A Dynamic User Concept Pattern Learning Framework for Content-Based Image Retrieval

Shu-Ching Chen, *Senior Member, IEEE*, Stuart H. Rubin, *Senior Member, IEEE*, Mei-Ling Shyu, *Senior Member, IEEE*, and Chengcui Zhang, *Member, IEEE*

***Abstract*—A rapid increase in the amount of image data and the inefficiency of traditional text-based image retrieval systems have served to make content-based image retrieval an active research field. It is crucial to effectively discover users' concept patterns through an acquired understanding of the subjective role played by humans in the retrieval process for such systems. A learning and retrieval framework is used to achieve this. It seamlessly incorporates multiple instance learning for relevant feedback to discover users concept patterns—especially in the region of greatest user interest. It also maps the local feature vector of that region to the high-level concept pattern. This underlying mapping can be progressively discovered through feedback and learning. The user guides the retrieval systems learning process using his/her focus of attention. Retrieval performance is tested to establish the feasibility and effectiveness of the proposed learning and retrieval framework.**

***Index Terms*—Content-based image retrieval (CBIR), multiple instance learning, neural network, relevance feedback.**

I. Introduction

THE VOLUME of multimedia data—typically image data—is increasing rapidly and shows every sign of continuing to do so. As a consequence, new techniques need to be discovered for efficient image retrieval. Content-based image retrieval (CBIR) has emerged and is dedicated to tackling such difficulties. The objective of a CBIR system is to enable the user to efficiently find and retrieve those images that he/she wants from a database while the image search is based on the visual contents of the images, rather than relying on human-supplied keywords or captions. In contrast to the text-based approach, CBIR operates on a totally different principle; i.e., to retrieve the stored images from a collection of images by comparing the features that were automatically extracted from the images themselves. CBIR involves matching a query image with the images stored in a database. The first step of this process involves extracting a feature vector to represent the unique characteristics of each image. The features used for retrieval can be either primitive or semantic, but the extraction process must be automatic. The retrieval process is highly dependent on the representational formalism used to characterize the feature set. A quantified similarity value between two images is obtained by comparing their feature vectors. The commonly used image features include color, shape, and texture. Queries are issued through query by image example (QBE), which can either be provided or constructed by the users, or randomly selected from the image database.

There have been several systems and techniques developed in both the academic and commercial domains such as the IBM's query by image content (QBIC) system [1], Virage's VIR engine [2], VisualSEEk [3], and PhotoBook [4]. Recent improvements made to CBIR include the personalization of 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. Often, the retrieval results are not very satisfactory especially when the level of satisfaction is closely related to user 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. Also, it may be difficult for the user to describe the pattern of, say, a Bengal tiger without necessarily including that of, say, a Monarch butterfly. Clearly, user subjectivity in image retrieval is a very complex issue. Thus, a CBIR system needs to have the capability to discover users concept patterns and adapt to them. The relevance feedback (RF) technique has been proposed and applied with the aim to discover the user's concept patterns by bridging the gap between semantic concepts and low-level image features, as in [5]–[7].

A learning and retrieval framework is proposed in this paper. Our proposed framework provides for the dynamic discovery of the concept patterns of a specific user so that the retrieval of images is based on the user's regional focus. One of the major challenges is to discover 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.

Manuscript received February 4, 2004; revised November 19, 2004. The work of S.-C. Chen was supported in part by the National Science Foundation under Grants EIA-0220562 and HRD-0317692. S. Rubin was supported by the ONR and SPAWAR Systems Center, San Diego, CA. This work was produced, in part, by a U.S. government employee as part of his official duties and is not subject to copyright. It is approved for public release with an unlimited distribution. The work of M.-L. Shyu was supported in part by NSF ITR (Medium) IIS-0325260. The work of C. Zhang was supported in part by SBE-0245090 and the UAB ADVANCE program of the Office for the Advancement of Women in Science and Engineering. The preliminary version of the current draft was published in the *Proceedings of the Third International Workshop on Multimedia Data Mining* (MDM/KDD'2002). This paper was recommended by Associate Editor D. Zhang.

S.-C. Chen is with the Distributed Multimedia Information System Laboratory, School of Computer Science, Florida International University, Miami, FL 33199 USA (e-mail: chens@cs.fiu.edu).

S. H. Rubin is with the Space and Naval Warfare Systems Center (SSC), San Diego, CA 92152-5001 USA (e-mail: stuart.rubin@navy.mil).

M.-L. Shyu is with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124 USA (e-mail: shyu@miami.edu).

C. Zhang is with the Department of Computer and Information Sciences, University of Alabama at Birmingham, Birmingham, AL 35294 USA (e-mail: zhang@cis.uab.edu).

Digital Object Identifier 10.1109/TSMCC.2006.855507

Our proposed framework seamlessly integrates several learning techniques. First, user relevance feedback is supported during the retrieval process, which means that users interact with the system by choosing the positive and negative examples from the retrieved images based on their own concepts. Then, the user's feedback is fed into the retrieval system and triggers the modification of the query criteria, which best matches the user's concepts [8]. Second, multiple instance learning (MIL) and neural network techniques are integrated into the query-refining process. Such integration provides the functionality to identify the most interesting region within the image.

