentation using multiresolution wavelet analysis and expectation-maximization (em) algorithm for digital mammography. *International Journal of Imaging Systems and Technology*, 8(5):491–504, 1997.

[4] M. L. Comer, S. Liu, and E. J. Delp. Statistical segmentation of mammograms. In *Proc. of the 3nd International Workshop on Digital Mammography*, pages 475–478, Chicago, June 9-12 1996.

[5] A. Dhawan et al. Radial-basis-function-based classification of mammographic microcalcifications using texture features. In *Proc. of the 17th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 535–536, 1995.

[6] H. Chan et al. Computerized analysis of mammographic microcalcifications in morphological and feature spaces. *Medical Physics*, 25(10):2007–2019, 1998.

[7] H. Li et al. Marcov random field for tumor detection in digital mammography. *IEEE Trans. Medical Imaging*, 14(3):565–576, 1995.

[8] H. Li et al. Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms. *IEEE Trans. Medical Imaging*, 16(6):785–798, 1997.

[9] I. Christoyianni et al. Fast detection of masses in computer-aided mammography. *IEEE Signal Processing Magazine*, pages 54–64, Jan 2000.

[10] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing, 2nd edition*. Addison-Wesley, 1993.

[11] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *ACM-SIGMOD*, Dallas, 2000.

[12] Jiawei Han and Micheline Kamber. *Data Mining, Concepts and Techniques*. Morgan Kaufmann, 2001.

[13] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison-Wesley, Reading, Mass., 1992.

[14] Breast Cancer in Men. A complete patient's guide. http://www.breastdoctor.com/breast/men/cancer.htm.

[15] Breast Cancer in Men. Male breast cancer information center. http://interact.withus.com/interact/mbc/.

[16] S. Lai, X. Li, and W. Bischof. On techniques for detecting circumscribed masses in mammograms. *IEEE Trans. Medical Imaging*, 8(4):377–386, 1989.

[17] George Wenmin Li. Classification based on multiple association rules. Master's thesis, Computing Science, Simon Fraser University, 2001.

[18] http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html.

[19] S. Singh and R. Al-Mansoori. Identification of regions of interest in digital mammograms. *Journal of Intelligent Systems*, 10(2):183–217, 2000.

[20] T. Wang and N. Karayiannis. Detection of microcalcification in digital mammograms using wavelets. *IEEE Trans. Medical Imaging*, 17(4):498–509, 1998.

[21] Osmar R. Zaïane and Maria-Luiza Antonie. Classifying text documents by associating terms with text categories. In *submitted to the Tenth International Conference on Information and Knowledge Management (CIKM 2001)*, Atlanta, GA, USA, November 2001.

Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIGKDD conference. San Francisco, USA, August 26, 2001

A Computer-aided Visual Exploration System for Knowledge Discovery from Images

Yusuke Uehara, Susumu Endo, Shuichi Shiitani, Daiki Masumoto, and Shigemi Nagata
Intelligent Systems Laboratory, Fujitsu Laboratories Ltd.
1-9-3 Nakase, Mihama-ku, Chiba-shi, Chiba 261-8588, Japan
{yuehara, endou.susumu-02, shiitani, masumoto.daiki, nagata.shigemi}@jp.fujitsu.com

ABSTRACT

Image databases contain knowledge that could be mined based on the relationships between visual feature of images and content of collateral text. For example, if popular handbags have a specific design, this knowledge can be discovered according to the relationships between the visual feature of these handbags and sales data described in text. However, in many cases, computers cannot discover such knowledge since it is often difficult to determine what type of visual feature of images are related to content of collateral text. In this paper, we propose a method to combine human cognitive abilities with the capabilities of computers for this knowledge discovery. In this method of knowledge discovery, there are two important steps. First, users view images and try to formulate hypotheses, such as association rules, about the knowledge for discovery. To aid this hypothesis formulation, a computer locates and displays images appropriately. Then, if a user can formulate a hypothesis, it should be verified. The computer can also aid users during this hypothesis verification process, but the method varies depending on the visual feature. From this point of view, we classified the verification process according to three classes of the visual feature: *auto-extractable*, *semi-auto-extractable*, and *un-extractable*. In this paper, we also present an experimental system named MIRACLES (*Multimedia Information RetrievAl, CLassification, and Exploration System*) that is designed to aid users during the hypothesis formulation process. To evaluate the capability of this system, we conducted two experiments in which we formulated association rules between visual feature of images and content of collateral text. We verified these rules by calculating the values of *support* and *confidence*. The results confirm that meaningful association rules can be discovered with this system.

