INCORPORATING REAL-VALUED MULTIPLE INSTANCE LEARNING INTO RELEVANCE FEEDBACK FOR IMAGE RETRIEVAL

Xin Hunag[1] *Shu-Ching Chen*[1] *Mei-Ling Shyu*[2]

[1]Distributed Multimedia Information System Laboratory, School of Computer Science
Florida International University, Miami, FL 33199, USA
[2]Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA

ABSTRACT

This paper presents a content-based image retrieval (CBIR) system that incorporates real-valued Multiple Instance Learning (MIL) into the user relevance feedback (RF) to learn the user's subjective visual concepts, 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 the user. RF provides a way to obtain the subjectivity of the user's high-level visual concepts, and MIL enables the automatic learning of the user's high-level concepts. The user interacts with the CBIR system by relevance feedback in a way that the extent to which the image samples retrieved by the system are relevant to the user's intention is labeled. The system in turn applies the MIL method to find user's most interested image region from the feedback. A multilayer neural network that is trained progressively through the feedback and learning procedure is used to map the low-level image features to the high-level concepts.

1. INTRODUCTION

A good CBIR system should have the capability of learning the users' visual concepts and adapting to them accordingly since different users may have different visual concepts or different intentions even on the same query image. For instance, by submitting a query image including multiple objects such as river and mountain, one user may look for those images with mountains; while another user may be more interested in the river and do not care about the mountain. The CBIR system, however, could not understand this kind of user subjectivity directly from the query image. In this paper, we propose a CBIR system that can dynamically learn the visual concept of a specific user from the user's relevance feedback. Especially, it can simultaneously find out the user's most interested image region and learn the mapping relation between the low-level image features of the images and the user's concept. The proposed CBIR system has the following three features.

First, the user has the opportunity to interact with the system by user Relevance Feedback (RF) during the retrieval process. Relevance feedback is an interactive and iterative process between the user and the retrieval system to bridge the gap between low-level image features and high-level concepts [1] and has been an active research field in CBIR [2][3]. Different from many other CBIR systems that use RF in a way that the user can give only positive or negative feedback to a sample image, our CBIR system provides a more precise RF mechanism by allowing the user to indicate the percentage that a sample image meets the user's concepts. This proposed CBIR system then undertakes the learning process to best match the user's concepts from the user's relevance feedback. This process iterates until a satisfactory result is obtained for the user.

Second, in our proposed CBIR system, the real-valued Multiple Instance Learning (MIL) is integrated into the query refining process to learn the region of interest from user relevance feedback and to tell the system to shift its focus of attention to that region. In the scenario of MIL, the labels of individual instances in the training data are not available; instead the labeled unit is a set of instances (*bag*). In other words, a training example is a labeled bag. The goal of learning is to obtain a hypothesis from the training examples that generates labels to the unseen bags. The MIL technique is originally used in molecule categorization in the context of drug activity prediction where each molecule (bag) is represented by a bag of possible conformations (instances) [6]. In image retrieval, each image is viewed as a bag of image regions (instances). However, though the user may only be interested in a specific region (instance) of an image (bag), he/she can only give relevance feedback on the whole image (bag). In this context, MIL can be applied to learn the region of interest. In our proposed CBIR system, the traditional 2-valued (Positive and Negative) MIL is extended to the real-valued MIL since the user's label on an image is actually a real value in the closed interval [0,1].

Third, 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 remaining of this paper is as follows. Section 2 introduces the details of the Multiple Instance Learning techniques used in our CBIR system. Section 3 describes the proposed CBIR system using user relevance feedback and the real-valued Multiple Instance Learning. The experimental results are analyzed in Section 4. Section 5 concludes this paper.

2. REAL-VALUED MULTIPLE INSTANCE LEARNING

In original Multiple Instance Learning, the label of each bag is either 1 (*Positive*) or 0 (*Negative*). A bag is labeled *Positive* if the bag has one or more positive instances and is labeled *Negative* if and only if all its instances are negative. The goal of learning is to generate a hypothesis from the labeled bags to

predict the labels of unseen bags. The original Multiple Instance Learning problem can be defined in formal way as follows.

