Mining High-Level User Concepts with Multiple Instance Learning and Relevance Feedback for Content-Based Image Retrieval

Xin Huang[1], Shu-Ching Chen[1]*, Mei-Ling Shyu[2], and Chengcui Zhang[1]

[1] Distributed Multimedia Information System Laboratory
School of Computer Science, Florida International University
Miami, FL 33199, USA
{xhuan001, chens, czhang02}@cs.fiu.edu
http://dmis.cs.fiu.edu/

[2] Department of Electrical and Computer Engineering, University of Miami
Coral Gables, FL 33124 USA
shyu@miami.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 book chapter, 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 the 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 to show the feasibility and effectiveness of the proposed multimedia data mining framework.

1 Introduction

The availability of today's digital devices and techniques offers people more opportunities than ever to create their own digital images. Moreover, Internet has become the major platform to get, distribute and exchange digital image data. The rapid increase in the amount of image data and the inefficiency of traditional text-based image retrieval have created great demands for new approaches in image retrieval. As a consequence of such fast growth of digital image databases, the development of efficient search mechanisms has become more and more important. Currently, Content-Based Image Retrieval (CBIR) emerges and

* Shu-Ching Chen gratefully acknowledges the support received from the NSF through grants CDA-9711582 and EIA-0220562 at Florida International University.

O.R. Zaïane et al. (Eds.): Mining Multimedia and Complex Data, LNAI 2797, pp. 50–67, 2003.

dedicates to tackling such difficulties. CBIR is an active research area where the image retrieval queries are based on the content of multimedia data.

Recently, many efforts have been made to CBIR in order to personalize the retrieval engine. A significant problem in CBIR is the gap between semantic concepts and low-level image features. 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. The relevance feedback (RF) technique has been proposed and applied with the aim to discover the users' concept patterns by bridging the gap between semantic concepts and low-level image features as in [6,14,16].

In this book chapter, a multimedia data mining framework is proposed that can dynamically discover the concept patterns of a specific user to allow the retrieval of images by the user's most interested region. The discovering and adapting processes aim 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. Especially, the user's interest in special regions can be discovered from the images. The proposed multimedia data mining framework seamlessly integrates several data mining techniques. First, it takes advantages of the user relevance 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 [17]. Second, in order to identify the user's most interested region within the image, the Multiple Instance Learning 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). Under the Multiple Instance Learning scenario, 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 by Multiple Instance Learning, and tells the system to shift its focus of attention to the region of interest. The neural network technology is applied to map the low-level image features to the user's concepts. The parameters in the neural network are dynamically updated according to the user relevance feedback during the whole retrieval process to best represent the user's concepts. In this sense, it is similar to the re-weighting techniques in the RF approach.

The remainder of this chapter is organized as follows. Section 2 briefly introduces the background and related work in Relevance Feedback and Multiple Instance Learning. An overview of our proposed multimedia data mining framework is given in Section 3. Section 4 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 5. The experimental results are analyzed in Section 6. Section 7 gives the conclusion and future work.

2 Background and Related Work

2.1 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 [15]. 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.

Most RF techniques in CBIR are based on the most popular vector model [3, 15,18] used in information retrieval [8]. 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 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]. The basic idea of query point movement is quite straightforward. It tries to move the estimation of "ideal query point" towards positive example points and away from negative example points specified by the user according to his/her subjective judgments. The Rocchio's formula [14] is the frequently used technique to iteratively update the estimation of "ideal query point". The re-weighting techniques, however, take user's query example as the fixed "ideal query point" and attempt to estimate the best similarity metrics by adjusting the weight associated with each low-level image feature [1,5,15]. The basic idea is to give larger weights to more important dimensions and smaller weights to unimportant ones.

2.2 Multiple Instance Learning

The Multiple Instance Learning problem is a special kind of supervised machine learning problem, which is getting more attention recently in the field of machine learning and has been applied to many applications such as drug activity prediction, stock prediction, natural scene image classification, and content-based image retrieval.