Keywords

computer-aided knowledge discovery, image, data mining, visualization, visual exploration, association rule

1. INTRODUCTION

Recently, vast amounts of multimedia data have been created and stored in databases for use in many different areas, including business, science, and technology. Huge multimedia databases contain not only raw data on many types of media, but also knowledge that can be discovered from the collection of alphanumeric data, text, and images and from their relationships to one another [11][12]. For example, in handbag marketing research, it is useful to know which handbag designs are popular with teenagers. This knowledge can be extracted according to the relationship between the visual feature of handbag images and alphanumeric sales data.

As revealed in this example, the knowledge extracted from the relationships between visual feature of images and feature of other media is especially useful in areas where images play important roles. Several excellent studies and effective applications for discovering this kind of knowledge were reported. Stolorz et al. designed "CONQUEST," a system that discovers knowledge about changes in the global climate from satellite images [11]. Kakimoto et al. introduced an algorithm for discovering the relationships between active areas in a brain and their functions from f-MRI images [5]. Zaïane et al. demonstrated a system called "MultiMediaMiner," which includes functions such as characterization, classification, and association for mining in image and video databases [12].

In these studies, the type of visual feature, such as color histogram, texture, and the area of attention, could be defined for each application. Then, the value of the visual feature can be extracted from each image, and the relationship between the visual feature of images and the feature of other media can be calculated with computers. However, in many cases, it is difficult to know what type of visual feature is related to the feature of other media. For example, to discover the design of popular handbags, many aspects can be used to define feature types, such as the following: "Which feature is related to sales data: color or shape? " and "Which is more important, the shape of handbag or the shape of the pocket of handbag? ". In addition, some visual features are too complex to implement feature extraction functions on computers but are easily recognized by humans.

Langley reported on the importance of human activity in processes for discovering scientific knowledge [8]. In his report, he introduced several successful cases that combined human cognitive abilities with the capabilities of computational discovery systems. This approach is also effective for knowledge discovery from images.

To discover knowledge based on the relationships between visual feature of images and feature of other media, human users must view these images and compare their visual features. However, if images are randomly located and displayed, users cannot analyze them efficiently. To aid people in this task, computers should provide a function to locate images appropriately.

In this paper, we introduce an experimental system, MIRACLES, that provides functions to aid the human tasks for discovering knowledge. By using MIRACLES, users can discover knowledge efficiently, if a certain relationship exists between visual feature of some images and content of collateral text. For example, to discover the designs of popular handbags, it would be helpful to know that the more similar two customer profiles are, the closer together their two corresponding handbag images are. If a group of customers usually bought handbags that have a similar visual feature, users can easily find it by looking at an area in which corresponding images are located.

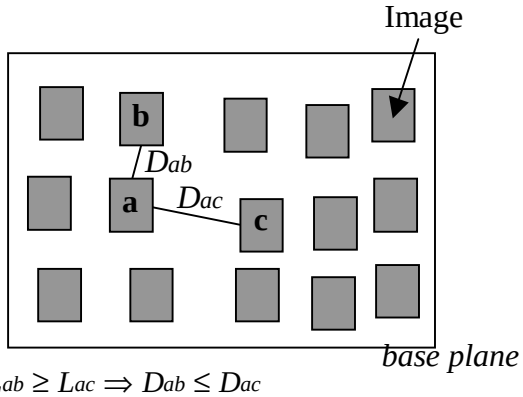
Section 2 explains the process of knowledge discovery from images. Section 3 describes the architecture and the functions of MIRACLES. Section 4 explains simple experiments to provide a detailed description of our knowledge discovery method. Lastly, Section 5 gives an example showing an application of MIRACLES.