MIL was first used to categorize molecules in the context of drug design. Each molecule (bag) is represented by a bag of possible conformations (instances). Under the MIL scenario, each image is viewed as a bag of image regions (instances) in image retrieval. In fact, user feedback guides the system learning process through the positive and negative examples using MIL and informs the system to shift its focus of attention to the region of interest. This neural version space technology is applied to map the low-level image features to the user's concepts. The parameters in the neural network are dynamically updated to best represent the user's concepts according to user relevance feedback obtained during the retrieval process. In this sense, it is similar to the reweighting techniques used in the RF approach.

The remainder of this paper is organized as follows. A literature review in relevance feedback and multiple instance learning is presented in Section II. Section III presents our proposed learning and retrieval framework with the details of the multiple instance learning and neural network techniques used in our framework. The proposed framework for content-based image retrieval with experimental results is presented in Section IV. This paper is concluded in Section V.

II. Literature Review

A. Relevance Feedback

A plethora of research has served to establish the base for CBIR. However, most of these efforts ignore two distinct characteristics of CBIR systems: 1) the gap between high-level concepts and low-level features and 2) the subjectivity of human perception of visual content. In order to overcome these shortcomings, the concept of RF associated with CBIR was proposed in [9]. Relevance feedback is an interactive process by 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. The 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 proven successful in many application areas. In addition, 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 [9], [10] used in information retrieval [11]. 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 previous RF research has been based on low-level image features such as color, texture, and shape and can be classified into two basic approaches: query point movement and reweighting techniques [11]. The essential idea of query point movement is quite straightforward. It represents an attempt to move the estimation of the "ideal query point" toward positive example points and away from negative example points specified by the user in accordance with his/her subjective judgments.

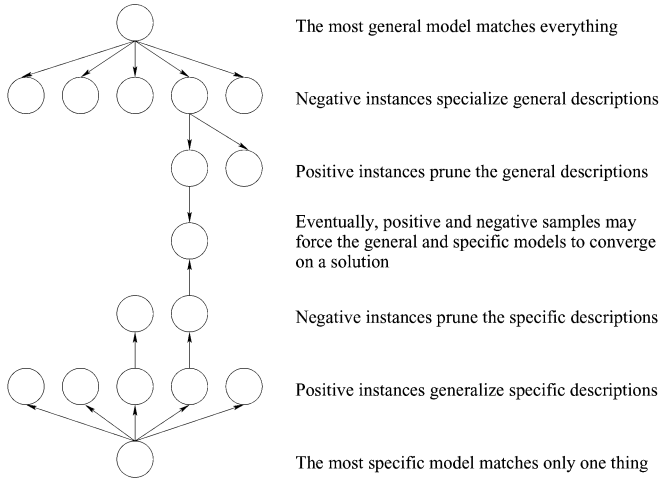


Fig. 1. Generalization and specialization in the version-space trees.

In particular, in the version space approach to concept learning [12], the set of all hypotheses consistent with the examples seen so far is maintained without remembering any of the examples. In a version space, a positive example may affect the most specific hypothesis by making it more general. Conversely, a negative example may affect the most general hypothesis by making it more specific. The true hypothesis is bounded by these two hypotheses. If the most specific and most general hypothesis are equal, then the concept has been uniquely identified from the training sequence. If these cross (that is, the most general becomes more specific than the most specific), then there is no hypothesis in the concept language consistent with the training sequence. As shown in Fig. 1, the G or generalization-tree is shown at the top, and the S or specialization-tree is shown at the bottom.

Rocchio's formula [13] is frequently used to iteratively update the estimation of the "ideal query point." The reweighting techniques, however, take the 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 [9], [14], [15]. The basic idea is to give larger weights to more important dimensions and smaller weights to less important ones.

One limitation of the low-level feature-based RF techniques in CBIR is that the captured mapping between low-level features and high-level semantics in one query using RF is discarded after the termination of that query session. It is then difficult for the CBIR system to learn the captured mapping information and further utilize the information in the following query sessions.

In addition, it is more natural and powerful to represent the user's concept with semantic terms such as "annotations" than by using low-level features. Thus, the new trend in CBIR is to incorporate semantic content into relevance feedback in addition to incorporating low-level features. The hidden annotation mechanism in the PicHunter system [5] provides one example. Lu *et al*. [6] proposed the idea of semantic propagation based on relevance feedback. In their approach, the progressive learning process is combined with user relevance feedback to propagate the keyword annotation from the labeled images to the unlabeled images so that a greater number of images are implicitly labeled by keywords.

B. Multiple Instance Learning

The multiple instance learning problem is a special kind of supervised machine learning problem, which has recently received more attention from the computational intelligence community. It has been applied to many applications such as drug activity prediction, stock prediction, natural scene image classification, and content-based image retrieval.

