LMS -- A Long term knowledge-based multimedia retrieval system for region-based image databases

Xin Chen and Chengcui Zhang*
Department of Computer and Information Sciences,
The University of Alabama at Birmingham,
CH 127 1530 3rd Ave S, Birmingham, AL 35294, USA
Email: chenxin@cis.uab.edu Email: zhang@cis.uab.edu
*Corresponding author

Shu-Ching Chen and Min Chen
School of Computing and Information Sciences
Florida International University
Miami, FL 33199, USA
Email: chens@cs.fiu.edu Email: mchen005@cs.fiu.edu

Abstract: In knowledge-based systems, human interaction usually refers to "expert knowledge". However, in a large system where no pre-defined knowledge from expert is available, we may learn from the users of the system, i.e. through users' queries and their feedbacks on the query results. The Content-Based Image Retrieval (CBIR) system is a special kind of knowledge-based multimedia retrieval system. In CBIR, there is a widely used technique for incorporating the user's knowledge with the learning process called Relevance Feedback (RF). As a supervised learning technique, RF has been shown to significantly increase the retrieval accuracy. However, as a CBIR system continues to receive user queries and user feedbacks, the information of user preferences across query sessions are often lost at the end of search, thus requiring the feedback process to restart for each new query. A few works targeting long-term learning have been done in general CBIR domain to alleviate this problem. However, none of them address the needs and long-term similarity learning techniques for region-based image retrieval. This paper proposes a long-term knowledge-based multimedia retrieval system based on Latent Semantic Indexing (LSI) and human interaction (Relevance Feedback). The proposed knowledge-based system for image region retrieval, LMS (Long Term Knowledge-based Multimedia Retrieval System), is constructed on a Multiple Instance Learning (MIL) framework with One-class Support Vector Machine (SVM) as its core. Experiments show that the proposed system can better utilize users' feedbacks of previous sessions, thus improving the performance of the learning algorithm.

Keywords: Human-centered Multimedia Retrieval System, Knowledge-based System, Region-based Image Retrieval, Latent Semantic Indexing, One-Class Support Vector Machine, Relevance Feedback.

Reference to this paper should be made as follows: Chen, X., Zhang, C., Chen, S.-C. and Chen, M. 'LMS – A Long term knowledge-based multimedia retrieval system for region-based image databases', *Int. J.,* Vol. X, No. Y, pp. 000-000.

Biographical notes:
Xin Chen received her Master's degree in Computer Science from University of Science and Technology Beijing, China, in 2002. From 2004 to present, she has been pursuing her Ph.D. degree in the Computer and Information Sciences Department at the University of Alabama at Birmingham. Her research interests include Content-based Image Retrieval, multimedia data mining, and spatio-temporal data mining.

Chengcui Zhang is an Assistant Professor of Computer and Information Science at the University of Alabama at Birmingham (UAB) since August, 2004. She received her Ph.D. from the School of Computer Science at Florida International University, Miami, FL, USA in August, 2004. She also received her bachelor and master degrees in Computer Science from Zhejiang University in China. Her research interests include multimedia databases, multimedia data mining, image and video database retrieval, bioinformatics, and GIS data filtering. She is the recipient of several awards, including the IBM Unstructured Information Management Architecture (UIMA) Innovation Award, UAB ADVANCE Junior Faculty Research Award from the National Science Foundation, UAB Faculty Development Award, and the Presidential Fellowship and the Best Graduate Student Research Award at FIU.

Shu-Ching Chen received his Ph.D. from the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA in December, 1998. He also received Masters degrees in Computer Science, Electrical Engineering, and Civil Engineering from Purdue University, West Lafayette, IN, USA. He has been an Associate Professor in the School of Computer Science (SCS), Florida International University (FIU) since August, 2004. Prior to that, he was an Assistant Professor in SCS at FIU dating from August, 1999. His main research interests include distributed multimedia database systems, data mining, and multimedia networking. Dr. Chen has authored and co-authored more than 130 research papers in journals, refereed conference/symposium/workshop proceedings, and book chapters. He was awarded University Outstanding Faculty Research Award from FIU in 2004. He also received Outstanding Faculty Research Award from SCS at FIU in 2002. He is the general co-chair of the IEEE International Conference on Information Reuse and Integration and program chair of several conferences.

Min Chen received her bachelor's degree in Electrical Engineering from Zhejiang University in China. She is currently a Ph.D. candidate in the School of Computing and Information Sciences at Florida International University (FIU). Her research interests include distributed multimedia database systems, image and video database retrieval, and multimedia data mining. She has authored and co-authored 17 technical papers published in various prestigious journals, refereed conference/workshop proceedings and book chapters. She is the recipient of several awards, including a Presidential Fellowship and the Best Graduate Student Research Award from FIU.

1. Introduction

While multimedia data becomes increasingly available, there is a need for an effective and efficient content retrieval system which integrates multimedia content analysis, information theory, artificial intelligence, and database technology. However, the conventional passive interaction mode, in which the search engine responds to a user's query according to some fixed search mechanism, is not sufficient to meet the user's perceptual needs on multimedia data. In a human-centered approach to multimedia system design, the retrieval and search engine must accommodate human perceptual and response capabilities and limitations. Several efforts are underway to develop interaction models that are more natural to everyday human experience. Recently, Relevance Feedback (RF) technique has been used to incorporate the user's knowledge with the learning process for Content-based Image Retrieval Systems (CBIR), in which the user is allowed to evaluate the quality of the returned results by providing 'positive' or 'negative' feedback. Specifically, in Content-Based Image Retrieval Systems, Relevance Feedback is a commonly used technique to overcome the "semantic gap" between high level user concepts and low level image features. In this paper, we explore the influence

of human interactions (Relevance Feedback) to the performance of knowledge based multimedia retrieval system in the long run, and apply the proposed platform to region-based Content-Based Image Retrieval. To our best knowledge, this is the first in the literature that studies the long-term knowledge discovery and utilization from user interactions for region-based image retrieval systems. The interactions between users and systems are important in understanding the big picture of a human-centered design for multimedia retrieval systems.

Most of the existing Relevance Feedback (RF) based approaches {Rui, Huang and Mehrotra, 1997; Su, Zhang, Li and Ma, 2003} consider each image as a whole, which is represented by a vector of N dimensional image features. However, the user's query interest is often focused on certain part of the query image i.e. a region in the image that has an obvious semantic meaning. Therefore, rather than viewing each image as a whole, it is more reasonable to view it as a set of semantic regions. In this context, the goal of image retrieval is to find the semantic region(s) of the user's interest. Since each image is composed of several regions and each region can be considered as an instance, region-based CBIR is then transformed into a Multiple Instance Learning (MIL) problem {Maron and Lozano-Perez, 1998}. Maron et al. applied MIL to natural scene image classification {Maron and Lozano-Perez, 1998}. Each image is viewed as a bag of semantic regions (instances). In the scenario of MIL, the labels of individual instances in the training data are not available. Instead, the bags are labeled. When applied to RF-based CBIR, this corresponds to the scenario that the user gives feedback on the whole image (bag) although he/she may be only interested in a specific region (instance) of that image. The goal of MIL is to obtain a hypothesis from the training examples that generates labels for unseen bags (images) based on the user's interest on a specific region.

