User Concept Pattern Discovery Using Relevance Feedback and Multiple Instance Learning for Content-Based Image Retrieval

Xin Huang
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199
USA
xhuan001@cs.fiu.edu

Shu-Ching Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199
USA
chens@cs.fiu.edu

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124
USA
shyu@miami.edu

Chengcui Zhang
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199
USA
czhang02@cs.fiu.edu

ABSTRACT

Understanding and learning the subjective aspect of humans in Content-Based Image Retrieval has been an active research field during the past few years. However, how to effectively discover users' concept patterns when there are multiple visual features existing in the retrieval system still remains a big issue. In this paper, we propose a multimedia data mining framework that incorporates Multiple Instance Learning into the user relevance feedback in a seamless way to discover the concept patterns of users, especially where the user's most interested region and how to map the local feature vector of that region to the high-level concept pattern of users. This underlying mapping can be progressively discovered through the feedback and learning procedure. The role user plays in the retrieval system is to guide the system mining process to his/her own focus of attention. The retrieval performance is tested under a couple of conditions.

Keywords

Multimedia Data Mining, Image Retrieval, Multiple Instance Learning, Relevance Feedback

1. INTRODUCTION

Recently, many efforts have been made to Content-Based Image Retrieval (CBIR) in order to personalize the retrieval engine. The subjectivity of human perception of visual content plays an important role in the CBIR systems. It is very often that the retrieval results are not very satisfactory especially when the level of satisfaction is closely related to user's subjectivity. For example, given a query image with a tiger lying on the grass, one user may want to retrieve those images with the tiger objects in them, while another user may find the green grass background more interesting. User subjectivity in image retrieval is a very complex issue and difficult to explain. Therefore, a CBIR system needs to have the capability to discover the users' concept patterns and adapt to them.

In this paper, we propose a multimedia data mining framework that can dynamically discovering the concept patterns of a specific user to allow the retrieval of images by the user's most interested region. The discovering and adapting process aims to find out the mapping between the local low-level features of the images and the concept patterns of the user with respect to how he/she feels about the images. The proposed multimedia data mining framework seamlessly integrates several data mining techniques. First, it takes advantages of the user feedback during the retrieval process. The users interact with the system by choosing the positive and negative samples from the retrieved images based on their own concepts. The user feedback is then fed into the retrieval system and triggers the modification of the query criteria to best match the users' concepts [14]. Second, in order to identify the user's most interested region within the image, the Multiple Instance Learning [16, 18] and neural network techniques are integrated into the query refining process. The Multiple Instance Learning technique is originally used in categorization of molecules in the context of drug design. Each molecule (bag) is represented by a bag of possible conformations (instances). In image retrieval, each image is viewed as a bag of image regions (instances). In fact, the user feedback guides the system mining through the positive and negative examples, and tells the system to shift its focus of attention to the region of interest. Compared with other Multiple Instance Learning methods used in CBIR, our methodology has the following advantages: 1) Instead of manually dividing each picture into many overlapping regions [16], we adopt the image segmentation method in [5] to partition the images in a more natural way; 2) In

other Multiple Instance Learning based image retrieval systems such as [18], the users are usually asked to provide the positive and negative samples by looking through a huge amount of images in the database. While in our framework, user feedback is used in the image retrieval process, which makes the process more efficient and precise. It is more efficient since it is easy for the user to find some positive samples among the initial retrieved results. It is more precise since among the retrieved images, the user can select the negative samples based on his/her subjective perception. The reason is that the selected negative ones have similar features/contents with the query image but they have different focuses of attention from the user's point of view. By selecting them as negative samples, the system can better distinguish the real needs of the users from the "noisy" or unrelated information via Multiple Instance Learning. As a result, the system can discover which feature vector related to a region in each image best represents the user's concept, and furthermore, it can determines which dimensions of the feature vector are important by adaptively reweighing them through the neural network technique.

This paper is organized as follows. Section 2 briefly introduces the related work in Relevance Feedback and Multiple Instance Learning. Section 3 introduces the details of the Multiple Instance Learning and neural network techniques used in our framework. The proposed multimedia data mining framework for content-based image retrieval using user feedback and Multiple Instance Learning is described in Section 4. The experimental results are analyzed in Section 5. Section 6 gives the conclusion and future work.