In the standard supervised machine learning methodology, each object in the set of training examples is labeled and the problem is to learn a hypothesis that can accurately predict the labels of the unseen objects. Conversely, in the multiple instance learning scenario, the labels of individual objects in the training data are not available; instead, the labeled unit is a set of objects called a *bag*. 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 generates labels for 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 where the labeling information is incomplete. There are two kinds of labels in the domain of multiple instance learning, 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 was originally used in the context of drug activity prediction. In this domain, the input object is a molecule, and the observed result pertains to whether the molecule binds to a target "binding site" or not. If a molecule binds to the target "binding site," then we label it as *positive*; otherwise, we label it *negative*. A molecule has many 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. The others typically have no effect on the binding. Unfortunately, the binding activity of a specific molecular conformation cannot be directly observed. Actually, only the binding activity of a molecule can be observed. Thus, 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 [16].

In addition to its application in drug discovery, multiple instance learning has also been applied to stock prediction to discover the relationship between a stock's behavior and its characterizing economic features. One main problem in stock prediction is the ambiguity in causality. multiple instance learning can discover the fundamental economic features that determine a stock's behavior by considering its monthly behavior as a bag. Thus, a reasonable prediction can be made despite the ambiguity.

The applications of multiple instance learning pertaining to this paper's theme include natural scene image classification and content-based image retrieval. In the first application, a natural image scene usually contains many different semantic regions, and its semantic category is usually only determined by one or more regions in the image. There may be some regions that do not fit the semantic meaning of the category. For example, assume 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 fact and, furthermore, only consider the features of the river object when learning the classifier, then better performance can be achieved than by using the features of the whole image instead. Maron *et al*. applied multiple instance learning into natural scene image classification [17] using this approach. 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*. The concept can be learned using multiple instance learning and the learned concept can be used for scene classification.

The multiple instance learning method can also be applied to CBIR for natural scene image classification. 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, while other objects are unrelated to it. The multiple instance learning method can discover the objects that are actually related to the user concept by considering each object as an instance and the image as a bag. 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. The multiple instance learning method has been applied in CBIR using this idea [18], [19].

In addition to the application of multiple instance learning, a significant amount of research has been carried out in multiple instance learning algorithms. Dietterich *et al*. [16] represent the target concept by an axis-parallel rectangle (APR) in the n-dimensional feature space and present several multiple instance learning algorithms for learning the axis-parallel rectangles. A MULTINST algorithm which is also an APR-based method for multiple instance learning was proposed in [20]. The concept of diversity density was introduced by Maron and Lozano-Perez [17] and a two-step gradient descent with multiple starting points was applied to find the maximum diversity density. The

EM-DD algorithm was proposed by Zhang and Goldman [21], based on diversity density. Their algorithm was predicated on the assumption that each bag has a representative instance that was treated as a missed value, and then, the expectation–maximization (EM) method and quasi-Newton method were used to simultaneously learn the representative instances and maximize the diversity density. Ray and Page [22] also used the EM method for multiple instance regression. Wang *et al*. [23] explored the lazy learning approaches in multiple instance learning. They developed two kNN-based algorithms: citation-kNN and Bayesian-kNN. In Zucker and Chevaleyre [24], the authors attempted to solve the multiple instance learning problem with decision trees and decision rules. Ramon *et al*. [25] proposed the multiple instance neural network.

III. Proposed Learning and Retrieval Framework

In our proposed framework, an open multiple instance learning approach is designed, where "open" means that different subalgorithms may be plugged into the learning framework for different applications. It provides the opportunity to select the most suitable subalgorithm to get the best performance for a specific application. In our proposed framework, the multilayer feedforward neural network and backpropagation algorithm are plugged into the multiple instance learning approach.

Our approach is amenable to massive and distributed parallel processing, which serves to partially offset the relative slowness of the backpropagation algorithm as well as the high spatial requirements implied by the power set methodology, which is common to Definition 2 and version spaces in general. This is especially the case when given relatively large feature sets. It follows that the selection of a good characteristic feature vector is fundamental to the tractability of the methodology. This note is in agreement with Lin and Vitter's [26] theoretical results on learning in neural networks having at least one hidden layer.

A. Open Multiple Instance Learning

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

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.

In multiple instance learning, unlike the case for traditional supervised 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 instances and is labeled *negative* if and only if all of its instances are negative. The multiple instance learning problem requires learning a function mapping from an instance to a label (either *positive* or *negative*) with the best possible approximation to the unknown real mapping function, which can be defined as follows.

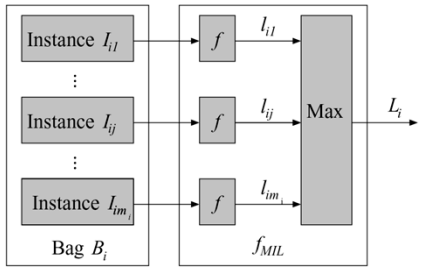


Fig. 2. Relationship between functions f and f_{MIL}.