We addressed the above mentioned problem using One-class Support Vector Machine {Schölkopf, Platt, et al., 1999} and built up a learning and retrieval framework in our previous work {Zhang, Chen, Chen, Chen and Shyu, 2005}. This framework applies MIL to learn the region of interest from users' relevance feedback on the whole image and tells the system to shift its focus of attention to that region. In particular, the learning algorithm concentrates on those positive bags (images) and uses the learned region-of-interest to evaluate all the other images in the image database. The choice of **One-class** Support Vector Machine is based on the observation that *positive* images are *positive* in the same way while *negative* images are *negative* in their own way. In other words, instead of building models for both positive class and negative class, it makes more sense to assume that all positive regions are in one class while the negative regions are outliers of the positive class. Therefore, we concentrate on *positive* image regions and try to model them by utilizing the extract knowledge from user interaction within our framework.

Chen et al. {Chen, Zhou, Tomas and Huang, 2001} and Gondra {Gondra and Heisterkamp, 2004} both use One-Class SVM in image retrieval, but it is applied to the image as a whole. In our system, One-Class SVM is used to model the non-linear distribution of image regions and separate positive regions from negative ones. Each region of the test images is given a score by the evaluation function built from the model. The higher the score, the more similar it is to the region of interest. The images with the highest scores are returned to the user as query results. Our comparative study shows the effectiveness of this framework with a high retrieval accuracy being achieved on average within 4 iterations {Zhang, Chen, Chen, Chen and Shyu, 2005}.

In our experiments, we also observed that it is highly possible that repetitive or similar queries are issued by different users. However, the learning mechanism ignores the previously acquired knowledge from users' relevance feedback and treats each query

as an independent and brand-new one. This raised a question i.e. how can we make full use of the relevance feedback information collected across query sessions? In this paper, we explore a Latent Semantic Indexing based method to analyze and extract useful knowledge from feedbacks stored in database access logs. Related work on using database log information can be found in {Hoi and Lyu, 2004; Shyu, Chen, Chen and Zhang, 2004; Fournier and Cord, 2007; Zhou, Zhang, Liu, Zhang and Shi, 2003; Lee, Ma and Zhang, 1999; Li, Chen and Zhang, 2002; He, King, Ma, Li and Zhang, 2003; Han, Ngan, Li and Zhang, 2005}. However, again their works are based on the whole image instead of image regions. Long-term learning techniques which try to propagate the feedback information across query sessions for region-based image retrieval still remain an open issue. The information discovery strategy adopted in this paper uses Latent Semantic Indexing in analyzing database log information. In this way, data dimensions are reduced and become more compact with only important information retained and noise removed. This differentiates our method from the methods proposed in {Hoi and Lyu, 2004; Shyu, Chen, Chen and Zhang, 2004}, in which log information is used in a more straightforward way. Our experiments demonstrate that the proposed method performs better in extracting useful knowledge from database access logs.

Section 2 reviews related work. Section 3 illustrates the mapping between region-based image retrieval and Multiple Instance Learning. Section 4 introduces the proposed learning mechanism based on One-class Support Vector Machine. In Section 5, the detailed learning and retrieval approach based on One-class Support Vector Machine is discussed. In Section 6, we propose to use Latent Semantic Indexing to extract information from database access log. The overall architecture of LMS is illustrated and the experimental results are presented in Section 7. Section 8 concludes the paper.

2. Related work

2.1. Multiple Instance Learning

As aforementioned, the system proposed in this paper is region-based. Instead of viewing each image as a whole, we treat each image as a set of regions with obvious semantic meanings. The user specifies a region of interest as the query region. The ultimate goal is then to find the images in the database which have at least one semantic region that matches the user's query region. This scenario is mapped to a Multiple Instance Learning (MIL) problem in this paper.

The Multiple Instance Learning problem (MIL) is a special kind of supervised machine learning problem, which has recently received a great amount of attention from the computational intelligence community. In the standard supervised machine learning methodology, each object in the set of training examples is labeled and the problem is to learn a hypothesis that can accurately predict the labels of the unseen objects. However, in MIL problem, the labeling information is incomplete. "Bag" is a term used in MIL to denote a set of labeled objects. "Object" corresponds to the term "instance". Therefore, in MIL, a training example is a labeled bag and the labels of the instances are unknown; although, each instance is actually associated with a label. The goal of learning is to obtain a hypothesis from the training examples that generate labels for the unseen bags and instances. There are two kinds of labels in MIL, namely Positive and Negative. A bag is labeled Positive if and only if the bag has one or more Positive instances and is labeled Negative if and only if all its instances are Negative.

Originally, MIL is applied to such applications as drug activity prediction and stock prediction. Recently, it has gained its popularity in natural scene image classification and Content-based Image Retrieval. A natural image scene usually contains a lot of different semantic regions and its semantic category is usually determined by one or more regions in the image. There might be some regions that do not fit into the semantic category of that image. Each image corresponds to a bag and the regions in that image correspond to the instances in that bag. An image is labeled Positive if it contains the concept of a specific semantic category; otherwise, it is labeled Negative. Multiple Instance Learning (MIL) can discover the regions that are actually related to the user concept. By filtering out the unrelated regions (which can be considered as "noise") and only considering the related regions in the query process, we can expect a better query performance.

There is a plethora of research done in MIL. A representative approach by learning the axis-parallel rectangles (APR) is first developed by Dietterich et al. {Dietterich, Lathrop and Lozano-Perez, 1997}. A MULTINST algorithm which is also an APR-based method for Multiple Instance Learning was proposed in {Auer, 1997}. Maron et al. {Maron and Lozano-Perez, 1998} applied Multiple Instance Learning to natural scene image classification. The concept of Diverse Density (DD) is introduced in their approach and a two-step gradient descent with multiple starting points is applied to find the maximum Diverse Density. The EM-DD algorithm is proposed by Zhang and Goldman in {Zhang and Goldman, 2002} based on Diverse Density. Its main difference from Maron's method is that it searches the maximum DD points by using Expectation Maximization. Their algorithm was predicated on the assumption that each bag has a representative instance that was treated as a missed value and then the EM (Expectation-Maximization) method and Quasi-Newton method were used to simultaneously learn the representative instances and maximize the Diversity Density. It is shown that EM-DD is more robust in dealing with high-dimensional data. Ray and Page {Ray and Page, 2001} also used the EM method for Multiple Instance Regression. Wang et al. {Wang and Zucker, 2000} explored the lazy learning approaches in Multiple Instance Learning. Zucker et al. developed two kNN-based algorithms, Citation-kNN and Bayesian-kNN, for Multiple Instance Learning. In {Zucker and Chevaleyre, 2001}, the authors attempted to solve the Multiple Instance Learning problem with decision trees and decision rules. Ramon et al. {Ramon and De Raedt, 2000} propose the Multiple Instance Neural Network. Andrews et al. use Support Vector Machines (SVMs) to solve MIL problem. Their method is called MI-SVM {Andrews, Tsochantaridis and Hofmann, 2003}. Chen et al. {Chen, Rubin, Shyu and Zhang, 2006} adopt Neural Network based learning algorithm. Zhang et al.'s EM-DD method {Zhang and Goldman, 2002} is used by Chen et al. {Chen and Wang, 2004} for image categorization, in which Chen et al. apply standard SVMs to solve a two-class classification problem. Their approach is called DD-SVM which is a region-based image retrieval algorithm. It is claimed in Chen et al.'s work that their approach is different from {Andrews, Tsochantaridis and Hofmann, 2003}'s MI-SVM in that DD-SVM defines several features for each bag according to instance prototypes (generated by DD) while MI-SVM only selects one instance to represent the entire positive bag.