2. PROCESS OF KNOWLEDGE DISCOVERY FROM IMAGES

In MIRACLES, the process of knowledge discovery from images consists of three steps: *data collection*, *hypothesis formulation*, and *hypothesis verification*. Figure 1 illustrates the outline of this process.

Data Collection

A large number of pairs of image and text are collected from local databases and/or the Internet as a source of knowledge. In addition to text, other types of data can be used, including alphanumeric data and voice data. In this paper, text and data of other media are called *related data* for the convenience of explanation. For example, handbag images and their *related data*, such as alphanumeric customer profile data and collateral text, are collected.



$L_{ab} \geq L_{ac} \Rightarrow D_{ab} \leq D_{ac}$

L_{ij} : Similarity between *base feature* of image i and j

D_{ij} : Distance between image i and j on the base plane

Figure 2. Method of locating on *base plane*.

Hypothesis Formulation

Users can view images to formulate hypotheses concerning the relationships between visual feature of the images and feature of *related data*. For example, an association rule [1] can be formulated as a hypothesis. For this purpose, images are located on a plane (called the "*base plane*") according to the similarity of feature (called the "*base feature*") of *related data*. Figure 2 illustrates the concept of the *base plane*. The gray rectangles indicate images. L_{ij} denotes the similarity between the *base feature* values of image i and j, and D_{ij} denotes the distance between the locations of image i and j on the *base plane*. The larger L_{ij} is, the smaller D_{ij} is.

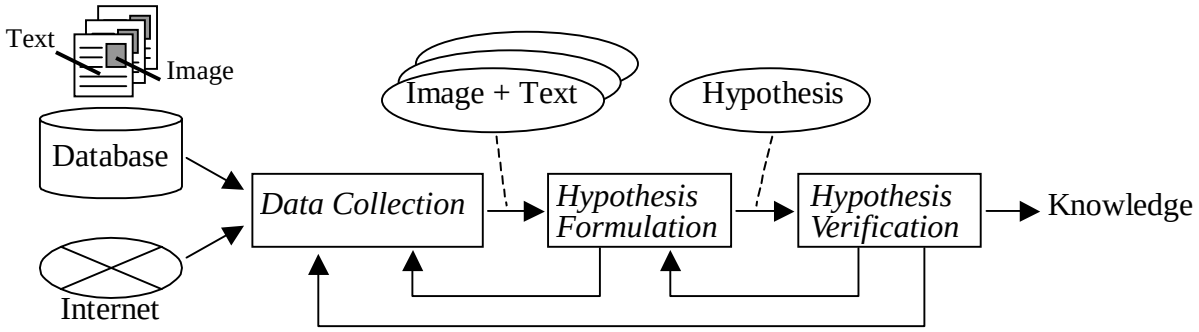


Figure 1. Process of knowledge discovery from images.

The *base feature* is represented as vectors. Thus, if the *base feature* is two-dimensional vectors, images are simply *mapped* onto the two-dimensional *base plane*. In contrast, if the *base feature* is high-dimensional vectors, images are *mapped* in a way that preserves as much as possible the topological relationship among these *base feature* values in the high-dimensional vector space as possible. As a result, if some images have a similar visual feature and *related data* also has similar *base feature*, users can easily formulate a hypothesis concerning this relationship because the similar visual feature of these images can be seen closer together in an area of the *base plane*. For example, if black handbags are popular among teenagers, users can assume this knowledge by viewing the results of *mapping* based on customer profile data as a *base feature*. It is also possible to use a pre-defined type of visual feature of images as a *base feature*. For example, if the color of handbags is used as a *base feature*, a hypothesis of "most red handbags have a rectangular shape" can be formulated.

If no hypothesis is formulated in this process, users return to the *Data Collection* process and collect data from other points of view.