2. RELATED WORK

2.1 Retrieval Using Relevance Feedback

While lots of research efforts establish the base of CBIR, most of them relatively ignore two distinct characteristics of the CBIR systems: (1) the gap between high-level concepts and low-level features, and (2) the subjectivity of human perception of visual content. To overcome these shortcomings, the concept of relevance feedback (RF) associated with CBIR was proposed in [13]. Relevance feedback is an interactive process in which the user judges the quality of the retrieval performed by the system by 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. This process iterates until a satisfactory result is obtained for the user.

In the past few years, the RF approach to image retrieval has been an active research field. This powerful technique has been proved successful in many application areas. Various ad hoc parameter estimation techniques have been proposed for the RF approaches. The method of RF is based on the most popular vector model [4] used in information retrieval. The RF techniques do not require a user to provide accurate initial queries, but rather estimate the user's ideal query by using positive and negative examples (training samples) provided by the user. The fundamental goal of these techniques is to estimate the ideal query parameters (both the query vectors and the associated weights) accurately and robustly. Most of the previous RF researches [1][6] are based on the low-level image features such as color, texture and shape and can be classified into two approaches: query point movement and re-weighting techniques [8]. More recently, the new trend towards taking advantages of the semantic contents of the images in addition to the low-level features has appeared.

2.2 Multiple Instance Learning

Dietterich et al. [7] introduced the Multiple Instance Learning problem and presented Multiple Instance Learning algorithms for learning axis-parallel rectangles (APR). In [3], Auer et al. proposed MULTIINST algorithm for Multiple Instance Learning that is also an APR based method. In [10], Maron et al. introduced the concept of Diversity Density and applied a two-step gradient ascent with multiple starting points to find the maximum Diversity Density. Based on the Diversity Density, Qi Zhang et al. [17] proposed EM-DD algorithm. In their algorithm, it was assumed that each bag has a representative instance and treated it as a missed value, and then the EM (Expectation-Maximization) method and Quasi-Newton method were used to learn the representative instances and maximize the Diversity Density simultaneously. [12] also used the EM method to do Multiple Instance Regression. Jun Wang et al. [15] explored the lazy learning approaches in Multiple Instance Learning. They developed two kNN-based algorithms: Citation-kNN and Bayesian-kNN. In [19], Jean-Daniel Zucker et al. tried to solve the Multiple Instance Learning problem with decision trees and decision rules. Jan Ramon et al. [11] proposed the Multiple Instance Neural Network. Stuart Andrews et al. [2] utilized the Support Vector Machine in Multiple Instance Learning.

In this paper, one of the main goals is to map the original visual feature space into a space that better describes the user desired high-level concepts. In other words, we try to discover the specific concept patterns for an individual user via user feedback and Multiple Instance Learning. In our method, we assume the user searches for those images close to the query image and responds to a series of machine queries by declaring the positive and negative sample images among the displayed images. Efficiency can be measured by the average number of queries necessary to locate the desired images. For this purpose, we introduce a multiple instance feedback model that accounts for various concepts/responses of the user. Each

new query is chosen to achieve the user expectation more closely given the previous user responses. Compared with the traditional RF techniques, our method differs in the following two aspects:

1. It is based on such an assumption that the users are usually more interested in one specific region (blob object) than other regions of the query image. However, to our best knowledge, the recent efforts in the RF techniques are based on the global image properties of the query image. In order to produce a higher precision, we use the segmentation method proposed in [5] to segment an image into regions (segments) that roughly correspond to objects, which provides the possibility for the retrieval system to discover the most interested region for a specific user based on his feedback.

2. In many cases, what the user is really interested in is just a region (an object) of the query image (example). However, the user's feedback is on the whole image. How to effectively identify the user's most interested region (object) and to precisely capture the user's high-level concepts based on his/her feedback on the whole image have not received much attention yet. In this paper, we apply Multiple Instance Learning method to discover the user's interested region and then mine the user's high-level concepts. By doing so, not only the region-of-interest can be discovered, but also the ideal query point of that query image can be approached within several iterations.