Definition 2: Given an object space Φ, a label space $\Psi = \{1(\mathrm{Positive}), 0(\mathrm{Negative})\}$, and a set of n bags $B = \{B_i \mid B_i \in P(\Phi), i = 1 \ldots n\}$, where $P(\Phi)$ 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.

1) Problem Definition: Let $T = \langle B, L \rangle$ denote a training set where $B = \{B_i, i = 1, \ldots, n\}$ is the set of n bags in the training set, $L = \{L_i, i = 1, \ldots, n\}$ 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} $(j = 1, \ldots, m_i)$. 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 our proposed multiple instance learning framework, the label space is transformed from a discrete space $\Psi = \{1(\mathrm{Positive}), 0(\mathrm{Negative})\}$ to a continuous space $\Psi' = [0, 1]$ and the label of a bag actually indicates the extent to which that bag is *positive*-instead of either 100% *positive* or *negative*. The label "1" (*positive*) means the bag is one hundred percent *positive*, while the label "0" (*negative*) indicates that the bag is zero percent *positive*. The same applies to the label of an instance. The goal of learning, subsequent to this transformation, is to generate a mapping function $\hat{f} : \Phi \to \Psi'$ from the training examples to predict the extent to which an instance is *positive*. Specifically, in the multiple instance learning scenario, the extent to which a bag is *positive* is determined by the maximum extent to which its instances are positive. In other words, the label of a bag is the maximum label of its instances. The relationship between the functions f and f_{MIL} is given in Fig. 2.

As can be seen from Fig. 2, 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 $L_i = f_{\mathrm{MIL}}(B_i) = \max_j\{l_{ij}\} = \max_j\{f(I_{ij})\}$. The multiple instance learning problem 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 $L = \{L_i, i = 1 \ldots n\}$. The corresponding approximation of f_{MIL} is $\hat{f}_{\mathrm{MIL}}(B_i) = \max_j\{\hat{f}(I_{ij})\}$.

In our framework, the minimum square error (MSE) is adopted, i.e., we try to find the function $\hat{f}$ that minimizes the

2-norm

$$\mathrm{SE} = \sum_{i=1}^{n}(L_i - \hat{f}_{\mathrm{MIL}}(B_i))^2 = \sum_{i=1}^{n}(L_i - \max_j\{\hat{f}(I_{ij})\})^2. \quad (1)$$

Let $\gamma = \{\gamma_k, k = 1, \ldots, N\}$ denote the N parameters of the function f (where N is the number of parameters). The multiple instance learning problem is transformed into the following unconstrained optimization problem:

$$\hat{\gamma} = \arg\min_\gamma \sum_{i=1}^{n}(L_i - \max_j\{\hat{f}(I_{ij})\})^2. \quad (2)$$

Unconstrained optimization methods include gradient descent search, Newton's method, quasi-Newton methods, and the backpropagation (BP) learning method within a multilayer feedforward neural network. The target optimization function needs to be differentiated to apply these methods. In our multiple instance learning framework, we need to differentiate the function $E = (L_i - \max_j\{\hat{f}(I_{ij})\})^2$. The differentiation of the **max** function needs to be calculated first in order to accomplish this.

*2) Differentiation of the **max** Function:* As mentioned in [27], the differentiation of the **max** function results in a "pointer" that specifies the source of the maximum. Let

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

where $U(\cdot)$ is a unit step function, i.e.,

$$U(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

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

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

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 (2), we need to further calculate the differentiation of the function $E = (L_i - \max_j\{\hat{f}(I_{ij})\})^2$ on the parameters $\gamma = \{\gamma_k\}$ of $\hat{f}$. The first partial derivative is as follows:

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

Suppose the sth instance of bag B_i has the maximum value; i.e., $\hat{f}(I_{is}) = \max_j\{\hat{f}(I_{ij})\}$. According to (4), (5) can be written as

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

Furthermore, the nth derivative of the target optimization function E can be written as

$$\frac{\partial^n E}{\partial {\gamma_k}^n} = \frac{\partial^n (L_i - \max_j\{\hat{f}(I_{ij})\})^2}{\partial {\gamma_k}^n} = \frac{\partial^n (L_i - \hat{f}(I_{is}))^2}{\partial {\gamma_k}^n} \quad (7)$$

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

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

4) Multiple Instance Learning to Traditional Supervised Learning: The traditional supervised learning problem can be converted to an unconstrained optimization problem as shown in (9). This is similar to the analysis of the multiple instance learning problem given in Section IV-A

$$\bar{\gamma} = \arg\min_\gamma \sum_{i=1}^{n}(L_i - \hat{f}(O_i))^2. \quad (9)$$

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

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

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

Notice that (10) is the same as the right-hand side of (7), and (11) is the same as the right-hand side of (8) except that O_i in (10) and (11) represents an object, while I_{is} in (7) and (8) represents an instance with the maximum label in bag B_i. This similarity provides us with an easy way of transforming multiple instance learning into the traditional supervised learning.

The steps for transformation are as follows.