Hypothesis Verification

Each hypothesis is verified by calculating some statistical parameters to define it as knowledge. In the case of association rules, *support* and *confidence* are evaluated. The method for hypothesis verification varies depending on the degree of abstraction of visual feature *types*. From this point of view, we roughly classified visual feature *types* into three classes: *auto-extractable*, *semi-auto-extractable*, and *un-extractable*. The *auto-extractable* feature corresponds directly to low-level surface features of images, such as a global color histogram. In contrast, the *un-extractable* feature corresponds to concepts, such as foods and birds. It is generally impossible to define visual feature to recognize these concepts with present pattern recognition techniques. The *semi-auto-extractable* feature also corresponds to objects, but these can be represented with a combination of low-level surface image features with present pattern recognition techniques.

With the *auto-extractable* feature, systems can automatically verify the relationship between visual feature of images and feature of *related data*. With the *semi-auto-extractable* feature, systems have difficulty extracting the visual feature from images automatically, but it is possible to extract them with a user's guidance indicating how to combine low-level surface features. For the *un-extractable* feature, it is too complex to be extracted by computers, so users have to verify it manually by, for example, checking all images on the *base plane*. Section 4 has more detailed descriptions for these three cases.

If no useful knowledge is discovered in this process, users return to the *Hypothesis Formulation* process or *Data Collection* process.

3. MIRACLES

3.1 Overview

The general architecture of MIRACLES is shown in Figure 3. The current version of MIRACLES deals with pairs of image and collateral text. In particular, primarily for convenience in conducting experiments, Collection module collects Web documents via the Internet. Users can indicate specific URLs (Uniform Resource Locators) or keywords to collect the desired Web documents. Collection module pairs each image with its collateral text in Web pages. Feature extraction module extracts feature value from image and text. Location module calculates the location of each image. Exploration module provides several functions for visual exploration on the *base plane*. GUI module provides an input/output interface for users.

3.2 Feature extraction module

Many proposals to date have attempted to represent

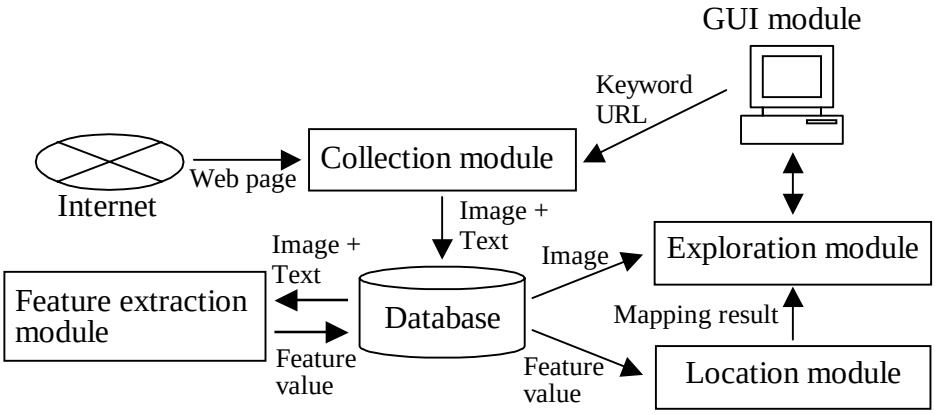


Figure 3. General architecture of MIRACLES.

feature of images and text for information retrieval. For ordinary feature *types*, we used several typical ones for *mapping* images onto the *base plane*. Brief descriptions of these feature *types* are given below.

Text feature
We used a text feature based on the basic *vector space model* [10]. A content of text is represented as a vector in which every component corresponds to the frequency of occurrence for a particular word in text. In addition, a weight is assigned to each component. This weight represents the importance of the word.

Color feature
A global color histogram based on the HSV color space is used as a color feature. Initially, the HSV color space is divided into eight hues, four saturations, four values, and four gray levels. This division results in the generation of 132 subspaces in the HSV space. The number of pixels in each subspace is then counted as a component of the histogram.

Shape feature
We used coefficients of the Mallat transformation, which is a kind of Wavelet transformation [2], as the shape feature. These coefficients capture the directions and spatial distributions of a sharp edge by high-frequency components and those of a dull edge by low-frequency components. In our implementation, low-frequency components are used because of their robustness for object displacement in images [9]. This feature can roughly capture the shape of an object in an image if the background color and texture are uniform in all images.