Our proposed learning algorithm concentrates on those positive images and uses the learned region-of-interest to evaluate all the other images in the image database. For this purpose, we applied One-Class Support Vector Machine (SVM) {Schölkopf, Platt, et al., 1999} to solve the MIL problem in CBIR. Chen et al. {Chen, Zhou, Tomas and Huang, 2001} and Gondra et al. {Gondra and Heisterkamp, 2004} also use One-Class SVM in image retrieval. However, it is applied to the image as a whole. An example of region-based image retrieval using One-Class SVM is Jing et al.'s work in {Jing, Li, Zhang,

Zhang and Zhang, 2003}. However, One-class SVM is only used as a distribution estimator and the classification is done by a two-class SVM.

2.2. Long-term learning

The concept of relevance feedback (RF) associated with CBIR is first proposed in {Rui, Huang and Mehrotra, 1997}. It is widely used to incorporate users' subjective concepts with the learning process. Relevance feedback is an interactive process by which the user judges the quality of the retrieval performed by the system by explicitly marking those images that the user perceives as truly relevant among the images retrieved by the system. This information is then used to refine the original query. The process iterates until a satisfactory result is obtained for the user. However, RF often suffers from lack of training samples. In a typical setting for RF based CBIR system, relevance feedback is conducted independently by each user in each query. Query results are not stored in database for further usage. However, it is a waste of resource by not taking into account the query history since these information may help improve the query accuracy immediately. In our proposed system, this issue is addressed by mining information from the query history.

There are some related works in long-term learning in CBIR. Lee et al. {Lee, Ma and Zhang, 1999} proposed a mechanism to gradually integrate relevance feedback information to the image retrieval system which will improve its performance over time. The system is based on clustering. A correlation matrix is constructed according to users' feedback which guides cluster re-adjustment by splitting/merging. In this way, starting with low-level features, high level semantics are gradually embedded into the system through the process of updating the correlation matrix by Radial Basis Function (RBF) transform. In Fournier and Cord's {Fournier and Cord, 2002} work, a long-term similarity learning model is used to refine the CBIR system. The system is based on image histogram similarities. User-provided information is an added factor in the long-term similarity measure. Therefore, the relation between two images depends on both the static similarity measure and the dynamic long-term similarity measure. Li et. al. {Li, Chen and Zhang, 2002} analyzes statistical correlations among images through unigram and bigram frequency. Bigram frequency is the number of times two images are co-relevant according to the user's relevance feedback and unigram frequency is the number of times an image is relevant. The retrieved images are then reordered based on the proposed correlation model. He et. al. {He, King, Ma, Li and Zhang, 2003} tried to infer a semantic space through the user's query. The semantic space is constructed by a semantic matrix (rows corresponds to all images and columns corresponds to image features), query image vectors and result vectors which is the multiplication of the first two. Singular Value Decomposition (SVD) is adopted to reduce the size of the semantic space. In {Zhou, Zhang, Liu, Zhang and Shi, 2003}, a correlation matrix is constructed such that rows are the queries and columns are the images. The proposed work applies Edit Distance to evaluate the similarity between feedbacks of current retrieval and prefixes of records in database log. This is integrated with Euclidean distance based similarity measurement. In Hoi et. al.'s work {Hoi and Lyu, 2004}, the correlation matrix is similar to that in {Zhou, Zhang, Liu, Zhang and Shi, 2003} except that it is its transpose and is called relevance matrix. A standard Support Vector Machine (SVM) is proposed to make use of the relevance matrix and classify the images into two classes – positive and negative. In {Shyu, Chen, Chen and Zhang, 2004}, based on database access log, the affinity relations among databases as well as among images within one database

are measured. For this purpose, local and integrated Markov Model Mediators are constructed, respectively. However, since this work focuses on long-term learning, it provides little support for instant learning. In {Han, Ngan, Li and Zhang, 2005}, Han et. al. proposed a knowledge memory model by accumulating user-provided feedbacks. Images are grouped into semantically correlated clusters based on the memorized correlations among images. Hidden semantic correlation between images and that between an image and the feedback examples are analyzed in this model. SVM is trained to learn these correlations. At the end, an annotation propagation scheme is proposed using both memorized and learned semantics.

All these previous long-term learning works in CBIR tries to explore the relationships among images with the aid of query history. Most of them either implicitly or explicitly use the concept of correlation matrix in their analysis. It is the method used to construct such matrix that varies. The proposed system is basically different from the others in that the proposed system is region based while the aforementioned works are all based on the whole image. In addition, we exploit Latent Semantic Indexing and One-Class Support Vector Machine for both short-term (instant) learning and long-term learning.

3. Multiple Instance Learning

In traditional supervised learning, each object in the training set has a label. The goal of learning is then to map a given object to its label according to the information learned from the training set. However, in Multiple Instance Learning (MIL), the label of an individual instance i.e. object is unknown. Only the label of a set of instances is known, which is called the label of the bag. Thus, the MIL problem becomes - how to map an instance to its label according to the information learned from the bag labels.

In RF-based CBIR systems, we have two types of labels – *Positive* and *Negative*. Each image is considered a bag of semantic regions (instances). In giving feedback to the retrieved images, a user labels an image as positive if it contains the region of interest; otherwise, it is labeled as negative. As a result, the label of each retrieved image, i.e. bag label, is available. However, the labels of the semantic regions in that image bag still remain unknown, because the user only gives feedback to an image as a whole, not to individual semantic regions in that image. The goal of MIL, in the context of CBIR, is to learn the label of each semantic region in the training set and use this information to estimate the similarity scores of the test image regions. In this way, the single object based CBIR problem can then be transformed to a MIL problem as defined below.

Given a set of training examples $T=<B, L>$ where $B=B_i(i=1,\ldots,n)$ is a set of n bags and $L=L_i(i=1,\ldots,n)$ is a set of labels of the corresponding bags. $L_i \in \{1(\text{Positive}), 0(\text{Negative})\}$. The goal of MIL is to identify the label of a given instance in a given bag.

In image retrieval, an image is labeled positive if it contains a region of interest. The relation between a bag (image) label and the labels of all its instances (regions) is defined as below.

$$L_i = 1 \quad if \quad \exists_{j=1}^{m} l_{ij} = 1 \tag{1}$$

$$L_i = 0 \quad if \quad \forall_{j=1}^{m} l_{ij} = 0 \tag{2}$$

Assume there are m instances in B_i. l_{ij} is the label of the j^{th} instance in the i^{th} bag. If the bag label is positive, there exists at least one positive instance in that bag. If the bag label is negative, all instances in that bag are negative.