3. THE PROPOSED MULTIPLE INSTANCE LEARNING FRAMEWORK

In a traditional supervised learning scenario, each object in the training set has a label associated with it. The supervised learning can be viewed as a search for a function that maps an object to its label with the best approximation to the real unknown mapping function, which can be described with the following:

Definition 1. *Given an object space* Ω*, a label space* Ψ*, a set of objects* $O=\{O_i \mid O_i \in \Omega\}$ *and their associated labels* $L=\{L_i \mid L_i \in \Psi\}$*, the problem of supervised learning is to find a mapping function* $\hat{f}:\Omega \to \Psi$ *so that the function* $\hat{f}$ *has the best approximation of the real unknown function* f*.*

Unlike the traditional supervised learning, in multiple instance learning, the label of an individual object is unknown. Instead, only the label of a set of objects is available. An individual object is called an instance and a set of instances with an associated label is called a bag. Specifically, in image retrieval there are only two kinds of labels which are Positive and Negative respectively. A bag is labeled Positive if the bag has one or more than one Positive instance and is labeled negative if and only if all its instances are Negative. The Multiple Instance Learning problem is to learn a function mapping from an instance to a label (either *Positive* or *Negative*) with the best approximation to the unknown real mapping function, which can be defined as follows:

Definition 2. *Given an instance space* Φ*, a label space* $\Psi=\{1\ (\text{Positive}), 0\ (\text{Negative})\}$*, a set of* n *bags* $B=\{B_i \mid B_i \in \mathrm{P}(\Phi),\ i=1...n\}$*, where* $\mathrm{P}(\Phi)$ *is the power set of* Φ*, and their associated labels* $L=\{L_i \mid L_i \in \Psi\}$*, the problem of Multiple Instance Learning is to find a mapping function* $\hat{f}:\Phi \to \Psi$ *so that the function* $\hat{f}$ *has the best approximation of the real unknown function* f*.*

3.1 Problem Definition

Let $T=\langle B,L\rangle$ denote a training set where $B=\{B_i\}(i=1...n)$ are the n bags in the training set; $L=\{L_i\}(i=1...n)$ are the set of labels of B and L_i is the label of B_i. A bag B_i contains m_i instances that are denoted by I_{ij} $(j=1,\cdots,m_i)$. The function f is the real unknown mapping function that maps an instance to its label, and the function f_{MIL} denotes the function that maps a bag to its label. In Multiple Instance Learning, a bag is labeled *Positive* if at least one of its instances is *Positive*. Otherwise, it has *Negative* label. Hence, the relationship between the functions f and f_{MIL} can be described in Figure 1.

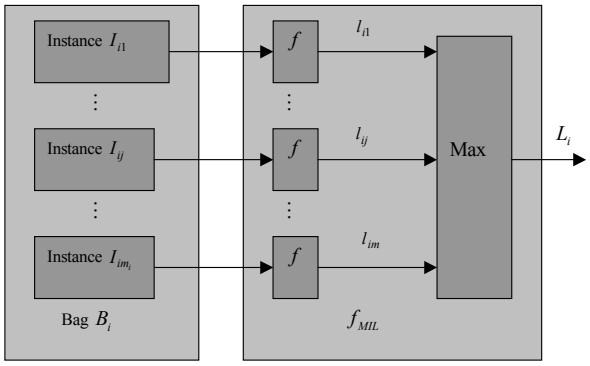

Figure 1. Relationship between functions f and f_{MIL}

As can be seen from this figure, the function f maps each instance I_{ij} in bag B_i to its label l_{ij}. The label L_i of the bag B_i is the maximum of the labels of all its instances, which means $L_i = f_{MIL}(B_i) = \underset{j}{MAX}\{l_{ij}\} = \underset{j}{MAX}\{f(I_{ij})\}$. The Multiple Instance Learning is to find a mapping function $\hat{f}$ with best approximation to function f given a training set $B=\{B_i\}$ and their corresponding labels $L=\{L_i\}$

$(i=1,\cdots,n)$. The corresponding approximation of f_{MIL} is $\hat{f}_{MIL}(B_i)=\underset{j}{MAX}\{\hat{f}(I_{ij})\}$.

