Multiple Object Retrieval for Image Databases Using Multiple Instance Learning and Relevance Feedback

Chengcui Zhang[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
{czhang02, chens}@cs.fiu.edu
[2]*Department of Electrical and Computer Engineering*
University of Miami, Coral Gables, FL 33124, USA
shyu@miami.edu

Abstract

This paper proposes a method to effectively discover users' concept patterns when multiple objects of interests (e.g., foreground and background objects) are involved in content-based image retrieval. The proposed method incorporates Multiple Instance Learning into the user relevance feedback in a seamless way to discover where the user's most interested objects/regions and how to map the local features of that region(s) to user's high-level concepts. A three-layer neural network is used to model the underlying mapping progressively through the feedback and learning procedure.

1. Introduction

The subjectivity of human perception of visual content plays an important role in content-based image retrieval (CBIR) systems. A fixed image similarity measure cannot meet the need to adapt to different focuses of attention of different users. The relevance feedback and region-based image retrieval are two techniques used to deal with this issue. The region-based retrieval systems segment an image into several homogenous regions, and then the features for each region can be extracted and compared. Relevance feedback (RF) [1] is an interactive process in which the user judges the quality of the retrieval results returned by the system. The user feedback information is then used to refine the original query. Recently, the research in integrating these two major techniques has gained many attentions. The representative is the RF based Multiple Instance Learning (MIL) mechanism proposed in [3, 4] which integrates RF and single object-based retrieval seamlessly.

In this paper, we propose a method that can dynamically discover the visual concept of a specific user from the user's relevance feedback when multiple objects of interest are involved in that user's focus of attention. Especially, it can simultaneously find out the multiple objects/regions of the user's interests and learn the mapping between the local image features of those objects and the user's concept. This method has the following distinctive features.

First, 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. Our method extends the existing MIL system [4] in a way that the user can provide feedback information to multiple objects instead of one, and the multiple objects of interests can be discovered simultaneously by feedback information fusion. In the scenario of MIL, each image is viewed as a bag of image regions (instances). The labels (relevant or irrelevant) of the individual regions in the training data are not available; instead the labeled unit is a set of instances (images). In other words, a training example is a labeled image. The goal of learning is to obtain a hypothesis from the training examples that generates labels to the unseen images. The original MIL technique has the assumption that the user's concept can be represented as a single "best" object. However, the discovery of multiple objects of interests is also very common and it is more natural to have one visual concept corresponding to more than one significant object. For example, one user may look for those images with red cars parked on the grassland; while another user may be more interested in red cars running on the highway. Second, 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 feedback to best represent the user's concepts. Third, a fast and unsupervised image segmentation method called WavSeg is proposed in this paper to automate the process of segmenting the image into multiple regions. The color and texture features are collected for each image region to form a feature vector for each instance (region) in a bag (image).

0-7803-8603-5/04/$20.00 ©2004 IEEE

The remaining of this paper is as follows. Section 2 introduces the Multiple Instance Learning techniques. Section 3 describes the image segmentation and feature extraction. Section 4 describes the retrieval process and the fusion of the relevance feedback information. The experimental results are analyzed in Section 5. Section 6 concludes this paper.

2. Multiple instance learning

In Multiple Instance Learning, the label of each bag is either *Positive* or *Negative*. A bag is labeled *Positive* if the bag has at least one positive instance and is labeled *Negative* if and only if all its instances are negative. The goal of learning is to generate a mapping function from the training data set to predict the labels of the unseen bags.

Definition 1. Given the instance space μ, the bag space ν, the label space $K = [0,1]$, a set of training examples $T = \langle B, L \rangle$ where $B = \{ B_i \mid B_i \in \nu,\ i = 1...n \}$ is a set of n bags and $L = \{ L_i \mid L_i \in K, i = 1...n \}$ 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 $h_B : \nu \to K = [0,1]$ 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 (Label 0 means *Negative.*), 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 $h_I : \mu \to K = [0,1]$ 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 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).

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

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 E.

3. Image segmentation and feature extraction

3.1. WavSeg: unsupervised segmentation

Instead of manually dividing each image into many overlapping regions [2], in this study, we propose a fast yet effective image segmentation method called WavSeg to partition the images. In Wavseg, a wavelet analysis in concert with the SPCPE algorithm [5] is used to segment an image into regions. By using wavelet transform and choosing proper wavelets (Daubechies wavelets), the high-frequency components will disappear in larger scale subbands and therefore, the possible regions will be clearly evident. In our experiments, the images are pre-processed by Daubechies wavelet transform because it is proven to be suitable for image analysis. The decomposition level is 1. Then by grouping the salient points from each channel, an initial coarse partition can be obtained and passed as the input to the SPCPE segmentation algorithm. Actually, even the coarse initial partition generated by wavelet transform is much closer to some global minima in SPCPE than a random initial partition, which means a better initial partition will lead to better segmentation results. In addition, wavelet transform can produce other useful features such as texture features in addition to extracting the region-of-interest within one entry scanning through the image data. Based on our initial testing results, the wavelet based SPCPE segmentation framework (WavSeg) outperforms the random initial partition based SPCPE algorithm in average. It is worth to point out that WavSeg is fast. The processing time for a 240*384 image is only about 0.33 sec in average.