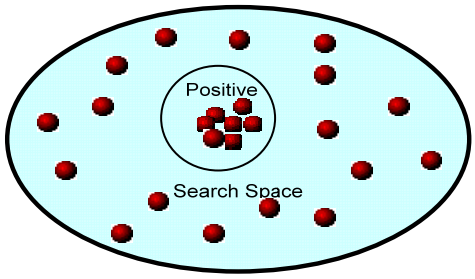
Given a query image, at each query iteration, the user's feedbacks on the whole images in the training set are collected and fed into learning algorithm. After studying these user-labeled positive and negative images, the proposed algorithm will use the knowledge just learned to perform image retrieval and return to the user the most similar ones. In this study, the One-Class SVM is adopted as the underlying learning algorithm.

4. One-Class Support Vector Machine

One-Class classification is a kind of unsupervised learning mechanism. It tries to assess whether a test point is likely to belong to the distribution underlying the training data. In our case, the training set is composed of positive image samples only. One-Class SVM has so far been studied in the context of SVMs {Schölkopf, Platt, et al., 1999}. The objective is to create a binary-valued function that is positive in those regions of input space where the data predominantly lies and negative elsewhere.

The idea is to model the dense region as a "ball" – hyper-sphere. In Multiple Instance Learning (MIL) problem, positive instances are inside the "ball" and negative instances are outside. If the origin of the "ball" is $\vec{\alpha}$ and the radius is *r*, a point $\vec{x_i}$, in this case an instance (image region) represented by an 32-feature vector, is inside the "ball" *iff* $\left\|\vec{x_i} - \vec{\alpha}\right\| \le r$. This is illustrated in Figure 1 with samples inside the circle being the positive instances.

Figure 1 One-Class classification



This "ball" is actually a hyper-sphere. The goal is to keep this hyper-sphere as "pure" as possible and include most of the positive objects. Since this involves a non-linear distribution in the original space, the strategy of Schölkopf's One-Class SVM is first to do a mapping θ to transform the data into a feature space *F* corresponding to the kernel *K*:

$$\theta(x_i) \cdot \theta(x_j) \equiv K(x_i, x_j) \tag{3}$$

where x_i and x_j are two data points. In this study, we choose to use Radial Basis Function (RBF) Machine below.

$$K(u, v) = \exp\left(\left\|x_i - x_j\right\| / 2\sigma\right) \tag{4}$$

Mathematically, One-Class SVM solves the following quadratic problem:

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\| - \rho + \frac{1}{\alpha n}\sum_{i=1}^{n}\xi_i \tag{5}$$

which is subject to

$$(w \cdot \theta(x_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \text{ and } i = 1,...,n \tag{6}$$

where ξ_i is the slack variable, and $\alpha \in (0,1)$ is a parameter that controls the trade off between maximizing the distance from the origin and enclosing most of the data samples in the "ball" formed by the hyper-sphere. The slack variable actually corresponds to the ratio of "outliers" in the training dataset. n is the total number of data points in the training set. If w and ρ are a solution to this problem, then the decision function is $f(x) = sign(w \cdot \theta(x) - \rho)$ and it will be 1 for most examples x_i contained in the training set.

5. Training by One-Class SVM

Given a query image, in initial query, the user needs to identify a semantic region of his/her interest. Since no training data is available at this point, we simply compute the Euclidean distances between the query semantic region and all the other semantic regions in the image database. The smaller the distance, the more likely this semantic region is similar to the query region. The similarity score for each image is then set to the inverse of the minimum distance between its regions and the query region. The training sample set is then constructed according to the user's feedback.

The user's feedback provides labels (positive/negative) for a small set of images (e.g., top 30 most similar images). Although the labels of individual instances remain unknown, which is our ultimate goal, we do know the relationship between images and their enclosed regions. If the bag is labeled positive, at least one instance in the bag is positive. If the bag is labeled negative, all the instances in the bag are negative. This is the particular Multiple Instance Learning problem we are trying to solve.

If an image is labeled positive, its semantic region that is the least distant from the query region is labeled positive. In this way, most of the positive regions can be identified. In our experiment, we choose to use Blob-world {Carson, Belongie, Greenspan and Malik, 2002} as our image segmentation method. For some images, Blob-world may "over-segment" such that one semantic region is segmented into two or more "blobs". In addition, some images may actually contain more than one positive region. Therefore, we cannot assume that only one region in each image is positive. Instead, we assume the number of positive images is *h* and the number of all semantic regions in the training set is *H*. Then the ratio of "outliers" in the training set is estimated as:

$$\alpha = 1 - (\frac{h}{H} + z) \tag{7}$$

z is a small number used to adjust α in order to alleviate the above mentioned problem. Our experiment results show that *z*= 0.01 is a reasonable value.

The training set as well as the parameter α are fed into One-Class SVM to obtain w and ρ, which are used to calculate the value of the decision function for the test data, i.e. all the image regions in the database. Each image region will be assigned a "score" by $w \cdot \theta(x) - \rho$ in the decision function. The higher the score, the more likely this region belongs to the positive class. The similarity score of each image is then set to the highest score of all its regions. It is worth mentioning that except for the initial query in which the

user needs to specify the query region in the query image, the subsequent iterations will only ask for the user's feedback on the whole image.

6. Latent Semantic Indexing (LSI)

Latent Semantic Indexing (LSI) was originally used as a mathematical/statistical technique for text mining. It is a novel information retrieval method developed by Deerwester et al. {Deerwester, Dumais, Landauer, Furnas and Harshman, 1990}. It is often the case that two documents may be semantically close even if they do not share a particular keyword. The "power" of LSI is that it can find and rank relevant documents even they do not contain the query keywords. The whole procedure is fully automatic.

The fact that LSI does not require an exact match to return useful results fits perfectly with the scenario of region-based image retrieval. Suppose there is a query image – a "tiger" in the grass and the user is interested in finding all images in the image database that contain "tiger". It is obviously not a good idea to use exact match since no "tiger" image would have exactly the same low-level features as the query image except the query image itself. If we consider an image as a "document", the "tiger" object is then one of the words in the document. The only difference is that the "tiger" object is not a word, but a multi-dimensional feature vector.

6.1 Constructing "Term-Document" Matrix

The first step in Latent Semantic Indexing is to construct the term-document matrix. It is a 2-D grid with documents listed along the horizontal axis, and content words along the vertical axis. For the image retrieval purpose, we construct a matrix A in a similar sense except that "documents" are images and "content words" are image regions. The matrix has all the training data (user feedbacks) collected by using the method presented in Section 5. Table 1 shows a part of the Matrix.

Table 1 The term-document matrix A

		I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	…
O_1		0	0	2	2	-2	-2	0	2	…
O_2		1	1	-1	-1	2	0	0	0	…
O_3		0	0	0	0	0	1	1	0	…
…	…									…

I_1, I_2, … represent images and O_1, O_2,… are trained data i.e. query image objects. Given an image object O_i, if it is queried by the user, the system depicted in Section 5 will return a series of results upon which user will provide feedback. These feedbacks are collected and stored in the "term-document" matrix. If a returned image I_j is "positive", the corresponding cell value $A_{ij}(O_i, I_j)$ will be incremented by 1. This corresponds to the situation in LSI-based text mining in which the corresponding cell value in the term-

document matrix will be increased by 1 if an exact match of a query keyword is found in a document. If it is "negative", the value of $A_{ij}(O_i, I_j)$ will be decreased by 1. Otherwise, i.e. its relevance to the query object is unknown (e.g., unlabeled images), the corresponding value in A_{ij} is "0".