1) For each bag $B_i (i = 1, \ldots, n)$ in the training set, calculate the label of each instance I_{ij} belonging to it.
2) Select the instance having 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 $\{O_i\}$ $(i = 1, \ldots, n)$ 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 $(\{B_i\}, \{L_i\})$ is converted to the traditional supervised learning problem with the input $(\{O_i\}, \{L_{O_i}\})$.

Gradient-based search methods used in the traditional supervised learning, such as the steepest descent method, can be applied to multiple instance learning subsequent to this transformation.

There still exists a major difference between multiple instance learning and traditional supervised learning despite the transformation from multiple instance learning to the traditional supervised learning. The training set is static and usually does not change during the learning procedure in the traditional supervised learning. However, in the transformed version of multiple instance learning, the training set may change during the learning procedure. The reason for this 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. The training set constructed along with the aforementioned transformation may also change during the learning procedure. The fundamental learning method remains the same despite such dynamic. The following pseudocode defines our multiple instance learning framework.

MIL$(B,\ L)$
Input: $B = \{B_i,\ i = 1, \ldots, n\}$ *is the set of* n *bags in the training set and* $L = \{L_i,\ i = 1, \ldots, n\}$ *is the set of labels where* L_i *is the label of bag* B_i.
Output: $\gamma = \{\gamma_k,\ k = 1, \ldots, N\}$ *is the set of parameters of the mapping function* $\hat{f}$ *where* N *is the number of parameters.*

1) *Set initial values for parameters* γ_k *in* γ.
2) *While the termination criterion has not been met Do*
   /* *The termination criterion can be based on MSE or the number of iterations.* */
   a) *Transform multiple instance learning to traditional supervised learning using the method described in this section*
   b) *Apply the gradient-based search method in traditional supervised learning to update the parameters in* γ.
3) *Return the parameter set* γ *of function* $\hat{f}$.

Obviously, the convergence of our multiple instance learning framework depends on what kind of gradient-based search method is applied at Step 2(b). Actually, it converges at the same rate as does the gradient-based search method.

B. Image Processing Techniques

It is tacitly assumed that the user is only interested in a specific region of the query image for the application of multiple instance learning for learning user concept patterns. Thus, we first need to perform image segmentation.

1) WavSeg Image Segmentation: Instead of manually dividing each image into many overlapping regions [18], in this study, we propose to use a fast yet effective image segmentation method called WavSeg [28] to partition the images. In WavSeg, a wavelet analysis in concert with the SPCPE algorithm [29] 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 preprocessed 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, the 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, on average. It is worth pointing out that WavSeg is fast. The processing time for a 240×384 image is only about 0.33 s, average.

2) Image Feature Extraction: Both the local color and local texture features are extracted for each image region.

a) Color Features: HSV color space and its variants are proven to be particularly amenable to color image analysis. Thus, 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 13 color features for each region in our method.

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

The 13 color features and six texture features of each region are extracted after image segmentation. Thus, for each bag (image), the number of its instances (regions) is equal to the number of regions within that image. Each instance has 19 features.

C. Neural Network Techniques

A three-layer feedforward neural network is used in our experiments as the function f to map an image region (including those 19 low-level texture and color features) into the user's

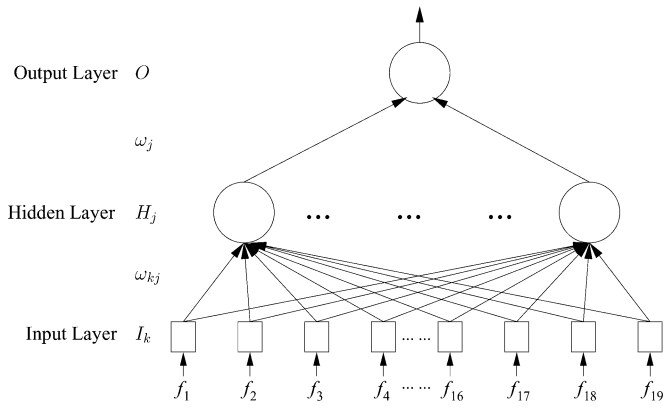


Fig. 3. Structure layout of the three-layer feedforward neural Network.

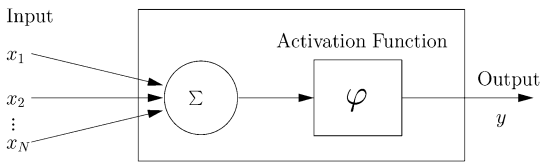


Fig. 4. Structure of the node.

high-level concept. Fig. 3 shows the structure of the three-layer feedforward neural network.

The input layer has 19 input units. Each of these units represents a low-level feature of an image region, and hence $\{f_1, f_2, \ldots, f_{19}\}$ are the 19 low-level image features mentioned in Section III-B. The hidden layer consists of 19 hidden nodes. ω_{kj} denotes the weight of the connection between the input unit I_k and the hidden node H_j. There is only one node in the output layer, and its output indicates the extent to which an image region satisfies the user's concept. The weight of the connection between the hidden node H_j, and the output node is denoted by ω_j.