Texture feature
We also used coefficients of the Mallat transformation as the texture feature. When both of low-frequency components and high-frequency components are used, these coefficients capture the texture feature of images [9].

Feature extraction module extracts these types of feature values as vectors and saves them in a database.

3.3 Location module
With Location module, images are *mapped* according to the similarity of the *base feature*, as discussed in Section 3.2. Before *mapping*, a user selects text, color, shape, or texture as a feature type. Then the similarity between two feature vectors is defined using the Euclidean distance between them. Thus, the smaller the Euclidean distance between the *base feature* vectors of two images, the closer these two images are on the *base plane*.

Implementation of this *mapping* function is based on the *Self-Organizing Map* (SOM) algorithm [3][4][6][7]. SOM is a neural network algorithm based on unsupervised learning. SOM can form illustrative two-dimensional projections of data distributions in a high-dimensional data space. Thus, SOM can approximate the mapping described in Section 2(Figure.2). In our implementation, the rectangular grid SOM is applied to the *base feature* vectors of images, and images are positioned at cells corresponding to the *base feature* vectors.

3.4 Exploration module
To explore *mapping* results visually, Exploration module provides a fly-through function. With this function, users can move and zoom in/out of images on the *base plane*. In addition to the fly-through function, MIRACLES provides several exploration functions that are described in Section 4.

4. KNOWLEDGE DISCOVERY WITH MIRACLES
This section demonstrates a process of knowledge discovery with MIRACLES by using a simple example. In addition, it shows an evaluation of the association rules discovered in this process. Since it is difficult to evaluate performance directly and quantitatively, we provide suggested performance values using the results obtained from examples.

As a data set, we collected Web documents consisting of images of national flags and collateral text that describes about each nation, with information such as its region in the world, population, language, and religion.

First, we tried to formulate a hypothesis that is expressed by the following association rules [1]:

$$H(n, f) \Rightarrow L(n, r), \qquad (1)$$

where $H(n, f)$ is a flag of nation n having visual feature f. If f is *auto-extractable*, it can be represented with visual feature that can be extracted by Feature extraction module, namely the color histogram or the Mallat transformation coefficients. If f is *semi-auto-extractable*, it denotes a concept initially. But it can be represented by the combination of implemented visual features. If f is *un-extractable*, it denotes a concept and cannot be extracted by Feature extraction module. $L(n, r)$ means that nation n has attribute r, such as its region in the world and religion. For this purpose, we *mapped* these flag images based on the region of each nation described in collateral text. Figure 4 shows the results of this *mapping*.

Subsequently, if a hypothesis could be found, we verified it by using *support* and *confidence*, which can measure the importance of this knowledge. The definition of *support* and *confidence* is as follows:

$$support = \frac{N(H(n, f) \wedge L(n, r))}{NT} \qquad (2)$$

Figure 4. Classification of national flag images.

$$confidence = \frac{N(H(n, f) \wedge L(n,r))}{N(H(n, f))}$$

where $N(P)$ is the number of data items that satisfy proposition P, and NT is the total number of data items. This verification is achieved manually by checking and counting images that are matched with $N(H(n, f))$ and $N(H(n, f) \wedge L(n,r))$.

As described in Section 2, visual features are categorized into three types, *auto-extractable*, *semi-auto-extractable*, and *un-extractable*, in terms of the characteristics of the visual features. The following three examples correspond to each type of visual feature.

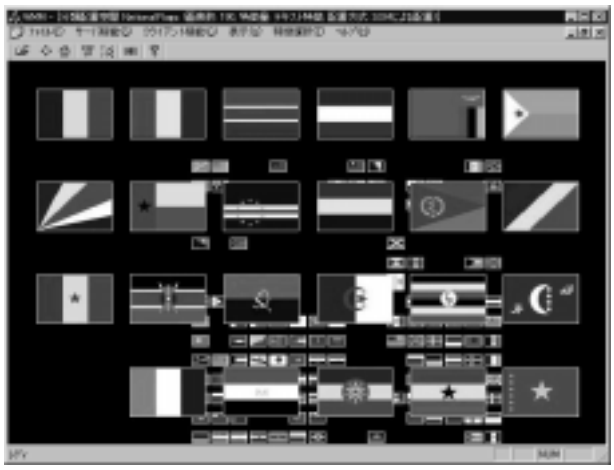


Figure 6. Flags of African nations.