In our experiment, there are in total 9800 images in the database. Therefore the final matrix A has 9800 columns. The training data are obtained from the access log information of this database i.e. query objects, retrieved results by the system mentioned in Section 5, and users' relevance feedbacks. User queries and feedbacks are collected and indexed continuously over a certain period of time. The final matrix A has 1493 rows, i.e. 1493 distinct query objects. At the end, we normalize the matrix by using the z-score method.

6.2 Singular Value Decomposition

The key step in Latent Semantic Indexing is to decompose the "term-document" matrix using a technique called Singular Value Decomposition (SVD). LSI works by projecting a large multidimensional space down onto a smaller number of dimensions. In doing so, images that are semantically similar will get squeezed together. The SVD preserves as much information as possible about the relative distance between images while collapsing them down into a much smaller set of dimensions. In doing so, noise data are removed and the latent semantic similarities are revealed. In other words, after SVD, similar things become more similar while dissimilar things remain distinct.

By SVD, the "term-document" matrix is first decomposed into a set of smaller components. Let $A_{m\times n}$ be the original "term-document" matrix that has m rows and n columns, the resulting components of A are $U_{m\times n}$, $S_{n\times n}$ and $V_{n\times n}$ as shown in Figure 2.

The columns of U are the left singular vectors which are made up by the eigenvectors of AA^{T}. V^{T} has rows that are the right singular vectors, which are made up by the eigenvectors of $A^{\mathrm{T}}A$. S is a diagonal matrix that has the same number of columns as A does. It has the singular values in descending order along the main diagonal of S. These values are square roots of the eigenvalues of $A^{\mathrm{T}}A$. The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal.

"The beauty of an SVD is that it allows a simple strategy for optimal approximate fit using smaller matrices {Deerwester, Dumais, Landauer, Furnas and Harshman, 1990}". Since the singular values in S are sorted in a non-increasing order, the most important information actually resides in the top k largest values. Therefore, by keeping these k values, we are trying to eliminate noise to the greatest extent. A new matrix, *ASVD*, is then constructed with the rank k (see Figure 3).

Figure 2 Singular Value Decomposition of the "Term-Document" Matrix

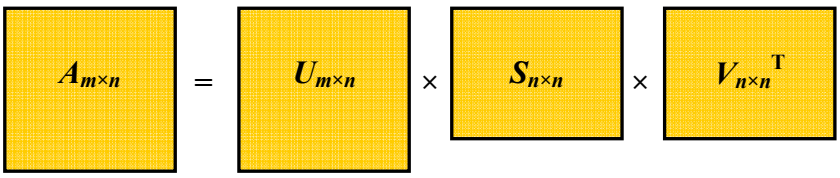
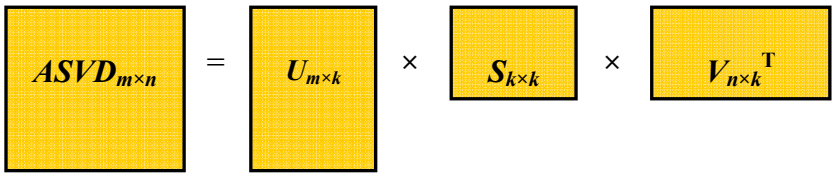


Figure 3 The reduced "Term-Document" Matrix after Singular Value Decomposition



In our experiment, we have a database of 9800 Corel images. They roughly fall into 100 categories. Therefore, the k shall be somewhere around 100. To estimate the value for k, we construct a histogram with all the singular values and find out that when k equals 172 the histogram experiences the sharpest drop. Therefore, in our case, k is chosen to be 172. It is worth mentioning that the estimation of appropriate k for different image databases can be done with the aid of image clustering which is beyond the scope of this paper. The detailed experimental results are presented in Section 7.

6.3 Region-based image retrieval by LSI

The matrix *ASVD*, as described above, contains the SVD-transformed access log information. The next key problem is how to make use of this matrix in our region-based image retrieval system.

In the initial query, we simply compute the Euclidean distance between the query object and all the other objects in the database. The top images are those whose objects have the shortest distances with the query object and are returned to the user for feedback. User feedbacks are either "positive" or "negative". In our previous study {Zhang, Chen, Chen, Chen and Shyu, 2005}, "positive" images are directly fed into the One-class SVM for learning purpose. In some cases, the number of "positive" images retrieved by the initial query is very small (e.g., 2~3 positive images out of the top 30 images). This lack of positive samples hinders the learning performance of One-class SVM. By using the access log information, it is expected that more positive samples can be provided to the One-class SVM.

Given the "positive" images, their relationships with other images in the database with respective to the given query object can be revealed by looking up their corresponding entries in the *ASVD* matrix. The similarity between two images in the matrix is measured by the dot product of the two column vectors that represent the two images. Since this is a region-based image retrieval system, we cannot simply treat each image as a column vector in full length. Instead, only those rows whose objects have short distances with the query object shall be considered. Therefore, a threshold needs to be set for the distance between the query object and all the other objects along rows of *ASVD*. The purpose is to limit the influence of the trained objects that are far different from the query object in terms of low-level features. In our experiment, we set this threshold to 3. If no object in *ASVD* is close enough to the query object given the threshold, we simply perform the query by using the original system {Zhang, Chen, Chen, Chen and Shyu, 2005} in which only instant (short-term) learning is performed. This capability is extremely important for a practical content-based image retrieval system as both short-term learning and long-term learning are needed.

Table 2(a) An example of Matrix *ASVD*

	I_1	I_2	I_3	I_4	I_5	I_6
O_1	- 1.3	0.5	0.8	7.2	- 0.1	-8.5
O_2	1.5	2.5	- 0.5	0.05	0.1	-7.5
O_3	- 4.3	1.5	0.8	4.3	0.55	20.5
O_4	2.9	3.5	- 6.7	0.09	-6.8	-0.3

Table 2(b) The reduced image vectors

	I_1	I_2	I_3	I_4	I_5	I_6
O_1	- 1.3	0.5	0.8	7.2	- 0.1	-8.5
O_3	- 4.3	1.5	0.8	4.3	0.55	20.5

An example of matrix *ASVD* is given in Table 2. Let the query image be I_1, and the query object is O_q. In Table 2(a), assume that the objects O_1 and O_3 are the objects whose distances to O_q are less than the threshold. Therefore, the similarity between I_1 and all the other images in the database with respect to O_q is measured by the dot product of column vectors as shown in Table 2(b). Further, in order to alleviate the problem of lacking positive samples, we expand the set of positive images marked by the user by adding those images that are close enough (with similarity value greater than 1) to ALL the positive images with respect to the query object O_q. For each of these image samples, its object that has the shortest Euclidean distance to the query object O_q is identified and fed into the One-class SVM. The Euclidean distance is used since the SVM model is not yet available (trained) at this point.