The internal structure of a node is shown in Fig. 4. In this figure, $\{x_1, x_2, \ldots, x_N\}$ are the N inputs of the node and y represents the output of the node. The Sigmoid function with slope parameter 1 is used as the activation function. In other words, the input–output relationship can be written as

$$y = \frac{1}{1 + \exp(-\sum_{i=1}^{N} x_i)}. \tag{12}$$

We take the three-layer feedforward neural network as the mapping function $\hat{f}$ and the backpropagation (BP) learning algorithm as the gradient-based search method in our multiple instance learning framework. Then, the neural network parameters such as the weights of all connections (namely ω_{kj} and ω_j) are the parameters in γ [given in (9)] that we want to learn (search). The BP learning method was applied with a learning rate of 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 $|\mathrm{MSE}^{(k)} - \mathrm{MSE}^{(k-1)}| < \alpha \times \mathrm{MSE}^{(k-1)}$, where $\mathrm{MSE}^{(k)}$ denotes the MSE at the kth iteration and α is a small constant. In our experiments, α was set to 0.005.

D. Discussion

Our proposed learning and retrieval framework differs in the following two aspects when compared with traditional RF techniques. First, it is based on the assumption that the users are usually more interested in one specific region (blob object) than in other regions of the query image. However, to the best of our knowledge, recent efforts in RF techniques are based on global image properties of the query image. In order to produce a higher accuracy, we use the segmentation method proposed in [28] to segment an image into regions (segments) that roughly correspond to objects, which provide the retrieval system with the possibility of discovering the most interesting region for a specific user based on his/her feedback. Second, in many cases, what the user is really interested in is just an object of the query image (example). However, the user's feedback pertains to the whole image. The question of how to effectively identify the user's most interested object and how to precisely capture the user's high-level concepts based on his/her feedback on the whole image have not yet received much attention. In this paper, the multiple instance learning method is applied to discover the user's region of interest and subsequently mine the user's high-level concepts. Not only can the region of interest be discovered by so doing, but the ideal query point of that query image can also be approached within several iterations.

Moreover, compared with other multiple instance learning methods used in CBIR, our methodology has the following advantages. First, instead of manually dividing each picture into many overlapping regions [18], we adopt the image segmentation method in [28] to partition the images in a more natural way. Second, in other multiple instance learning-based image retrieval systems such as [19], the users are usually asked to provide the positive and negative examples by searching through a large number of images in the database. In our method, user feedback is efficiently and precisely applied to the image retrieval process. It is more efficient, since it is easy for the user to find some positive examples among the initial retrieved results. It is more precise, since the user can select the negative examples, from among the retrieved images, based on his/her subjective perception. Though the selected negative examples have similar features/contents in common with the query image, they have different foci of attention from the user's point of view. Our proposed system can better distinguish the real needs of the users from the "noisy" or unrelated information by selecting them as negative examples using multiple instance learning. As a result, it can discover that feature vector, related to a region in each image, which best represents the user's concept. Furthermore, the system can determine which dimensions of the feature vector are important by adaptively reweighting them using the neural network.

IV. Image Retrieval Using the Proposed Framework

In a CBIR system, the most common query method is "query-by-example," which means that the user submits a query example (image) and the CBIR system retrieves the images in the image database that are most similar to the query image. However, when a user submits a query image, in many cases, he/she

is only interested in a region of the image. The image retrieval system proposed by Blobworld [30] first segments each image into a couple of regions and then allows the user to specify the region of interest in the segmented query image. Unlike the Blobworld system, we apply the user's feedback and multiple instance learning to automatically capture the region of user's interest during the query-refining process. Another advantage provided by 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 process.

A. CBIR System Description

We have constructed a content-based image retrieval system, based on the proposed framework, using our own image repository, which includes 10 000 images from the Corel library. These images represent various categories for testing purposes.

In Zhang *et al.* [19], multiple instance learning is applied to CBIR. As a necessary step, prior to actual image retrieval, the user must first submit a set of images for 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 is obtained from the user's feedback on the initial retrieval results. In addition, the user's target concept is iteratively refined 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 region and $\Psi = \{1(\text{positive}), 0(\text{negative})\}$, where *positive* means that the feature vector representing this region satisfies the user's concept and *negative* means that it does not. An image is *positive* if there exists one or more regions in the image that can satisfy the user's concept. An image is $negative$ if none of the regions can satisfy the user's concept. Thus, an image can be viewed as a bag and its regions are the instances of the bag in the multiple instance learning scenario. The user's feedback can provide the labels (*positive* or *negative*) for the retrieved images during the image retrieval procedure. The labels are assigned to the individual images—not to individual regions. Thus, the image retrieval task can be viewed as a multiple instance learning task aiming to discover the mapping function f—using it to mine the user's high-level concept from the low-level features.

The user only needs to submit a query image. There are no training examples available at the outset of the retrieval process. This means that the learning method is not applicable at the current stage. Hence, a metric based on color histogram comparisons is applied to measure the similarity of a chosen pair of images. For each color, the two most significant bits of each R, G, B color component are extracted to compose a six-bit color code [31], [32]. The six-bit code provides 64 bins. Each image can be converted to a histogram with 64 bins and thus can be represented by a point in the 64-dimension feature space. Both the Manhattan distance and the Euclidean distance between two points are used for a measure of the dissimilarity between the two images represented by those two points.