In our framework, the Minimum Square Error (MSE) criterion is adopted, i.e., we try to find the function $\hat{f}$ that minimizes

$$SE=\sum_{i=1}^{n}\left(L_i-\hat{f}_{MIL}(B_i)\right)^2=\sum_{i=1}^{n}\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2 \quad (1)$$

Let $\gamma=\{\gamma_k\},(k=1,\cdots,N)$ denote the N parameters of the function f (where N is the number of parameters), the Multiple Instance Learning problem is transformed to the following unconstrained optimization problem:

$$\hat{\gamma}=\arg\underset{\gamma}{Min}\sum_{i=1}^{n}\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2 \quad (2)$$

One class of the unconstrained optimization methods is the gradient search method such as steepest descent method, Newton method, Quasi-Newton method and Back-propagation (BP) learning method in the Multilayer Feed-Forward Neural Network. To apply those gradient-based methods, the differentiation of the target optimization function needs to be calculated. In our Multiple Instance Learning framework, we need to calculate the differentiation of the function $E=\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2$. In order to do that, the differentiation of the MAX function needs to be calculated first.

3.2 Differentiation of the MAX Function

As mentioned in [9], the differentiation of the MAX function results in a 'pointer' that specifies the source of the maximum. Let

$$y=MAX(x_1,x_2,\cdots x_n)=\sum_{i=1}^{n}x_i\prod_{j\neq i}U(x_i-x_j), \quad (3)$$

where $U(\cdot)$ is a unit step function, i.e., $U(x)=\begin{cases}1 & x>0\\ 0 & x\le 0\end{cases}$

The differentiation of the MAX function can be written as:

$$\frac{\partial y}{\partial x_i}=\prod_{j\neq i}U(x_i-x_j)=\begin{cases}1 & \text{if } x_i \text{ is maximum}\\ 0 & \text{otherwise}\end{cases} \quad (4)$$

3.3 Differentiation of the Target Optimization Function

Equation (4) provides a way to differentiate the MAX function. In order to use the gradient-based search method to solve Equation (2), we need to further calculate the differentiation of the function $E=\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2$ on the parameters $\gamma=\{\gamma_k\}$ of function $\hat{f}$. The first partial derivative is as follows:

$$\begin{aligned}\frac{\partial E}{\partial\gamma_k}&=\frac{\partial\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2}{\partial\gamma_k}\\&=2\left(\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}-L_i\right)\times\frac{\partial\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}}{\partial\gamma_k}\\&=2\left(\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}-L_i\right)\\&\quad\times\sum_{j=1}^{m_i}\left(\frac{\partial\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}}{\partial\hat{f}(I_{ij})}\times\frac{\partial\left\{\hat{f}(I_{ij})\right\}}{\partial\gamma_k}\right)\end{aligned} \quad (5)$$

Suppose the s^{th} instance of bag B_i has the maximum value, i.e., $\hat{f}(I_{is})=\underset{j}{MAX}\{\hat{f}(I_{ij})\}$. According to Equation (4), Equation (5) can be written as:

$$\begin{aligned}\frac{\partial E}{\partial\gamma_k}&=2\left(\hat{f}(I_{is})-L_i\right)\times\sum_{j=1}^{m_i}\left(\frac{\partial\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}}{\partial\hat{f}(I_{ij})}\times\frac{\partial\left\{\hat{f}(I_{ij})\right\}}{\partial\gamma_k}\right)\\&=2\left(\hat{f}(I_{is})-L_i\right)\times\frac{\partial\left\{\hat{f}(I_{is})\right\}}{\partial\gamma_k}=\frac{\partial\left(L_i-\hat{f}(I_{is})\right)^2}{\partial\gamma_k}\end{aligned} \quad (6)$$

Furthermore, the n^{th} derivative of the target optimization function E can be written as

$$\frac{\partial^n E}{\partial{\gamma_k}^n}=\frac{\partial^n\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2}{\partial{\gamma_k}^n}=\frac{\partial^n\left(L_i-\hat{f}(I_{is})\right)^2}{\partial{\gamma_k}^n} \quad (7)$$

and the mixed partial derivation of function E can be written as:

$$\begin{aligned}\frac{\partial^{(\sum_k n_k)}E}{\prod_k\partial{\gamma_k}^{n_k}}&=\frac{\partial^{(\sum_k n_k)}\left(L_i-\underset{j}{MAX}\left\{\hat{f}(I_{ij})\right\}\right)^2}{\prod_k\partial{\gamma_k}^{n_k}}\\&=\frac{\partial^{(\sum_k n_k)}\left(L_i-\hat{f}(I_{is})\right)^2}{\prod_k\partial{\gamma_k}^{n_k}}\end{aligned} \quad (8)$$