It should be noted here that, in the initial query, we did not simply fetch all the relevant images that have been previously marked positive by users from the log file (Matrix *A*). Instead, we use the same initial query method as adopted in our prototype system {Zhang, Chen, Chen, Chen and Shyu, 2005}. We did this simply for comparison purpose. As its initial query mechanism is the same as that of the prototype system, we can easily measure the effectiveness of the proposed mechanism by examining the performance gains in the subsequent learning and retrieval cycles when log information is used. Then another question may be raised – which matrix is better, in terms of providing useful information to One-class SVM, *A* or *ASVD*? As shown in our experiment in Section 7, *ASVD* provides more useful information regarding positive samples and therefore One-class SVM performs better in the next round. Another advantage of using *ASVD* is that it does not require exact match with the query object while *A* does require it. In other words, only those images that have been previously marked positive can be retrieved when matrix *A* is used, while using *ASVD* can discover potentially positive images.

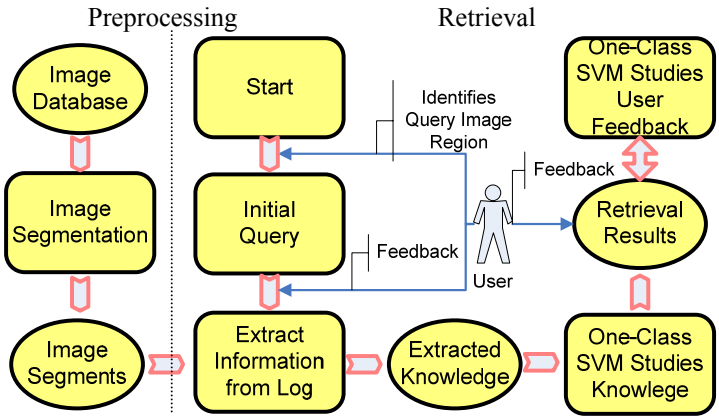
7. Experimentation

7.1 Image segmentation and feature extraction

Figure 4 shows the flowchart of the proposed system. In the preprocessing phase, images are first segmented into semantic regions, with each represented by a 32-feature vector – three texture features, two shape features and 27 color features.

In the initial query, the system gets the feature vector of the query region and compares it with all the other image regions in the database using Euclidean distance. After the initial query, user gives feedback to the retrieved images and these feedbacks are returned to the system. In the first round of query, these feedbacks are used to find relevant information from database logs as detailed in Section 6.3. If useful knowledge is found, i.e. that object has been queried before, our One-Class SVM based algorithm learns from this knowledge otherwise it learns directly from the user's current feedbacks and refines the retrieval results. Then another round of retrieval starts.

Figure 4 The LMS architecture



7.2 System performance evaluation

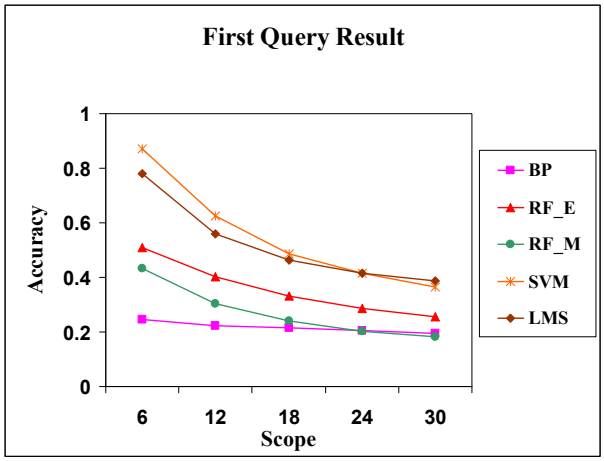
The experiment is conducted on a Corel image database consisting of 9,800 images from 98 categories. After segmentation, there are in total 82,552 image segments. Fifty images are randomly chosen from 20 categories as the query images. In our database log, we have collected altogether 1493 distinct queries.

In order to test the performance of LMS, we compare our algorithm with the one that does not consider log information {Zhang, Chen, Chen, Chen and Shyu, 2005}. We also compare the performance of our system with two other relevance feedback algorithms: 1) Neural Network based Multiple Instance Learning (MIL) algorithm with relevance feedback {Chen, Rubin, Shyu and Zhang, 2006}; 2) General feature re-weighting algorithm {Rui, Huang and Mehrotra, 1997} with relevance feedback. For the latter, both Euclidean and Manhattan distances are tested.

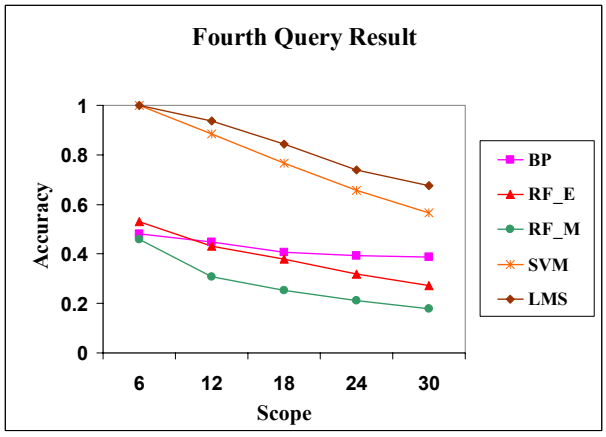
Five rounds of relevance feedback are performed for each query image - Initial (no feedback), First, Second, Third, and Fourth. The accuracy rates within different scopes, i.e. the percentage of positive images within the top 6, 12, 18, 24 and 30 retrieved images,

are calculated. Figure 5(a) shows the result from the First Query while Figure 5(b) shows the result after the Fourth Query. "BP" represents the Neural Network based MIL which uses both positive and negative examples. "RF_E" is a feature re-weighting method using Euclidean Distance while "RF_M" uses Manhattan Distance. "LMS" is the proposed system and "SVM" refers to the same retrieval mechanism except that database log information is not used in the retrieval process {Zhang, Chen, Chen, Chen and Shyu, 2005}.

Figure 5 (a) The retrieval accuracy after the 1st query; (b) the retrieval accuracy after the 4th query



(a)



(b)

It can be seen from Figure 5 that the accuracy of the proposed system outperforms all the other 3 algorithms. Especially, the proposed algorithm shows better performance over "SVM" – the one that does not consider log information. It also can be seen that the Neural Network based MIL (BP), although not as good as the feature re-weighting method (RF_E and EF_M) in the First Query, demonstrates a better performance than that

of general feature re-weighting algorithm after 4 rounds of learning.

In order to answer the question raised in Section 6.3 i.e. which matrix is better, in terms of providing useful information to One-class SVM, *A* or *ASVD*, we further compare the retrieval results using *A* and *ASVD*. In the experiment with matrix *A*, we extract the information of positive images directly from *A*. If the query region has been trained before, all the relevant image regions marked by users can be found directly from *A*. These relevant regions are directly returned as the query result of the initial query. However, with *ASVD*, we use Euclidean distance measure to first find the relevant image regions in the initial query. The user's feedbacks on the initial query results are then used to extract relevant information from *ASVD* in the way as described in Section 6.3. That is to say, with *A,* we start to use log information from the initial query whereas with *ASVD*, we start to use log from first query. However, this does not give much advantage to *A*. This is evidenced by the fact that after the first two rounds of retrieval, the performance gain in terms of accuracy drops sharply. It can be expected that, in the long run, *ASVD* may outperform *A*, since *ASVD* is better in discovering the correlations between images by providing a more compact, less noisy representation of data logs. This is evident in our experiment. Figure 6 shows the fourth round retrieval result of a "horse" region using *A*. Figure 7 shows the fourth round retrieval result of the same region using *ASVD*. As shown in Figures 6 and 7, the upper-left image is the query image. This image is segmented into 8 semantic regions (outlined by red lines). The user identifies the "horse" region as the region of interest (the 4th image from left in the 1st row, outlined by a blue rectangle). For this specific query object, the retrieved result using *ASVD* is better than that using *A*.