3.2. Image feature extraction

Both the local color and local texture features are extracted for each image region.

Color Features: HSV color space and its variants are proven to be particularly amenable to color image analysis. Therefore, we quantize the color space using color categorization based on H S V value ranges. Twelve representative colors are identified. They are black, white, red, red-yellow, yellow, yellow-green, green, green-blue, blue, blue-purple, purple, and purple-red. The Hue is divided into five main color slices and five transition color slices. Each transition color slice such as yellow-green is considered in both adjacent main color slices. We disregard the difference between the bright chromatic colors and the chromatic colors. Each transition color slice is treated as a separate category instead of being combined into both adjacent main color slices. A new category "gray" is

added so that there are totally thirteen color features for each region in our method.

Texture Features: One-level wavelet transformation using Daubechies wavelets are used to generate four subbands of the original image. They include the horizontal detail sub-image, the vertical detail sub-image, and the diagonal detail sub-image. For the wavelet coefficients in each of the above three subbands, the mean and variance values are collected respectively. Therefore, totally six texture features are generated for each image region in our method.

4. Learning and retrieval process

In the content-based image retrieval process, 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 according to some similarity measures. However, in many cases, when a user submits a query image, what the user is really interested in is just one or two region(s) of the image. For example, "Find all the images that contain a brown horse object and a white horse object." In this study, we target the retrieval of multiple objects of interests by integrating multiple instance learning into the user relevance feedback. We also realized that the number of user interested objects is usually about 2~3 (If more than that, the whole image query is more appropriate.), and therefore, the two-object retrieval scenario is used in this study to illustrate the basic idea. Our proposed method first segments the image into multiple regions by using WavSeg and then uses the user's relevance feedback and Multiple Instance Learning to automatically capture the user-interested regions during the query refining process. Another advantage of our method is that the underlying mapping between the local visual feature vectors of the regions of interests and the user's high-level concept can be progressively discovered through the feedback and learning procedure.

Taking the two-object retrieval scenario as an example, there exist two mapping functions for objects 1 and 2 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 to a real value in [0, 1], which represents how much the region meets the user's concept. The extent to which an image belongs 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. During the image retrieval procedure, the users are asked to provide relevance feedback (relevant/irrelevant) to the whole image for each interested object (objects 1 and 2). For each object of interest, there are a set of positive relevant images as well as a set of negative images. Since the labels are assigned to the individual images, not on the individual regions, the image retrieval task can be viewed as a MIL task. In the two-object retrieval case, two neural networks are learned, which can identify the user's two most interested regions and capture the user's high-level concepts from the low-level features.

At the beginning of the .retrieval, the learning method is not available since there are no training examples. Hence, we use a simple distance-based metric to measure the similarity of two images. Assume Image Q is the query image and consists of nq regions and Image I consists of ni regions, where $Q=\{Q_i\}(i=1,\cdots,nq)$ and $I=\{I_j\}(j=1,\cdots,ni)$. The difference between Images $I1$ and $I2$ is defined as:

$$Dist(Q,I)=\sum_{1\le i\le nq} \underset{1\le j\le ni}{Min}\left\{\left\|Q_i-I_j\right\|\right\} \quad (3)$$

Upon the first round of retrieving those "most similar" images according to Equation (3), users can give their feedbacks by labeling each retrieved image, and a set of training examples can be constructed for each object of interest based on the user feedbacks. Then the MIL is applied to train the neural networks for the two objects of interests. Each image in the database will be passed as an input to the two trained neural networks respectively, and the outputs for each image are two similarity scores, one for each object of interest. The retrieval system will rank the images according to the similarity function which is a combination of the two scores, and present the most similar images to the user. Currently, we use the sum of the two scores as the similarity function. However, other methods of combination and fusion can also be tested. The feedback and learning are executed iteratively, and the capturing of user's high-level concept is refined until the user satisfies.

5. Experimental results

We select 2,100 images of various categories from the Corel image library to build our testing image database. In our experiments, a three-layer Feed-Forward Neural Network is used. Specifically, the input layer has nineteen neurons with each of them corresponding to one of the nineteen 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 nineteen.