3.4 Multiple Instance Learning to Traditional Supervised Learning

Similar to the analysis on Multiple Instance Learning problem in Section 3.1, the traditional supervised learning problem can also be converted to an unconstrained optimization problem as shown in Equation (9).

$$\bar{\gamma}=\arg\underset{\gamma}{Min}\sum_{i=1}^{n}\left(L_i-\left\{\hat{f}(O_i)\right\}\right)^2 \quad (9)$$

The partial derivative and mixed partial derivative of the function $\left(L_i - \hat{f}(O_i)\right)^2$ are shown in Equations (10) and (11), respectively.

$$\frac{\partial^n \left(L_i - \hat{f}(O_i)\right)^2}{\partial \gamma_k{}^n} \tag{10}$$

$$\frac{\partial^{(\sum_k n_k)} \left(L_i - \hat{f}(O_i)\right)^2}{\prod_k \partial \gamma_k{}^{n_k}} \tag{11}$$

Notice that Equation (10) is the same as the right side of Equation (7), and Equation (11) is the same as the right side of Equation (8) except that O_i in Equations (10) and (11) represents an object while I_{is} in Equations (7) and (8) represents an instance with the maximum label in bag B_i. This similarity provides us an easy way to transform Multiple Instance Learning to the traditional supervised learning.

The steps of transformation are as follows:

1. For each bag $B_i (i = 1, \cdots, n)$ in the training set, calculate the label of each instance I_{ij} belonging to it.

2. Select the instance with maximum label in each bag B_i. Let I_{is} denote the instance with the maximum label in bag B_i.

3. Construct a set of objects $\{O_i\}(i = 1, \cdots, n)$ using all the instances I_{is} where $O_i = I_{is}$.

4. For each object O_i, construct a label Lo_i that is actually the label of bag B_i.

5. The Multiple Instance Learning problem with the input $\langle \{B_i\}, \{L_i\} \rangle$ is converted to the traditional supervised learning problem with the input $\langle \{O_i\}, \{Lo_i\} \rangle$.

After this transformation, the gradient-based search methods used in the traditional supervise learning such as the steepest descent method can be applied to Multiple Instance Learning.

Despite the above transformation from Multiple Instance Learning to the traditional supervised learning, there still exists a major difference between Multiple Instance Learning and traditional supervised learning. In the traditional supervised learning, the training set is static and usually does not change during the learning procedure. However, in the transformed version of Multiple Instance Learning, the training set may change during the learning procedure. The reason is that the instance with the maximum label in each bag may change with the update of the approximated function $\hat{f}$ during the learning procedure and therefore the training set constructed along with the aforementioned transformation may change during the learning procedure. In spite of such a dynamic characteristic of the training set, the fundamental learning method remains the same. The following is the pseudo code describing our Multiple Instance Learning framework.

MIL (B, L)

Input: $B = \{B_i\}(i = 1, \cdots, n)$ *is the set of* n *bags in the training set.*

$L = \{L_i\}(i = 1, \cdots, n)$ *is the set of labels where* L_i *is the label of bag* B_i

Output: $\gamma = \{\gamma_k\}(k = 1, \cdots N)$ *is the set of parameters of the mapping function* $\hat{f}$ *where* N *is the number of parameters.*

1 Set initial values to parameters γ_k in γ.

2 If the stop criterion has not been met, go to step 3; else return the parameter set γ of function $\hat{f}$.

/* *The stop criterion can be based on MSE or the number of iterations.* */

3 Transform Multiple Instance Learning to traditional supervised learning using the method described in this section.

4 Apply the gradient-based search method in traditional supervised learning to update the parameters in γ.

5 Go to Step 2.

Obviously, the convergence of our Multiple Instance Learning framework depends on what kind of gradient-based search method is applied at Step 4. Actually, it has the same convergence property as the gradient-based search method applied

4. IMAGE RETRIEVAL USING RELEVANCE FEEDBACK AND MULTIPLE INSTANCE LEARNING

In a CBIR system, the most common way is 'Query-by-Example' which means the user submits a query example (image) and the CBIR system retrieves the images that are most similar to the query image from the image database. However, in many cases, when a user submits a query image, what the user really interested in is just a region of the image. The image retrieval system proposed by [5] first segments each image into a couple of regions, and then allows the user to specify the region of interest on

the segmented query image. Unlike the Blobworld system, we use the user's feedback and Multiple Instance Learning to automatically capture the user-interested region during the query refining process. Another advantage of our method is that the underlying mapping between the local visual feature vector of that region and the user's high-level concept can be progressively discovered through the feedback and learning procedure.