Figure 6 A sample retrieval result by using Matrix *A*

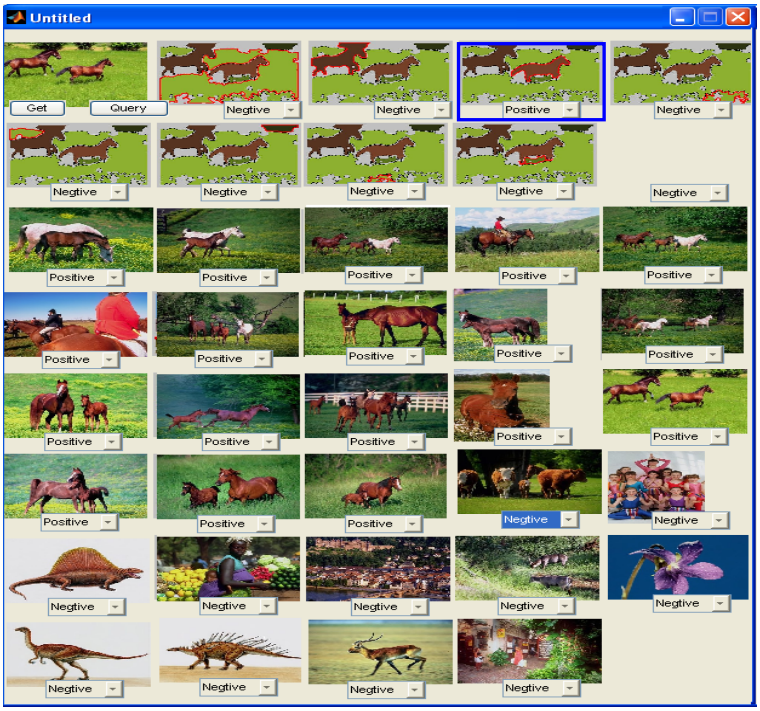


Figure 7 A sample retrieval result by using Matrix *ASVD*

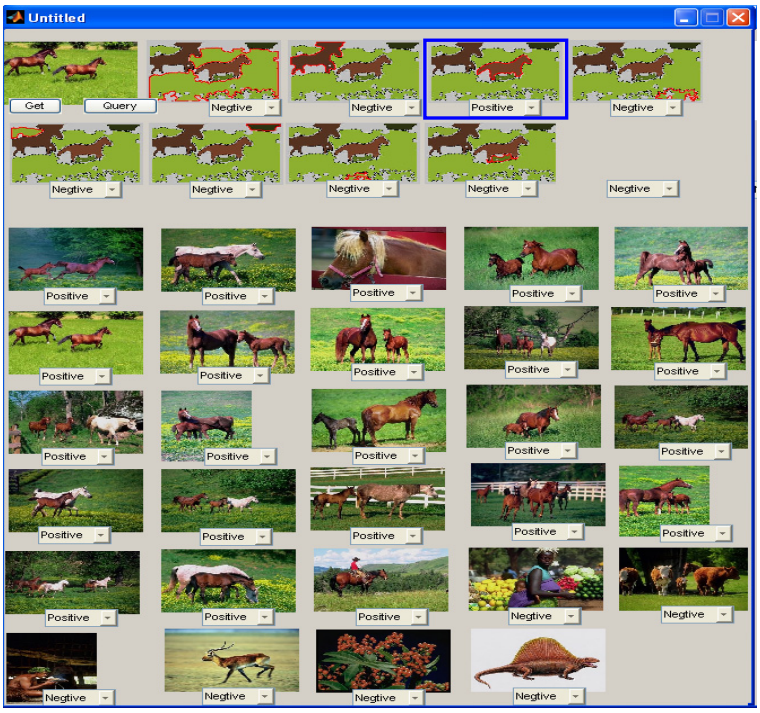
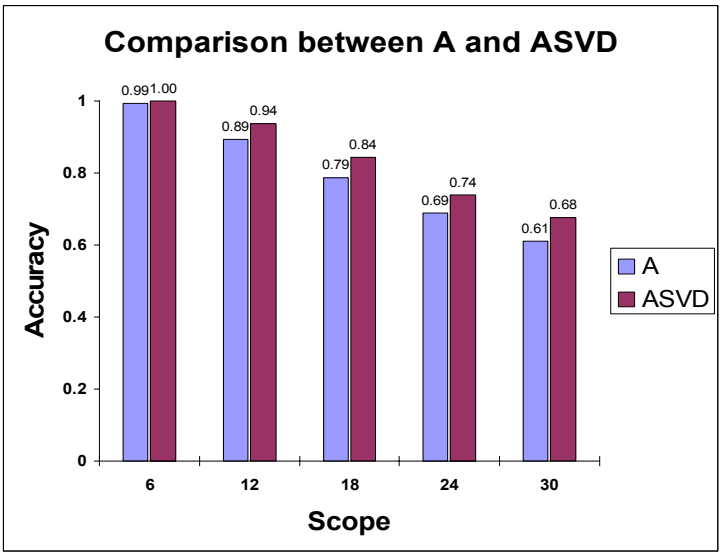


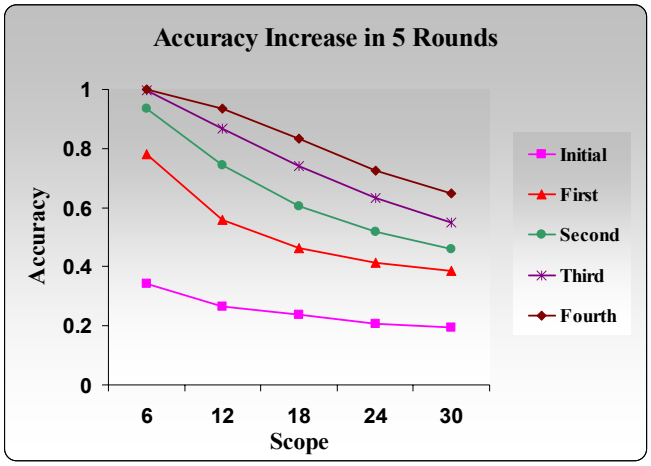
Figure 8 shows the average accuracy after the fourth query of the 50 query images using matrices *A* and *ASVD*, respectively. On average, the performance of the latter is also consistently better than that of the former.

Figure 8 The comparison between *A* and *ASVD* on the 4th query



In Figure 9, the accuracy rates of our system across 5 iterations are illustrated. Through each iteration, the number of positive images increases steadily.