Unlike the standard supervised machine learning where each object in the training examples is labeled and the problem is to learn a hypothesis that can predict the labels of the unseen objects accurately, in the scenario of Multiple Instance Learning, the labels of individual objects in the training data are not available; instead the labeled unit is a set of objects. A set of objects is called a *bag* and an individual object in a bag is called an *instance*. In other words, in Multiple Instance Learning, 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 to the unseen bags and instances. In this sense, the Multiple Instance Learning problem can be regarded as a special kind of supervised machine learning problem in the condition of incomplete labeling information. In the domain of Multiple Instance Learning, there are two kinds of labels, namely *Positive* and *Negative*. A label of an instance is either *Positive* or *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*.

The Multiple Instance Learning technique is originally used in the context of drug activity prediction. In this domain, the input object is a molecule and the observed result is whether the molecule binds to a target "binding site" or not. If a molecule binds to the target "binding site," we label it as *Positive*; else we label it *Negative*. A molecule has a lot of alternative conformations, and only one or a few of the different conformations of each molecule (bag) are actually bound to the binding site and produce the observed result, while the others typically have no effect on the binding. Unfortunately, the binding activity of a specific molecule conformation cannot be directly observed. Actually, only the binding activity of a molecule can be observed. Therefore, the binding activity prediction problem is a multiple instance learning problem. In this sense, each bag is a molecule and the instances of a bag (molecule) are the alternative conformations of the molecule [7] .

The applications of Multiple Instance Learning related to the topic of this chapter are natural scene image classification and content-based image retrieval. In the first application, a natural scene image usually contains a lot of different semantic regions and its semantic category is usually only determined by one or more regions in the image. There may be some regions which do not fit the semantic meaning of the category. For example, we have an image which contains a wide river and a hill beside it. This image can be classified into the "river" category because of the existence of the river. In this case, the hill has nothing to do with the classification. If the image classification system can discover this kind of fact and only consider the features of the river object when learning the classifier, better performance can be achieved than using the features of the whole image. Based on that basic idea, Maron et al. applied Multiple Instance Learning into natural scene image classification [10]. In their approach, each image is represented by a bag and the regions (subimages) in the image correspond to the instances in the bag. An image is labeled *Positive* if it somehow contains the concept of a specific semantic category (i.e., one of its regions contains the concept); otherwise it is labeled *Negative*. From the labeled training images, the concept can be learned by Multiple Instance Learning and the learned concept can be used for scene classification.

With the same idea in natural scene image classification, the Multiple Instance Learning can also be applied to CBIR. In CBIR, the user expresses the visual concept he/she is interested in by submitting a query image example representing the concept to the system. It is often the case that only one or more regions in the query example represent that concept and other objects are unrelated to it. Considering each object as an instance and the image as a bag, Multiple Instance Learning can discover the objects really related to the user concept. By filtering out the unrelated objects (which can be considered as "noise") and only applying the related objects in the query process, we can expect better query performance. Based on this idea, both [20] and [22] applied Multiple Instance Learning in CBIR.

In addition to the application of Multiple Instance Learning, a lot of research has been done in Multiple Instance Learning algorithms. Dietterich et al. [7] represented the target concept by an axis-parallel rectangle (APR) in the n-dimensional feature space and presented several Multiple Instance Learning algorithms for learning the axis-parallel rectangles. In [2], the authors proposed the MULTIINST algorithm for Multiple Instance Learning that is also an APR based method. In [10], the concept of Diversity Density was introduced and a two-step gradient ascent with multiple starting points was applied to find the maximum Diversity Density. Based on the Diversity Density, EM-DD algorithm was proposed [21]. In their algorithm, it assumed 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 learn the representative instances and maximize the Diversity Density simultaneously. [13] also used the EM method to do Multiple Instance Regression. Jun Wang et al. [19] explored the lazy learning approaches in Multiple Instance Learning. They developed two kNN-based algorithms: Citation-kNN and Bayesian-kNN. In [23], the authors tried to solve the Multiple Instance Learning problem with decision trees and decision rules. Jan Ramon et al. [12] proposed the Multiple Instance Neural Network.