In [18], Multiple Instance Learning is applied on CBIR. As a necessary step before actual image retrieval, the user has to first submit a set of images as the training examples that are used to learn the user's target concept. However, it is usually difficult for the user to provide such a training set. In our method, the first set of training examples are obtained from the user's feedback on the initial retrieval results. In addition, the user's target concept is refined iteratively during the interactive retrieval process.

It is assumed that user is only interested in one region of an image. In other words, there exists a function $f \in F : S \to \Psi$ that can roughly map a region of an image to the user's concept. S denotes the image feature vector space of the regions and $\Psi = \{1 \text{ (Positive)}, 0 \text{ (Negative)}\}$ where *Positive* means that the feature vector representing this region meets the user's concept and Negative means not. An image is *Positive* if there exists one or more regions in the image that can meet the user's concept. An image is *Negative* if none of the regions can meet the user's concept. Therefore, an image can be viewed as a bag and its regions are the instances of the bag in Multiple Instance Learning scenario. During the image retrieval procedure, the user's feedback can provide the labels (*Positive* or *Negative*) for the retrieved images and the labels are assigned to the individual images, not on individual regions. Thus, the image retrieval task can be viewed as a Multiple Instance Learning task aiming to discover the mapping function f and thus to mine the user's high-level concept from the low-level features.

At the beginning of retrieval, the user only submits a query image, and there are no training examples available, which means the learning method is not applicable at the current stage. Hence, we use the following metric to measure the similarity of two images. Assume Image A consists of n regions and Image B consists of m regions, i.e., $A = \{A_i\}\ (i = 1, \cdots, n)$ and $B = \{B_j\}\ (j = 1, \cdots, m)$, where A_i is a region of Image A and B_j is a region of Image B. The distance (difference) between Images A and B is defined as:

$$D(A, B) = \min_{1 \le i \le n, 1 \le j \le m} \left\{ \left\| A_i - B_j \right\| \right\} \qquad (12)$$

where $\left\| A_i - B_j \right\|$ is the Euclidean distance between two feature vectors of region A_i and B_j. The larger the $D(A, B)$, the less the similarity between Images A and B. This similarity metric implies that the similarity between two images is decided by the maximum similarity between any two regions of these two images.

Upon the first round of retrieving those "most similar" images, according to Equation (12), the users can give their feedbacks by labeling each retrieved image as *Positive* or *Negative*. Based on the user feedbacks, a set of training examples $\{B+, B-\}$can be constructed where $B+$ consists of all the Positive bags (i.e., the images the user assigns Positive labels) and $B-$ consists of all the Negative bags (i.e., the images the user assigns Negative labels). Given the training examples $\{B+, B-\}$, our Multiple Instance Learning framework can be applied to discover the mapping function f in a progressive way and thus can mine the user's high-level concept.

The feedback and learning are performed iteratively. Moreover, during the feedback and learning process, the capturing of user's high-level concept is refined until the user satisfies. At that time, the query process can be terminated by the user.

5. EXPERIMENTS AND RESULTS

In this section, the experimental setup and the experimental results are presented.

5.1 Image Repository

We created our own image repository using images from the Corel image library. There are 2,500 images collected from various categories for our testing purpose.

5.2 Image Processing Techniques

To apply Multiple Instance Learning on mining users' concept patterns, we assume that the user is only interested in a specific region of the query image. Therefore, we first need to perform image segmentation. The automatic segmentation method proposed in the Blobworld system [5] is used in our system. The joint distribution of the color, texture and location features is modeled using a mixture of Gaussian. The Expectation-Maximization (EM) method is used to estimate the parameters of the Gaussian Mixture model and Minimum Description Length (MDL) principle is used to select the best number of components in Gaussian Mixture model. The color, texture, shape and location characteristics of each region are extracted after image segmentation. Thus, each region is represented by a low-level feature vector. In our experiments, we used three texture features, three color features and two shape features as the representation of an image segment. Therefore, for each

bag (image), the number of its instances (regions) is the number of regions within that image, and each instance has eight features.