Example 1

By viewing entire images on the *base plane*, we found that many flags in the area of the upper-half of the *base plane* have both green and yellow. This area corresponds to nations on the African continent. In such case, it is usually necessary to compare features of images in this area with features of images in another area. To help with this comparison, MIRACLES has the *pop-up* function illustrated in Figure 5. With this function, users can retrieve desired images from among displayed images. Then, the retrieved images are moved from the *base plane* to a location closer to the user. With this function, users can focus on their desired images and compare them with other images at the same time. Figure 6 shows a snapshot in which the flags of African nations are retrieved.

From this exploration, the following hypothesis can be formulated:

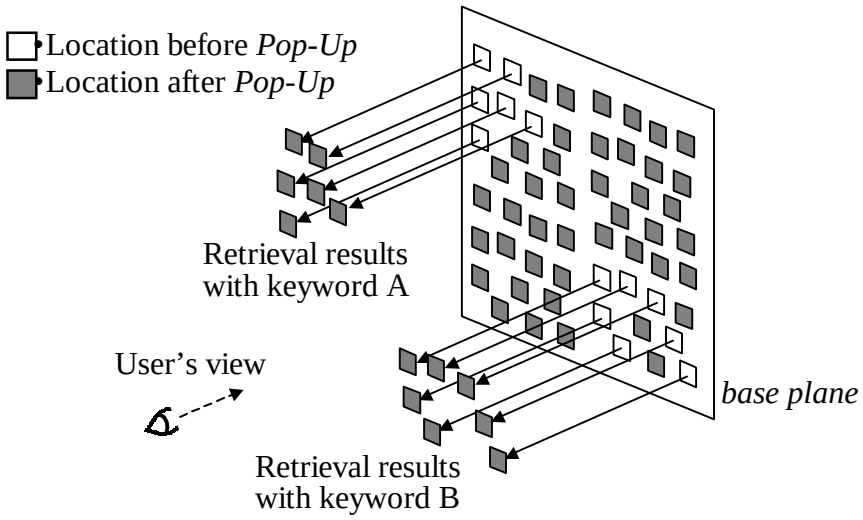


Figure 5. *Pop-up* function from *base plane*.

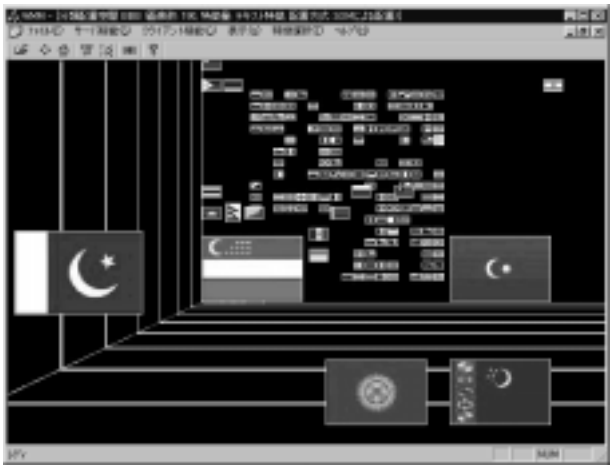


Figure 7. Flags of Asian nations where many Muslims live.

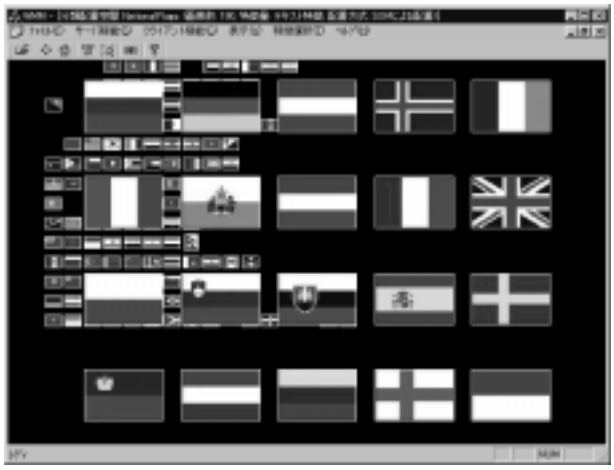


Figure 8. Flags of European nations.