Definition 1. *Given the instance space* α, *the bag space* β, *the label space* $Q=\{0(Negative),1(Positive)\}$, *a set of training examples* $T=\langle B,L\rangle$ *where* $B=\{B_i \mid B_i\in\beta,\ i=1...n\}$ *is a set of n bags and* $L=\{L_i \mid L_i\in Q, i=1...n\}$ *is the set of their associated labels with* L_i *being the label of* B_i, *the problem of Multiple Instance Learning is to generate a hypothesis* $h:\beta\to Q=\{0,1\}$ *which can predict the labels of unknown bags accurately.*

The original Multiple Instance Learning problem can be extended to a real-valued Multiple Instance Learning problem by converting the discrete label space $Q=\{0,1\}$ to a continuous label space $Q^*=[0,1]$ where the label of a bag indicates the degree in which the bag is Positive. Label "1" means the bag is Positive one hundred percent and label "0" indicates that the bag is impossible to be Positive. After this conversion, the goal of the learner changes to generate a hypothesis $h_B:\beta\to Q^*=[0,1]$ that can be described as follows.

Definition 2. *Given the instance space* α, *the bag space* β, *the label space* $Q^*=[0,1]$, *a set of training examples* $T=\langle B,L\rangle$ *where* $B=\{B_i \mid B_i\in\beta,\ i=1..n\}$ *is a set of* n *bag and* $L=\{L_i \mid L_i\in Q^*, i=1..n\}$ *is the set of their associated labels with* L_i *being the label of* B_i, *the problem of real-value labeled Multiple Instance Learning is to generate a hypothesis* $h_B:\beta\to Q^*=[0,1]$ *which can predict the labels of unknown bags accurately.*

Actually, each instance in a particular bag has a label in the closed interval $[0,1]$, which represents the degree of that instance being Positive, although it is unknown. Given the labels of all the instances in a bag, the label of the bag (i.e., the degree of the bag being Positive) can be represented by the maximum of the labels of all its instances. In other words, $L_i = \underset{j}{MAX}\{l_{ij}\}$ where the label L_i is the label of bag B_i and l_{ij} is the label of the j^{th} instance I_{ij} in B_i. Let $h_I:\alpha\to Q^*=[0,1]$ denote the hypothesis that predicts the label of an instance. The relationship between hypotheses h_B and h_I can be depicted in Equation (1)

$$L_i = h_B(B_i) = \underset{j}{MAX}\{l_{ij}\} = \underset{j}{MAX}\{h_I(I_{ij})\}. \qquad (1)$$

In our proposed real-valued Multiple Instance Learning framework, the Minimum Square Error (MSE) criterion Is adopted. That is, we try to learn the hypotheses $\hat{h}_B$ and $\hat{h}_I$ to minimize the function shown in Equation (2).

$$SE=\sum_{i=1}^{n}\left(L_i-\hat{h}_B(B_i)\right)^2=\sum_{i=1}^{n}\left(L_i-\underset{j}{MAX}\{\hat{h}_I(I_{ij})\}\right)^2 \qquad (2)$$

In addition, in our algorithm, the Multilayer Feed-Forward Neural Network is used as the hypothesis $\hat{h}_I$ and the Back-propagation (BP) learning method is used to train the neural network to minimize SE. Hence, we need to calculate the first partial derivative of function $E=\left(L_i-\underset{j}{MAX}\{\hat{h}_I(I_{ij})\}\right)^2$ on the neural network parameters $\gamma=\{\gamma_k\}$.

In order to differentiate the function E, we first need to calculate the differentiation of the MAX function. As mentioned in [7], the differentiation of the MAX function results in a 'pointer' that specifies the source of the maximum. Let

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

where $U(\cdot)$ is the 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:

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

Equation (4) gives the formula to differentiate the MAX function and thus the first partial derivative of function E on the neural network parameters $\gamma=\{\gamma_k\}$ can be calculates as follows:

$$\begin{aligned}\frac{\partial E}{\partial\gamma_k}&=\frac{\partial\left(L_i-\underset{j}{MAX}\{\hat{h}_I(I_{ij})\}\right)^2}{\partial\gamma_k}\\ &=2\left(\underset{j}{MAX}\{\hat{h}_I(I_{ij})\}-L_i\right)\\ &\quad\times\sum_j\left(\frac{\partial\underset{j}{MAX}\{\hat{h}_I(I_{ij})\}}{\partial\hat{h}_I(I_{ij})}\times\frac{\partial\{\hat{h}_I(I_{ij})\}}{\partial\gamma_k}\right)\\ &=2\left(\hat{h}_I(I_{is})-L_i\right)\times\frac{\partial\{\hat{h}_I(I_{is})\}}{\partial\gamma_k}\end{aligned} \qquad (5)$$

where it is assumed that the s[th] instance of bag B_i has the maximum value, i.e., $\hat{h}_I(I_{is})=\underset{j}{MAX}\{\hat{h}_I(I_{ij})\}$.