3 Overview of the Proposed Multimedia Data Mining Framework

In this chapter, 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. For this purpose, we introduce a multiple instance feedback model that accounts for various concepts/responses of the user. 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. After getting user's relevance feedback, Multiple Instance Learning is applied to capture the objects the user is really interested in and the mapping between the low-level feature and high-level concept simultaneously. Each new query is chosen to achieve the user expectation more closely given the previous user responses.

In our multimedia data mining framework, the Multiple Instance Learning algorithm is a key part. It determines the performance of the framework in a significant degree. To meet this requirement, an open Multiple Instance Learning framework is designed, where an "open" framework means that different sub-algorithms may be plugged into the learning framework for different applications. Hence, it provides the opportunity to select the suitable sub-algorithm for a specific application to get the best performance in a reasonable scope. In our multimedia data mining framework, the multi-layer feedforward neural network and back-propagation algorithm are plugged into the Multiple Instance Learning framework.

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 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 [4] to segment an image into regions 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/her feedback.
2. In many cases, what the user is really interested in is just 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 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 chapter, the Multiple Instance Learning method is applied 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.

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 [20], we adopt the image segmentation method in [4] to partition the images in a more natural way; 2) In other Multiple Instance Learning based image retrieval systems such as [22], it is not very clear how the user interacts with the CBIR system to provide the training images and the associated labeling information for Multiple Instance Learning. 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 determine which dimensions of the feature vector are important by adaptively reweighing them through the neural network technique.

4 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 $O = \{O_i | O_i \in \Omega\}$ and their associated labels $L = \{L_i | L_i \in \Psi\}$, the problem of supervised learning is to find a mapping function $\hat{f} : \Omega \rightarrow \Psi$ 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, namely $Positive$ and $Negative$. 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 object space Φ, a label space $\Psi = \{1\,(Positive), 0\,(Negative)\}$, a set of n bags $B = \{B_i | B_i \in P(\Phi), i = 1...n\}$, where $P(\Phi)$ is the power set of Φ, and their associated labels $L = \{L_i | L_i \in \Psi\}$, the problem of Multiple Instance Learning is to find a mapping function $\hat{f} : \Phi \rightarrow \Psi$ so that the function $\hat{f}$ has the best approximation of the real unknown function f.*

4.1 Problem Definition

Let $T =< B, L >$ denote a training set where $B = \{B_i, i = 1, ..., n\}$ is the set of n bags in the training set, $L = \{L_i, i = 1, ..., n\}$ is 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} $(j = 1, ..., m_i)$. The function f is the real unknown mapping function that maps an instance to its label, and 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 a *Negative* label. Hence, the relationship between the functions f and f_{MIL} can be described in Figure 1.

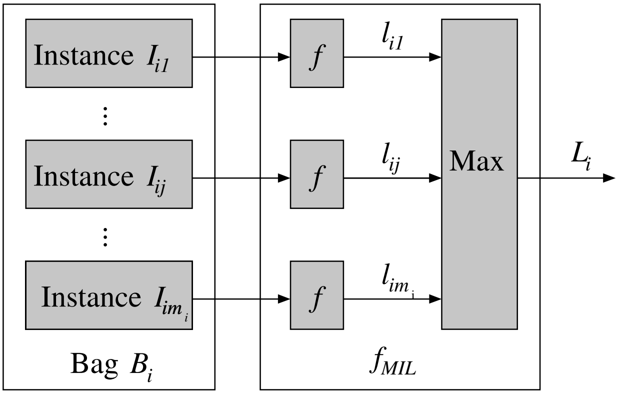

Fig. 1. Relationship between functions f and f_{MIL}.