$$H(n, f_{green\ \&\ yellow}) \Rightarrow L(n, r_{africa}), \qquad (3)$$

where $f_{green\ \&\ yellow}$ denotes that a flag has both green and yellow, and r_{africa} denotes that a nation is located on the African continent. Next, we verified this hypothesis by checking all images. As a result, *support* was found to be 0.10 and *confidence* was 0.68.

In this example, a visual feature ($f_{green\ \&\ yellow}$) can be represented by a global color histogram. Namely, this is an *auto-extractable* feature. MIRACLES can also has a function for extracting this feature from images. Thus, it will be possible to verify the hypothesis automatically.

Example 2

The next example concerns a visual feature that is categorized as a *semi-auto-extractable* feature. At first, we noticed that many flags have illustrations of the moon and stars in the lower area of the *base plane*, an area corresponding to Asian nations. Therefore, we concentrated these flags into one area by using the *pop-up* function with the keyword "Asia." We then noticed that several nations that are populated by many Muslims have flags designed with the moon and stars. Therefore, we used the *pop-up* function with the keyword "Muslims" on the previous *pop-up* results, as shown in Figure 7. This example shows that *Hypothesis Formulation* can be done step-by-step visually. As a result, we formulated the following hypothesis:

$$H(n, f_{moon\ \&\ star}) \Rightarrow L(n, r_{asia\ \&\ muslim}), \qquad (4)$$

where $f_{moon\ \&\ star}$ denotes that a flag has an illustration of the moon and stars, and $r_{asia\ \&\ muslim}$ denotes that a nation is located in Asia and many Muslims live in this nation. We verified this hypothesis in the same way as in Example1. As a result, *support* was found to be 0.03 and *confidence* was 0.70.

It is difficult for the system to find that illustrations of the moon and stars are useful visual feature. Thus, automatic verification on the system is difficult to accomplish. However, the feature of an illustration of the moon and stars is relatively simple. A feature extraction function that finds them can be developed with present pattern recognition techniques. As shown in this example, hypotheses about this type of visual feature can be verified automatically if users give appropriate guidance according to the respective hypotheses.

Example 3

After looking at the European nations' flags located in the lower-right area of the *base plane*, we formulated the following hypothesis:

$$H(n, f_{crest}) \Rightarrow L(n, r_{europe}), \qquad (5)$$

where f_{crest} denotes that a crest is drawn on a flag, and r_{europe} denotes that the nation is located in Europe. Figure 8 shows some of these flags. We verified this hypothesis in the same way as in Example 1. As a result, *support* was found to be 0.06 and *confidence* was 0.41.

Even though human users can easily notice that an important visual feature is a crest, it is almost impossible for the system to find it without knowledge of the crest. Moreover, the crest is too complex to develop a feature extraction module that models the visual feature of the crest. Therefore, users have to verify the hypothesis manually if the visual feature is *un-extractable*.

5. EXPERIMENTS

We conducted more practical experiments than the ones described in Section 4. We used Web documents that deal with traditional craft, especially chinaware, as a

topic. These Web documents are composed of images

Figure 9. Chinaware images.

illustrating the appearance of chinaware and collateral text describing both attribute and history of the chinaware.

First, images were *mapped* and displayed according to the similarity among features of the collateral text. Then, in an exploration process, we found several images of chinaware that are located in a certain local area and have the same visual feature, and we found that certain objects are painted with blue on a white background, as shown on the right side of Figure 9. This area of the *base plane* corresponds to the text about "Sometsuke," which is a special technique for making chinaware. We thus formulated the hypothesis described by formula (1), but here $H(n, f)$ means that chinaware n is white and something is drawn in blue on it, and having visual feature f and $L(n, r)$ means that chinaware n is made with the "Sometsuke" technique.