5.3 Neural Network Techniques

In our experiments, a three-layer Feed-Forward Neural Network is used as the function f to map an image region (including those eight low-level texture, color and shape features) into the user's high-level concept. By taking the three-layer Feed-Forward Neural Network as the mapping function $\hat{f}$ and the back-propagation (BP) learning algorithm as the gradient-based search method in our Multiple Instance Learning framework, the neural network parameters such as the weights of all connections and biases of neurons are the parameters in γ that we want to learn (search). Specifically, the input layer has eight neurons with each of them corresponding to one low-level image feature. The output layer has only one neuron and its output indicates the extent to which an image segment meets the user's concept. The number of neurons at the hidden layer is experimentally set to eight. The biases to all the neurons are set to zero, and the used activation function in the neuron is Sigmoid Function. The BP learning method was applied with learning rate 0.1 and no momentum. The initial weights of the connections in the network are randomly set with relatively small values. The termination condition of the BP algorithm is based on $\left|MSE^{(k)} - MSE^{(k-1)}\right| < \alpha \times MSE^{(k-1)}$, where $MSE^{(k)}$ denotes the MSE at the k^{th} iteration and α is a small constant. In our experiments, α is set to 0.005.

5.4 CBIR System Description

Based on the proposed framework, we have constructed a content-based image retrieval system. Figure 2 shows the interface of this system. As can be seen from this figure, the query image is the image at the top-left corner. The user can press the 'Get' button to select the query image and press the 'Query' button to perform a query. The query results are listed from top left to bottom right in decreasing order of their similarities to the query image. The user can use the pull down list under an image to input his/her feedback on that image (Negative or Positive). After the feedback, the user can carry out the next query. The user's concept is then learned by the system in a progressive way through the user feedback, and the refined query will return a new collection of the matching images to the user.

5.5 Experimental Results

A number of experiments are conducted to test our proposed framework. Usually, it converges after 6 iterations of the user feedbacks. Also, in many cases, the user's most interested region of the query image can be discovered, and therefore the query performance can be improved.

Figure 2. The interface of the proposed CBIR system and query results by using a simple distance-based metric of image similarity

Figure 3. The query results after 5 iterations of user feedback

As shown in Figure 2, there is one horse on the lawn in the query image. Assume the horse object (not the lawn) is what the user is really interested in. Figure 3 shows the initial retrieval results using a simple distance-based metric of image similarity according to Equation (12). As can be seen from this figure, many retrieved images contain lawns or green mountains without any animal object in them. The reason why they are considered more similar to the query image is that they have regions (e.g., lawn) very similar to the lawn region of the query image. However, what the user really needs are the images with the horse object in them. By integrating the user's feedback with Multiple Instance Learning, the proposed

CBIR system can solve the above problem since the user can provide his/her relevant feedback to the system by labeling each image as Positive or Negative. In Figure 2, those images with bounding boxes are labeled Positive, while the others are labeled Negative by the user. Such feedback information is then fed into the Multiple Instance Learning method to discover user's real interest and thus capture the user's high-level concept. Figure 3 shows the query results after 5 iterations of user feedback. The image repository includes eight images with the horse object in them. In addition to the query image, all the remaining seven images are successfully retrieved by the system. Especially, all of them have higher ranks than other retrieved images. Another interesting result is that some of the retrieved images, such as the sunset images, have been retrieved because of their similarity in color to the horse region of the query image. On the other hand, all the images with the pure lawn or the green mountain are filtered out during the feedback and learning procedure. Therefore, this example illustrates that our proposed framework is effective in identifying the user's specific intention and thus can mine the user's high-level concepts.

6. CONCLUSIONS

In this paper, we presented a multimedia data mining framework to discover user's high-level concepts from low-level image features using Relevance Feedback and Multiple Instance Learning. Relevant Feedback provides a way to obtain the subjectivity of the user's high-level vision concepts, and Multiple Instance Learning enables the automatic learning of the user's high-level concepts. Especially, Multiple Instance Learning can capture the user's specific interest in some region of an image and thus can discover user's high-level concepts more precisely. In order to test the performance of the proposed framework, a content-based image retrieval (CBIR) system using Relevant Feedback and Multiple Instance Learning was developed and several experiments were conducted. The experimental results demonstrate the effectiveness of our framework.