Figure 1 shows the two-object retrieval interface of this system and the initial retrieval results using the similarity function defined in Equation (3). The query image is at the top-left corner. 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 two pull-down menus under each image to input his/her feedback on that image and carry out the next round of retrieval. The first pull-down menu contains the feedback for object 1, while the second pull-down menu collects the feedback for object 2. The user's concept is then learned 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 the user's two most interested regions can be discovered within 4 iterations in most cases.

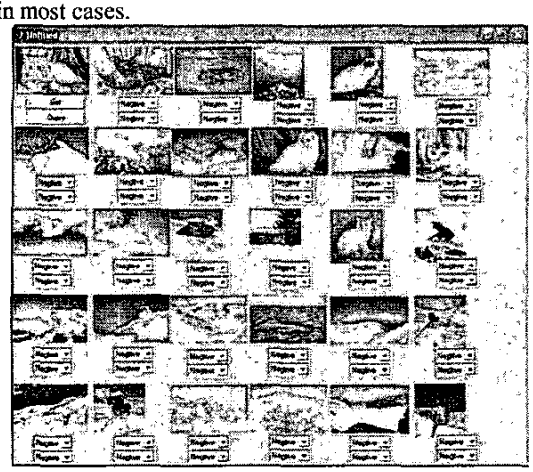

Figure 1. The CBIR retrieval interface and the initial query results

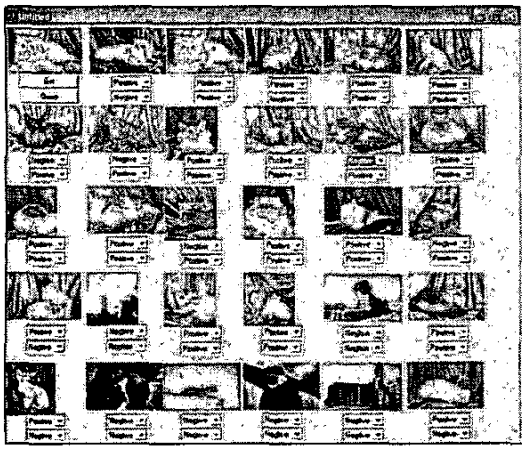

Figure 2. The query results after 4 iterations of user feedback

As shown in Figure 1, assume the two objects of interests are "white-yellow cat" and "blue-tone silken background." In the initial retrieved images, many of them contain snow scenes with blue skies without any of the two required objects in them. Figure 2 shows the retrieved images after 4 iterations of user feedback, and 24 out of 30 retrieved images contain both of the two query objects. We also conducted a number of other experiments on different image categories such as horses, mountain scenes, snow scenes, leopards, apes, owls, and race cars. The averaged accuracy within the top 30 retrieved images is around 70%, which demonstrates the effectiveness of our multiple object retrieval method using MIL and RF.

6. Conclusions

In this paper, we propose a method to discover user's high-level concepts from the low-level image features using RF and MIL. Compared with other MIL-based CBIR systems, our system has the following advantages: 1) Instead of manually dividing each image into many overlapping regions, we proposed the WavSeg image segmentation method to partition the images in a more natural way; 2) Instead of discovering one single object of interest, the proposed method can deal with the multiple object retrieval scenarios in which 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; and 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.

7. Acknowledgement

For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562, NSF HRD-0317692, and the office of the Provost/FIU Foundation. For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260.

8. References

[1] Y. Rui, T.S. Huang, et al., "Relevance Feedback: A Power Tool in Interactive Content-based Image Retrieval," *IEEE Trans. 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] C. Yang and T. Lozano-Pérez, "Image Database Retrieval with Multiple-Instance Learning Techniques," *Proc. of the 16th Intl. Conf. on Data Engineering*, pp. 233-243, 2000.

[3] Q. Zhang, S. A. Goldman, W. Yu and J. Fritts "Content Based Image Retrieval Using Multiple-Instance Learning," *Proc. of the 19th Intl. Conf. on Machine Learning*, 2002.

[4] X. Huang, S.-C. Chen, and M.-L. Shyu, "Incorporating Real-Valued Multiple Instance Learning Into Relevance Feedback For Image Retrieval," *Proc. of the IEEE Intl. Conf. on Multimedia & Expo (ICME)*, vol. I, pp. 321-324, 2003.

[5] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "An Indexing and Searching Structure for Multimedia Database Systems," *Proc. of the IS&T/SPIE Conf. on Storage and Retrieval for Media Databases 2000*, pp. 262-270, 2000.