Users can provide feedback, after the first round of retrieving those "most similar" images, by labeling each retrieved image as *positive* or *negative*. A set of training examples $\{B+, B-\}$ can be constructed, based on user feedback, where $B+$ consists of all the positive bags (i.e., the images to which the user assigns *positive* labels) and $B-$ consists of all of the negative bags (i.e., the images to which 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. Feedback and learning are iterative processes. Thus, the user's high-level concept is iteratively refined (i.e., by way of user feedback) until the user is satisfied. The query process may then be terminated by the user. In addition, similar to the idea as proposed by Su *et al.* [33], the negative images collected at each iteration are given a "punishment" value because the user is not interested in them. Thus, those images will be placed lower in the ranked list of images returned to the user in the next iteration.

Fig. 5 depicts the user interface for this system. As 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 similarity to the query image. The user may use the pull-down list under an image to input his/her feedback pertaining to that image (i.e., negative or positive). Subsequently, the user can affect the next query. The user's concept is then progressively acquired by the system by way of feedback. The refined query will return a new collection of matching images to the user.

B. Performance Analysis Using Query Examples

Here, some query examples will be given to illustrate how our CBIR system works, as well as for the purpose of comparing our system's performance with that of a color-histogram-based CBIR system in the absence of multiple instance learning.

1) Histogram-Based CBIR System: A histogram-based CBIR system was constructed for the purpose of performance comparison with our system. In a histogram-based CBIR system, each image is converted to a 64-dimensional feature vector with 64 bins, representing its color histogram. Each image can then be represented by a set of points in the 64-dimension space. Both the Manhattan metric and the Euclidean distance are used to measure the distance between two points (namely the dissimilarity between two corresponding images) as discussed in Section IV-A. The relevance feedback technique is also implemented in the system. Specifically, the reweighting method, known as the standard deviation method [9], is adopted. Assume that $(a_1, a_2, \ldots, a_{64})$ and $(b_1, b_2, \ldots, b_{64})$ are feature vectors for images A and B respectively and ω_i is the weight of the ith feature. All of the features have the same weight for the initial query, namely, $\omega_i = 1$.

The negative examples provided by user relevance feedback are not used, since the source negative images are usually quite irrelevant. However, negative examples are nonetheless useful in relevance feedback techniques categorized as, "query point movement." This is because those techniques try to move the

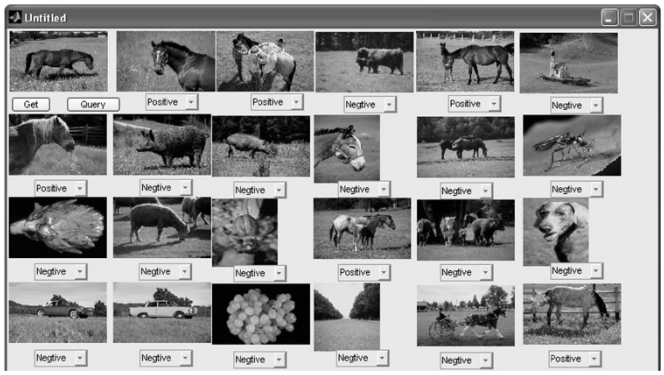


Fig. 5. Interface of the proposed CBIR system and initial query results.

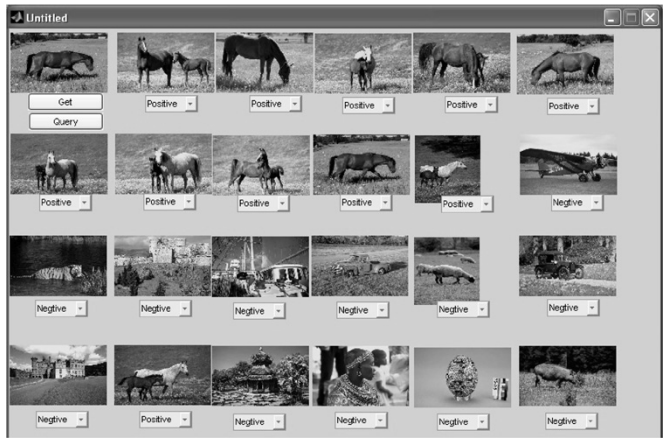


Fig. 6. Query results of our CBIR system after four iterations of user feedback.