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 $L_i{=}f_{MIL}(B_i){=}\max\limits_{j}\{l_{ij}\}{=}\max\limits_{j}\{f(I_{ij})\}$. The Multiple Instance Learning is to find a mapping function $\hat{f}$ with the best approximation to f given a training set $B = \{B_i\}$ and their corresponding labels $L = \{L_i, i = 1...n\}$. The corresponding approximation of f_{MIL} is $\hat{f}_{MIL}(B_i){=}\max\limits_{j}\{\hat{f}(I_{ij})\}$.

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

$$SE = \sum_{i=1}^{n}\big(L_i - \hat{f}_{MIL}(B_i)\big)^2 = \sum_{i=1}^{n}\big(L_i - \max_{j}\{\hat{f}(I_{ij})\}\big)^2 \tag{1}$$

Let $\gamma = \{\gamma_k, k = 1, ..., N\}$ 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:

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{n}\big(L_i - \max_{j}\{\hat{f}(I_{ij})\}\big)^2 \tag{2}$$

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 $E = \big(L_i - \max\limits_j\{\hat{f}(I_{ij})\}\big)^2$. In order to do that, the differentiation of the **max** function needs to be calculated first.

4.2 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

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

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

The differentiation of the **max** function can be written as:

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

4.3 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 $E = \big(L_i - \max\limits_j\{\hat{f}(I_{ij})\}\big)^2$ on the parameters $\gamma = \{\gamma_k\}$ of $\hat{f}$. The first partial derivative is as follows:

$$\frac{\partial E}{\partial \gamma_k} = \frac{\partial \big(L_i - \max\limits_j\{\hat{f}(I_{ij})\}\big)^2}{\partial \gamma_k} \tag{5}$$

$$= 2 \times \big(\max_j\{\hat{f}(I_{ij})\} - L_i\big) \times \frac{\partial \max\limits_j\{\hat{f}(I_{ij})\}}{\partial \gamma_k} \tag{6}$$

$$= 2 \times \big(\max_j\{\hat{f}(I_{ij})\} - L_i\big) \times \textstyle\sum_{j=1}^{m_i} \left(\frac{\partial \max\limits_j\{\hat{f}(I_{ij})\}}{\partial \hat{f}(I_{ij})} \times \frac{\partial\{\hat{f}(I_{ij})\}}{\partial \gamma_k} \right) \tag{7}$$

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

$$\frac{\partial E}{\partial \gamma_k} = 2 \times \big(\hat{f}(I_{is}) - L_i\big) \times \textstyle\sum_{j=1}^{m_i} \left(\frac{\partial \max\limits_j\{\hat{f}(I_{ij})\}}{\partial \hat{f}(I_{ij})} \times \frac{\partial\{\hat{f}(I_{ij})\}}{\partial \gamma_k} \right) \tag{8}$$

$$= 2 \times \big(\hat{f}(I_{is}) - L_i\big) \times \frac{\partial\{\hat{f}(I_{is})\}}{\partial \gamma_k} = \frac{\partial \big(L_i - \hat{f}(I_{is})\big)^2}{\partial \gamma_k} \tag{9}$$

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

$$\frac{\partial^n E}{\partial {\gamma_k}^n} = \frac{\partial^n \big(L_i - \max_j\{\hat{f}(I_{ij})\}\big)^2}{\partial {\gamma_k}^n} = \frac{\partial^n \big(L_i - \hat{f}(I_{is})\big)^2}{\partial {\gamma_k}^n} \tag{10}$$

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

$$\frac{\partial^{(\sum_k n_k)} E}{\prod_k \partial {\gamma_k}^{n_k}} = \frac{\partial^{(\sum_k n_k)} \big(L_i - \max_j\{\hat{f}(I_{ij})\}\big)^2}{\prod_k \partial {\gamma_k}^{n_k}} = \frac{\partial^{(\sum_k n_k)} \big(L_i - \hat{f}(I_{is})\big)^2}{\prod_k \partial {\gamma_k}^{n_k}} \tag{11}$$