Equation (5) provides a way to calculate the first partial derivative of the function SE on the neural network parameters where $\frac{\partial\{\hat{h}_I(I_{ij})\}}{\partial\gamma_k}$ is easy to calculate although the particular formula depends on the specific structure of the neural network used in practice. In addition, it is worth to point out that the instance with the maximum label in each bag (i.e., I_{is}) may be different at each iteration during the neural network training procedure, whereas the fundamental formula remains the same.

3. THE PROPOSED CBIR SYSTEM

In a CBIR system, 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 (an object) of the image. Our proposed CBIR system first segments the image into multiple regions and then uses the user's relevance feedback and the proposed real-valued 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.

The automatic segmentation method proposed in the Blobworld system [8] is used in our system to segment the image into multiple regions. In Blobworld, 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 the Gaussian Mixture model. Three texture features, three color features and two shape features are extracted from each region after image segmentation, and are used as the low-level feature vector to represent each region.

It is assumed that the user is only interested in one region of an image. In other words, there exists a mapping between a region of an image and the user's concept. Our system uses the Multilayer Feed-Forward Neural Network to map a low-level feature vector a real value in [0,1] which represents how much the region meets the user's concept. The extent to which an image belonging to the user's concept is the maximum one of all its regions. Therefore, an image can be viewed as a bag and its regions are the instances of the bag in Multiple Instance Learning (MIL). During the image retrieval procedure, the user's feedback can provide the labels for the retrieved images based on the user's own concept about the images. Since the labels are assigned to the individual images, not on individual regions, the image retrieval task can be viewed as a MIL task aiming to learn the neural network, identify the user's most interested region and capture the user's high-level concept from the low-level features.

At the beginning of retrieval, the learning method is not applicable since there are no training examples available. Hence, we use a simple distance-based metric to measure the similarity of two images. Assume Image A consists of n regions and Image B consists of m regions, where $A = \{A_i\}\ \ (i = 1, \cdots, n)$ and $B = \{B_j\}\ \ (j = 1, \cdots, m)$. The difference between Images A and B is defined as:

$$D(A, B) = \underset{1 \le i \le n, 1 \le j \le m}{Min} \left\{ \left\| A_i - B_j \right\| \right\} \qquad (6)$$

where $\left\| A_i - B_j \right\|$ is the Euclidean distance between two feature vectors of region A_i and B_j. This 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 (6), the user can give their feedbacks by labeling each retrieved image and a set of training examples can be constructed based on the user feedbacks. Then the real-valued MIL is applied to train the neural network and the trained neural network generates a similarity measurement for each image in the database and retrieves the most similar ones to the user. The feedback and learning are executed 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.

Compared with other CBIR systems using the RF techniques, our method differs in the following two aspects. 1) Unlike other recent efforts in the RF techniques that deal with the global image properties of the query image, our method first segments the image into regions that roughly correspond to objects and takes those regions as basic processing unit. 2) In many cases, what the user is really interested in is just a region of the image. However, the user can only provide the feedback on the whole image. How to effectively identify the user's most interested region (object) to more precisely capture the user's high-level concepts based on the feedback on the whole image have not received much attention yet. To achieve that, our system applies the real-valued MIL technique to discover the user's interested region from user relevance feedback.