ACKNOWLEDGMENT

Shu-Ching Chen gratefully acknowledges the support received from the National Science Foundation through grant CDA-9711582 at Florida International University.

REFERENCES

1. Aksoy, S., and Haralick, R.M. A Weighted Distance Approach to Relevance Feedback. *Proceedings of the International Conference on Pattern Recognition (ICPR00).*
2. Andrews, S., Hofmann, T., and Tsochantaridis, I. Multiple Instance Learning with Generalized Support Vector Machines. *The Learning Workshop.* (Snowbird, Utah, 2-5 Apr. 2002).
3. Auer, P. On Learning From Multi-instance Examples: Empirical Evaluation of a Theoretical Approach. *Proc. of 14th International Conference on Machine Learning*. (San Francisco, CA), 21-29.
4. Buckley, C., Singhal, A., Miltra, M. New Retriecal Approaches Using SMART: TREC4. *Text Retrieval Conference, Sponsored by National Institute of Standard and Technology and Advanced Research Projects Agency*. (Nov. 1995).
5. Carson, C., Belongie, S., Greenspan, H., and Malik, J. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, submitted to *PAMI*. (Available at: http://elib.cs.berkeley.edu/carson/papers/pami.html).
6. Chang, C.-H. and Hsu, C.-C. Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW. *IEEE Transactions on Knowledge and Data Engineering*, 11(4), 595-609.
7. Dietterich, T.G., Lathrop, R. H., and Lozano-Perez, T. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence Journal*, 89(1-2), 31-71.
8. Lu, Y., Hu, C.H., Zhu, X.Q., Zhang, H.J., and Yang, Q. A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems. *ACM Multimedia*. (2000), 31-37.
9. Marks II, R.J., Oh, S., Arabshahi, P., Caudell, T.P., Choi, J.J., and Song, B.G. Steepest Descent Adaptation of Min-Max Fuzzy If-Then Rules. *In Proc. IEEE/INNS International Conference on Neural Networks*. (Beijing, China, Nov. 1992).
10. Maron, O., and Lozano-Perez, T.. Multiple-Instance A Framework for Multiple-Instance Learning. *In Advances in Neural Information Processing System 10.* Cambridg, MA, MIT Press, 1998.
11. Ramon, J., and De Raedt, L. Multi-Instance Neural Networks," *ICML 2000 Workshop on Attribute-value and Relational Learning*. (2000).
12. Ray, S., and Page, D. Multiple-Instance Regression. *Proc. Of 18th International Conference on Machine Learning*. (San Francisco, CA), 425-432.
13. Rui, Y., Huang, T.S., Mehrotra, S. Content-based image retrieval with relevance feedback in MARS. *Proceedings of the 1997 International Conference on Image Processing (ICIP '97)* (3-Volume Set).

14. Rui, Y., and Huang, T.S. Optimizing Learning In Image Retrieval. *Proc. of IEEE Intl. Conf on Computer Vision and Pattern Recognition (CVPR00)*. (Hilton Head, SC, Jun. 2000), 236-243.

15. Wang, J., and Zucker, J.-D. Solving the Multiple-Instance Learning Problem: A Lazy Learning Approach. *Proc. Of 17th International Conference on Machine Learning*. (San Francisco, CA), 1119-1125.

16. Yang, C., and Lozano-Pérez, T. Image Database Retrieval with Multiple-Instance Learning Techniques. *Proceedings of the 16th International Conference on Data Engineering*. (2000), 233-243.

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

18. Zhang, Q., Goldman, S.A., Yu, W. and Fritts, J. Content-Based Image Retrieval Using Multiple-Instance Learning. *The Nineteenth International Conference on Machine Learning*. To be published, (Jul. 2002).

19. Zucker, J.-D., and Chevaleyre, Y. Solving Multiple-instance and Multiple-part Learning Problems with Decision Trees and Decision Rules. Application to the Mutagenesis Problem. *14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001*. (Ottawa, Canada, 7-9 Jun. 2001), 204-214.