4.4 Multiple Instance Learning to Traditional Supervised Learning

Similar to the analysis on Multiple Instance Learning problem in Section 4.1, the traditional supervised learning problem can also be converted to an unconstrained optimization problem as shown in Equation (9).

$$\overline{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{n} \big(L_i - \hat{f}(O_i)\big)^2 \tag{12}$$

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

$$\frac{\partial^n \big(L_i - \hat{f}(O_i)\big)^2}{\partial {\gamma_k}^n} \tag{13}$$

$$\frac{\partial^{(\sum_k n_k)} \big(L_i - \hat{f}(O_i)\big)^2}{\prod_k \partial {\gamma_k}^{n_k}} \tag{14}$$

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 $(i = 1, ..., n)$ in the training set, calculate the label of each instance I_{ij} belonging to it.
2. Select the instance with the 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 $\{O_i\}$ $(i = 1, ..., n)$ using all the instances I_{is} where $O_i = I_{is}$.

4. For each object O_i, construct a label L_{O_i} that is actually the label of bag B_i.
5. The Multiple Instance Learning problem with the input $\left(\{B_i\},\ \{L_i\}\right)$ is converted to the traditional supervised learning problem with the input $\left(\{O_i\},\ \{L_{O_i}\}\right)$.

After this transformation, the gradient-based search methods used in the traditional supervised 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($\boldsymbol{B},\ \boldsymbol{L}$)
Input : $B = \{B_i,\ i = 1, ..., n\}$ is the set of n bags in the training
set and $L = \{L_i,\ i = 1, ..., n\}$ is the set of labels where $L_i$ is the
label of bag $B_i$.
Output : $\gamma = \{\gamma_k,\ k = 1, ..., N\}$ is the set of parameters of the
mapping function $\hat{f}$ where $N$ is the number of parameters.

  1. Set initial values to parameters $\gamma_k$ in $\gamma$.
  2. If the termination criterion has not been met, go to Step 3; else
     return the parameter set $\gamma$ of function $\hat{f}$.
     /* The termination criterion can be based on MSE or the num-
     ber 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 super-
     vised learning to update the parameters in $\gamma$.
  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.

5 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, he/she is only interested in a region of the image. The image retrieval system proposed by Blobworld [4] 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.

To apply Multiple Instance Learning into CBIR, a necessary step before an actual image retrieval is to acquire a set of images as the training examples that are used to learn the user's target concept. In our method, the first set of training examples is 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 the user is only interested in one region of an image. In other words, there exists a function $f \in F : S \rightarrow \Psi$ 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 $\Psi = \{1\ (Positive),\ 0\ (Negative)\}$ 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, a metric based on color histogram comparison is applied to measure the similarity of two images. For each color, the two most significant bits of each R, G, B color component are extracted to compose a 6-bit color code [11]. The 6-bit code provides 64 bins. Each image can be converted to a histogram with 64 bins and therefore can be represented by a point in the 64-dimension feature space. The Manhattan distance between two points is used as the measurement of the dissimilarity between the two images represented by those two points respectively. Assume the color histograms of image A and image B are represented by two 64-dimension vectors $(a_1, a_2, ..., a_{64})$

and $(b_1, b_2, ..., b_{64})$ respectively. The dissimilarity (difference) between images A and B is defined as

$$D(A, B) = \sum_{i=1}^{64} |a_i - b_i| \tag{15}$$

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.

6 Experiments and Results

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.

6.1 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 [4] 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 then applied to select the best number of components in the 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 region.

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.

6.2 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 region 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 with 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 $|MSE^{(k)} - MSE^{(k-1)}| < \alpha \times MSE^{(k-1)}$, 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.

6.3 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.