Figure 9 The retrieval accuracies of LMS across 5 iterations



8. Conclusions

In this paper, we propose a long term learning system for region-based Content-based Image Retrieval, which is a human-centered and knowledge-based system. This system solves a Multiple Instance Learning (MIL) problem for single region based CBIR system. In view of the fact that the information acquired by Relevance Feedback is often lost at the end of search, we propose a method to make use of the previously acquired feedbacks for region-based image retrieval. The proposed method uses Latent Semantic Indexing which can fully exploit the feedback information and its effectiveness is demonstrated by experiments. We also adopt One-Class SVM as the actual learning mechanism to solve the mapping problem when MIL is applied to region-based CBIR. A particular advantage of the proposed system is that it targets on both short-term learning and long-term learning for region-based image retrieval, which is desired by contemporary CBIR systems since the user is often interested in specific region(s) in an image and often expects satisfactory results to be returned in a fast fashion. Our experiments demonstrate that the proposed system can better identify user's real need in region-based image retrieval.

9. References

Andrews, S., Tsochantaridis, I. and Hofmann, T. (2003) 'Support Vector Machines for Multiple-Instance Learning', *Advances in Neural Information Processing Systems 15*, Cambridge, MA: MIT Press, pp. 561-568.

Auer, P. (1997) 'On learning from multi-instance examples: Empirical evaluation of a theoretical approach', *Proceedings of the 14th International Conference on Machine Learning*, San Francisco, CA, USA, pp. 21–29.

Carson, C., Belongie, S., Greenspan, H. and Malik, J. (2002) 'Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No.8, pp.1026-1038.

Chen, S.-C., Rubin, S. H., Shyu, M.-L., and Zhang, C. (2006) 'A Dynamic User Concept Pattern Learning Framework for Content-Based Image Retrieval', *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, Vol. 36, Issue 6, pp. 772-783.

Chen, Y. X. and Wang, J. Z. (2004) 'Image Categorization by Learning and Reasoning with Regions', *Journal of Machine Learning Research*, Vol. 5, pp. 913-939.

Chen, Y., Zhou, X., Tomas, S. and Huang, T. S. (2001) 'One-Class SVM for Learning in Image Retrieval', *Proceedings of IEEE International Conference on Image Processing*.

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. (1990) 'Indexing by Latent Semantic Analysis', *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391- 407.

Dietterich, T. G., Lathrop, R. H. and Lozano-Perez, T. (1997) 'Solving the Multiple-Instance Problem with Axis-Parallel Rectangles', *Artificial Intelligence Journal*, Vol. 89, pp. 31-71.

Fournier, J. and Cord, M. (2002) 'Long-term similarity learning in content-based image retrieval', *Proceedings of the 2002 International Conference on Image Processing*, pp. 441-444, Rochester, New-York, USA.

Gondra, I. and Heisterkamp, D. R. (2004) 'Adaptive and Efficient Image Retrieval with One-Class Support Vector Machines for Inter-Query Learning', *WSEAS Transactions on Circuits and Systems*, Vol. 3, No. 2, pp. 324-329.

Han, J. K., Ngan, N., Li, M. and Zhang, H.-J. (2005) 'A Memory Learning Framework for Effective Image Retrieval', *IEEE Transactions on Image Processing*, Vol. 14, No. 4, pp. 511–524.

He, X., King, O., Ma, W. Y., Li, M. and Zhang, H. (2003) 'Learning a semantic space from user's relevance feedback for image retrieval', *IEEE Transactions on Circuits System and Video Technology*, Vol. 13, No. 1, pp. 39–48.

Hoi, C.-H. and Lyu, M. R. (2004) 'A Novel Log-based Relevance Feedback Technique in Content-based Image Retrieval', *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 24 – 31, New York, USA.

Jing, F., Li, M. M., Zhang, L., Zhang, H.-J. and Zhang, B. (2003) 'Learning in Region-based Image Retrieval', *Proceedings of International Conference on Image and Video Retrieval*.

Jing, F., Li, M. M., Zhang, L., Zhang, H.-J., and Zhang, B. (2003) 'Support Vector Machines for Region-Based Image Retrieval', *Proceedings of IEEE International Conference on Multimedia & Expo*.

Lee, C., Ma, W. Y. and Zhang, H. (1999) 'Information embedding based on user's relevance feedback for image retrieval', *Proceedings of the SPIE Conference on Multimedia Storage and Archiving Systems IV*, Boston, MA.

Li, M., Chen, Z. and Zhang, H. (2002) 'Statistical correlation analysis in image retrieval', *Pattern Recognition*, Vol. 35, pp. 2687–2693.

Maron, O. and Lozano-Perez, T. (1998) 'A Framework for Multiple Instance Learning', *Advances in Natural Information Processing System 10*. Cambridge, MA, MIT Press.

Ramon, J. and De Raedt, L. (2000) 'Multi-Instance Neural Networks', *Proceedings of the ICML 2000 Workshop on Attribute-value and Relational Learning*.

Ray, S. and Page, D. (2001) 'Multiple-instance regression', *in Proc. 18th International Conference on Machine Learning*, pp. 425–432.

Rui, Y., Huang, T. S. and Mehrotra, S. (1997) 'Content-based Image Retrieval with Relevance Feedback in MARS', *Proceedings of the International Conference on Image Processing*, pp. 815-818.

Schölkopf, B., Platt, J.C. et al (1999) 'Estimating the Support of a High-Dimensional Distribution', Microsoft Research Corporation Technical Report MSR-TR-99-87.

Shyu, M.-L., Chen, S.-C., Chen, M. and Zhang, C. (2004) 'Affinity Relation Discovery in Image Database Clustering and Content-based Retrieval', *Proceedings of the 2004 ACM Multimedia Conference*, pp. 372-375, October 10-16, New York, USA.

Su, Z., Zhang, H. J., Li, S. and Ma, S. P. (2003) 'Relevance Feedback in Content-based Image Retrieval: Bayesian Framework, Feature Subspaces, and Progressing Learning', *IEEE Transactions on Image Processing*, Vol. 12, No. 8, pp. 924-937.

Wang, J. and Zucker, J.-D. (2000) 'Solving the Multiple Instance Learning Problem: A Lazy Learning Approach', *Proceedings of the 17th International Conference on Machine Learning*, pp. 1119-1125.

Zhang, C., Chen, X., Chen, M., Chen, S.-C. and Shyu, M.-L. (2005) 'A Multiple Instance Learning Approach for Content Based Image Retrieval Using One-Class Support Vector Machine', *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, pp. 1142-1145, July 6-8, 2005, Amsterdam, The Netherlands.

Zhang, Q. and Goldman, S. A. (2002) 'EM-DD: An Improved Multiple-Instance Learning Technique', *Advances in Neural Information Processing Systems (NIPS)*.

Zhou, X., Zhang, Q., Liu, L., Zhang, L. and Shi, B. (2003) 'An image retrieval method based on analysis of feedback sequence log', *Pattern Recognition Letters*, Vol. 4, No. 14, pp. 2499-2508.

Zucker, J.-D. and Chevaleyre, Y. (2001) 'Solving Multiple-Instance and Multiple-part Learning Problems with Decision Trees and Decision Rules. Application to the Mutagenesis Problem', *Proceedings of the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, AI 2001, pp. 204-214.