Compared with other MIL-based CBIR systems, our system has the following advantages. 1) Instead of manually dividing each image into many overlapping regions [4], we adopt the Blobworld image segmentation method [8] to partition the images in a more natural way. 2) [5] also proposed an approach to apply MIL into content-based image retrieval. However, it is not very clear how the user interacts with the CBIR system to provide the training images and the associated labeling information. While in our system, the user gives a real-valued label on the sample images through relevance feedback and thus provides the training examples for MIL. This way is very efficient since the user can easily generate the examples from the initial retrieved results. It also makes the retrieval process more precise because the retrieved images have similar features/contents with the query image and the users may have different focuses of attention. By putting negative feedback on those images that do not meet user's specific concepts despite of the similar image features, the system can better distinguish user's real needs from the "noisy" or unrelated information via MIL. 3) In our system, the neural network is used to map the low-level image features to the user's concepts. The parameters of the neural network are adaptively updated during the feedback process. This has the similar function of the feature reweighing in the traditional RF techniques.

4. EXPERIMENTAL RESULTS

We select 2,500 images of various categories from the Corel image library to build our own image repository for the system testing purpose. In our experiments, a three-layer Feed-Forward Neural Network is used. Specifically, the input layer has eight neurons with each of them corresponding to one of the eight low-level image features. 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.

We have constructed a CBIR system and conducted a number of experiments. Figure 1 shows the interface of this system and the initial retrieval results using the distance-based metric of image similarity defined in Equation (6). The query image is at the top-left corner. The interface allows the user to press the 'Get' button to select the query image or the 'Query' button to execute 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 also use the scroll bar under each image to input his/her feedback on that image and carry out the next round of retrieval. The user's concept is then learned by the

system in a progressively way through the user feedback, and the refined query will return a new collection of the matching images to the user. It needs to be noted that it usually converges after 5 iterations of the relevant feedbacks and in many cases, the user's most interested region (object) of the query image can be discovered from our experiments. Therefore, the query performance can be improved.

Figure 1. The interface of the proposed CBIR system and the query results by using a simple distance-based metric.

Figure 2. The query results after 5 iterations of user feedback.

As can be seen from Figure 1, there is a horse on the lawn in the query image and it is assumed that the object the user is really interested in is the horse (not the lawn). In the initial retrieved images, many of them contain lawns or green mountains without any animal object in them. These images are retrieved because they have regions very similar to the lawn region of the query image. However, what the user really needs is the images with the horse object in them. Figure 2 shows the retrieved images after 5 iterations of user feedback. In fact, the image repository consists of eight images with the horse object and all of them have been successfully retrieved by the system and especially with higher similarities, whereas those images with only lawn or the green mountain are filtered out during the feedback and learning procedure. This demonstrates the effectiveness of our proposed CBIR system with the real-valued MIL and RF.

5. CONCLUSION

A CBIR system with real-valued Multiple Instance Learning and Relevance Feedback to learn user's high-level concepts from low-level image features is developed. The real-valued Multiple Instance Learning allows the user to specify the degree of preference for his/her specific interested region in an image to more precisely capture his/her high-level concepts. In order to test the performance of the proposed CBIR system, several experiments were conducted and the experimental results demonstrate the effectiveness of our proposed CBIR system.

6. ACKNOWLEDGEMENT

This research was supported in part by NSF CDA-9711582 and NSF EIA-0220562.

REFERENCES

[1] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra. "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,* 8(5): pp. 644-655, September 1998.

[2] S. Aksoy and R.M. Haralick, "A Weighted Distance Approach to Relevance Feedback," *Proceedings of the International Conference on Pattern Recognition (ICPR00)*.

[3] C.-H. Chang and C.-C. Hsu, "Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, No. 4, pp. 595-609, July/August 1999.

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

[5] Q. Zhang, S. A. Goldman, W. Yu and J. Fritts "Content-Based Image Retrieval Using Multiple-Instance Learning," *Proceeding of the Nineteenth International Conference on Machine Learning*, July 2002.

[6] T.G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the Multiple-Instance Problem with Axis-Parallel Rectangles," *Aritificial Intelligence Journal*, 89, pp. 31-71, 1997

[7] R.J. Marks II, S. Oh, P. Arabshahi, T.P. Caudell, J.J. Choi, and B.G. Song, *"Steepest Descent Adaptation of Min-Max Fuzzy If-Then Rules,"* In Proc. IEEE/INNS International Conference on Neural Networks, Beijing, China, November 1992.

[8] C. Carson, S. Belongie, H. Greenspan, and J. Malik, *"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.http://elib.cs.berkeley.edu/carson/papers/pami.htm.