6.4 Performance Analysis by a Query Example

In this section, a query example is conducted to illustrate how our CBIR system works. As shown in Figure 2, the query example is on the top-left corner. There is one tiger on the gray ground in the query image. Assume the tiger object (not the background) is what the user is really interested in. Figure 2 also shows the initial retrieval results using a simple color-histogram-based metric of image similarity according to Equation (12). As can be seen from this figure, many retrieved images have no tiger object in them. The reason why they are considered more similar to the query image is that they are similar in terms of the color distribution on the whole image. However, what the user really needs are the images with the tiger 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. 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 4 iterations of user feedback. Many more images containing the tiger object are successfully retrieved by the system. Especially, almost all of them have higher ranks than the other retrieved images. Another interesting result is that some of the retrieved images, such as the images containing horse on green lawn, have been retrieved although they are much different with the query image in terms of the color distribution on the whole image. The reason is that the horse object is more similar to the tiger object in the query image in terms of the color distribution on those objects. On the other hand, the irrelevant images with the similar color distribution on the whole images such as the building image and the cave image 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 to mine the user's high-level concepts.

A number of experiments have been conducted on our CBIR system. Usually, it converges after 4 or 5 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.

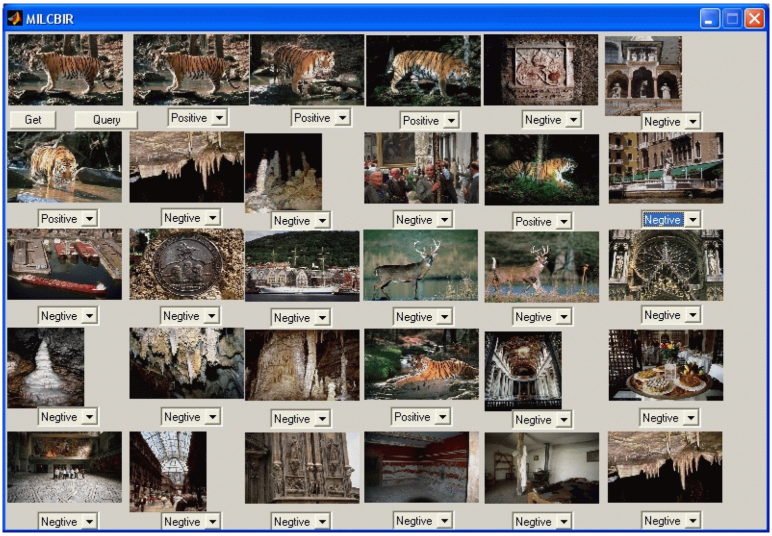

Fig. 2. The interface of the proposed CBIR system and initial query results

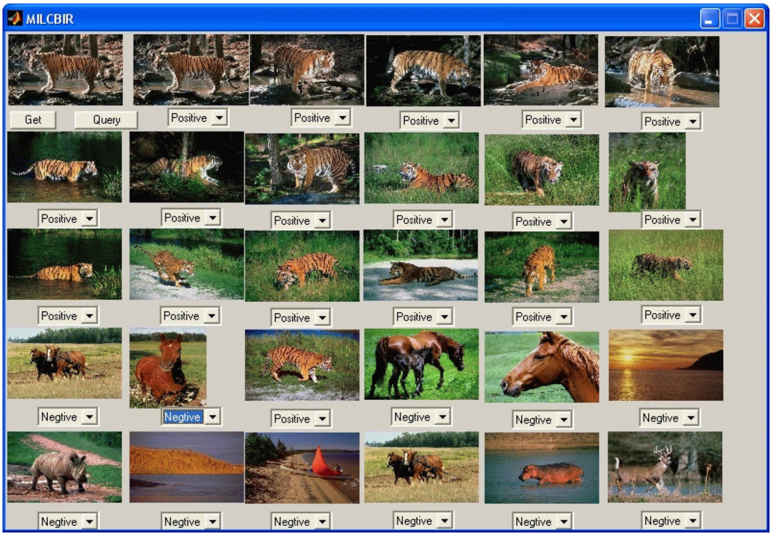

Fig. 3. The query results of our CBIR system after 4 iterations of user feedback

7 Conclusions

In this book chapter, we presented a multimedia data mining framework to discover user's high-level concepts from the 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.