To verify this hypothesis, we counted the number of images and calculated *support* and *confidence*. As a result, *support* was found to be 0.07 and *confidence* was 0.82. This example shows that we can formulate a hypothesis that has high *confidence* in this experiment. The visual feature in this experiment can be extracted by using a global color histogram and texture features where users should indicate appropriate weights for each feature. Therefore, these are *semi-auto-extractable* features.

6. CONCLUSIONS

In this paper, we have presented a method for knowledge discovery from images. This method consists of three steps: *Data Collection*, *Hypothesis Formulation*, and *Hypothesis Verification*. In *Hypothesis Formulation*, users have to formulate a hypothesis for the relationships between visual feature of images and feature of *related data*. We have developed an experimental system named MIRACLES that facilitates the human tasks in this process. With the *mapping* function, images are grouped and displayed in a local area, so users can easily look for a visual feature of images that are related to similar topics. With the visual exploration function, users view *mapping* results by visualizing the images in a 3-D space. Users can move the display focus back-and-forth from a global view of entire images to a local view of a specific image. In the *Hypothesis Formulation* process, it is very useful that users can view a huge number of images from multiple perspectives.

Moreover, we indicated that the *Hypothesis Verification* process is categorized into three types according to the difficulty of extracting visual feature from images, namely cases involving the *auto-extractable*, *semi-auto-extractable*, and *un-extractable* feature. In the first case, hypotheses can be verified automatically. In the second case, hypotheses can also be verified if users give appropriate instructions. In the last case, however, hypotheses cannot be verified automatically.

We showed the concept of auto verification for *auto-extractable* and *semi-auto-extractable* visual feature, but the current version of MIRACLES does not yet have this function. We plan to implement this function and evaluate it as future work.

7. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases, In *Proc. of ACM SIGMOD Conference on Management of Data*, pp.207-216, 1993.

[2] I. Daubechies, Ten Lectures on Wavelets, Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia, 1992.

[3] G. Deboeck and T. Kohonen. Visual explorations in finance using Self-Organizing Maps, Springer-Verlag, London, 1998.

[4] T. Honkela, S. Kaski, K Lagus, and T. Kohonen. WEBSOM-self-organizing maps of document collections, In *Proc. of Workshop on Self-Organizing Maps 1997(WSOM'97)*, pp.310-315, 1997.

[5] M. Kakimoto, C. Morita, and H. Tsukimoto. Data mining from functional brain images. In *Proc. of ACM MDM/KDD2000*, pp.91-97, 2000.

[6] T. Kohonen. Self-Organizing Maps. Springer-Verlag, Berlin, Heidelberg, 1995.

[7] J. C. Lamirel, J. Ducloy, and H. Kammoun. A self organizing map (SOM) extended model for information discovery in a digital library context. In *Proc. of ACM MDM/KDD2000*, pp.60-66.

[8] P. Langley. The computer-aided discovery of scientific knowledge. In *Proc. of the First International Conference on Discovery Science*. Fukuoka, Japan: Springer, 1998.

[9] K. Murao and A. Ando. Content-based image retrieval system. In *Proc. of the 1998 Information and Systems Society Conference of IEICE*, D-11-60, p.175. (in Japanese)

[10] G. Salton and M. J. McGill. Introduction to modern

information retrieval, *McGraw-Hill*, NewYork, 1983.
[11] P. Stolorz, H. Nakamura, E. Mesrobian, R. Muntz, E. Shek, J. Santos, J. Yi, K. Ng, S. Chien, C. Mechoso, and J. Farrara. Fast spatio-temporal data mining of large geophysical datasets. In *Proc. Int. Conf. on KDD*, pp.300-305, 1995.
[12] O. R. Zaïane, J. Han, Z. -N. Li, and J. Hou. Mining Multimedia Data. In *Proc. CASCON'98: Meeting of Minds*, Toronto, Canada, November, pp.83-96, 1998.

Author Index

-- NOTES --

-- NOTES --

-- NOTES --

-- NOTES --

-- NOTES --

-- NOTES --

-- NOTES --

-- NOTES --