References

1. Aksoy, S. and Haralick, R.M. : A Weighted Distance Approach to Relevance Feedback. Proceedings of the International Conference on Pattern Recognition, (2000) 812–815.
2. Auer, P.: On Learning From Multi-instance Examples: Empirical Evaluation of a Theoretical Approach. Proceedings of 14^{th} International Conference on Machine Learning. (San Francisco, CA), (1997) 21–29.
3. Buckley, C., Singhal, A., and Miltra, M.: New Retrieval Approaches Using SMART: TREC4. Text Retrieval Conference, Sponsored by National Institute of Standard and Technology and Advanced Research Projects Agency, (1995) 25–48.

4. Carson, C., Belongie, S., Greenspan, H., and Malik, J.: Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2002) 24(8), 1026–1038.
5. 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** (1999) 595–609.
6. Cox, I. J., Minka, T. P., Papathomas, T. V., and Yianilos, P. N.: The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments. IEEE Transactions on Image Processing –special issue on digital libraries, **9**(1) (2000) 20–37.
7. Dietterich, T.G., Lathrop, R. H., and Lozano-Perez, T.: Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence Journal, **89** (1997) 31–71.
8. Ishikawa, Y., Subramanya R., and Faloutsos, C.: Mindreader: Query Databases Through Multiple Examples. Proceedings of the 24th International Conference on Very Large Databases, (1998)
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. Proceedings of the IEEE/INNS International Conference on Neural Networks, Beijing, China, **3** (1992) 471–477.
10. Maron, O. and Lozano-Perez, T.: Multiple-Instance A Framework for Multiple-Instance Learning. In Advances in Neural Information Processing System 10. Cambridge, MA, MIT Press, (1998).
11. Nagasska, A. and Tanaka, Y.: Automatic Video Indexing and Full Video Search for Object Appearance. IFIP Trans. Visual Database Systems II, (1992) 113–127.
12. Ramon, J. and De Raedt, L.: Multi-Instance Neural Networks. Proceedings of the ICML 2000 Workshop on Attribute-value and Relational Learning, (2000)
13. Ray, S. and Page, D.: Multiple-Instance Regression. Proceedings of the 18^{th} International Conference on Machine Learning, (San Francisco, CA), (2001) 425–432.
14. Rocchio, J.J.: Relevance Feedback in Information Retrieval. The Smart System experiments in automatic document processing, Englewood Cliffs, NJ: Prentice Hall Inc. (1971) 313–323.
15. Rui, Y., Huang, T.S., and Mehrotra, S.: Content-based image retrieval with relevance feedback in MARS. Proceedings of the 1997 International Conference on Image Processing, (1997) 815–818.
16. Rui, Y., Huang, T.S., Ortega, M., and Mehrotra, S.: Relevance Feedback: A Power Tool in Interactive Content-Based Image Retrieval. IEEE Transaction on Circuits and Systems for Video Technology , Special Issue on Segmentation, Description, and Retrieval of Video Content, **18**(5) (1998) 644–655.
17. Rui, Y. and Huang, T.S.: Optimizing Learning In Image Retrieval. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, (2000) 236–243.
18. Salton, G. and McGill, M. J.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company, (1983).
19. Wang, J. and Zucker, J.-D.: Solving the Multiple-Instance Learning Problem: A Lazy Learning Approach. Proceedings of the 17^{th} International Conference on Machine Learning, (2000) 1119–1125.
20. Yang, C. and Lozano-Prez, T.: Image Database Retrieval with Multiple-Instance Learning Techniques. Proceedings of the 16^{th} International Conference on Data Engineering, (2000) 233–243.

21. 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.
22. Zhang, Q., Goldman, S.A., Yu, W., and Fritts, J.: Content-Based Image Retrieval Using Multiple-Instance Learning. Proceedings of the 9^{th} International Conference on Machine Learning, (2002).
23. 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. Proceedings of the 14^{th} Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2001, (2001) 204–214.