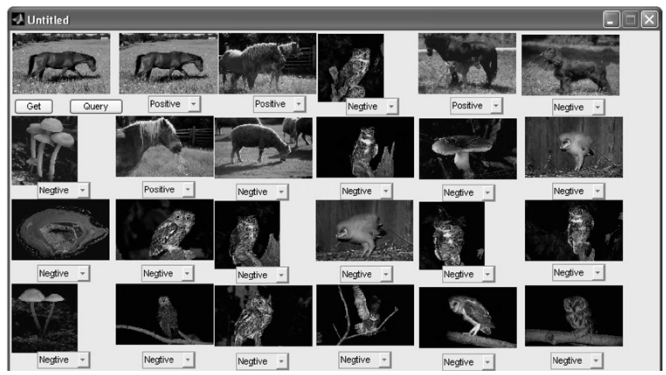


Fig. 7. Query results of the histogram-based CBIR system after four iterations of user feedback.

estimation of the "ideal query point" away from the negative examples. It would be better not to apply the negative examples in the reweighting relevance feedback techniques in view of the potentially high degree of irrelevancy among the negative examples. The standard deviation method is described as follows. The standard deviation dev_i is calculated for the ith feature of those positive examples. If dev_i of the positive examples is high, then it is safe to conclude that the values of the ith feature in those positive examples are not very relevant to the input query, and a corresponding low weight ω_i is assigned to it as a result. Thus, the inverse of dev_i is used as the basic indication to update the weight ω_i.

2) Query Examples: In this section, query examples are given to illustrate how our CBIR system works, as well as to compare the query results provided by our system with those provided by the histogram-based CBIR system described previously.

As shown in Fig. 5, the example query is located at the top-left corner. There is one horse on the grass in the query image. Assume that the horse object (not the grass) is what the user is really interested in. Fig. 5 also provides the initial retrieval results using a simple color histogram-based Manhattan metric of image similarity. Since the histogram-based CBIR system uses the same method on the initial query, its initial query results are the same as referred to in Fig. 5. As can be seen from this figure, many retrieved images have no horse 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 images with the horse object embedded. The proposed CBIR system can solve the problem by integrating the user's feedback with multiple instance learning. This follows because 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 the user's real interest and thus capture the user's high-level concept. Fig. 6 shows the query results after four iterations of user feedback. As shown in this figure, more images containing the horse object are successfully retrieved by the system. In particular, almost all of them have higher ranks than those of the other retrieved images. On the other hand, the irrelevant images having a similar color distribution on the whole image-such as the cow image and the dog image-are filtered out during the feedback and learning process. Thus, this example serves to illustrate that our proposed framework is effective in identifying the user's specific intention. In this manner, it can be used to mine the user's high-level concepts. Fig. 7 shows the query results of the histogram-based CBIR system after four iterations of user feedback. As can seen from this figure, these results are inferior to those obtained from our CBIR system. While the relevance feedback from this system can automatically adjust the weights of its features, it is not capable of capturing the user's interest in specific objects (e.g., the horse object).

Another query example is shown in Fig. 8. There is a green lawn with mountain views under a blue sky in the target image. Assume that the user is more interested in the green lawn than in the mountains and the blue sky. Fig. 8 shows the initial query results retrieved by the color histogram-based method. As can be seen from this figure, a couple of images without the green lawn were retrieved. The reason for this is that the histogram-based CBIR system has no concept pertaining to the user's subjectivity, because any similarity between an image and the target image is totally determined by the global color histogram of the images. In our CBIR system, the user can provide his/her judgments on the query results via relevance feedback.

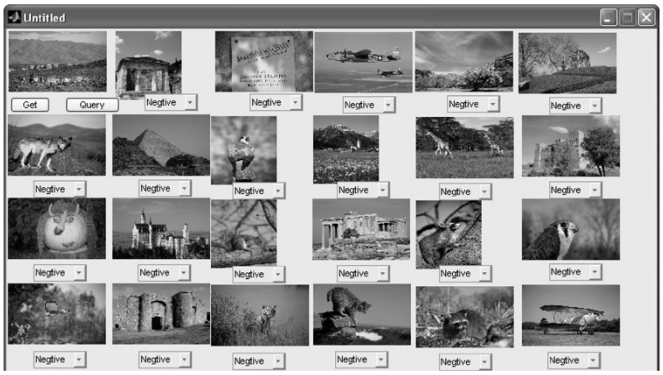


Fig. 8. Another query example: the initial query results.

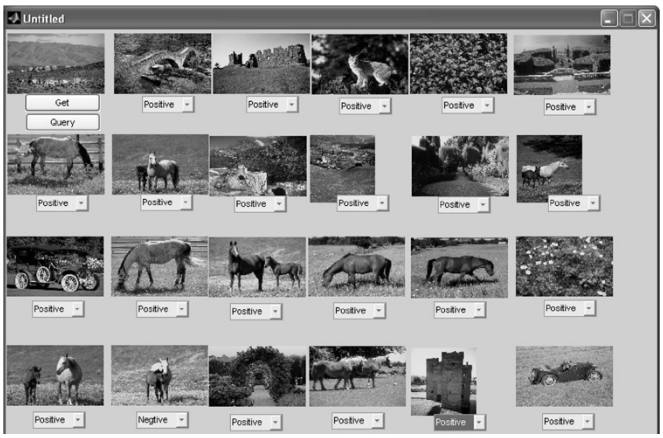


Fig. 9. Another query example: the query results of our CBIR system after four iterations of user feedback.

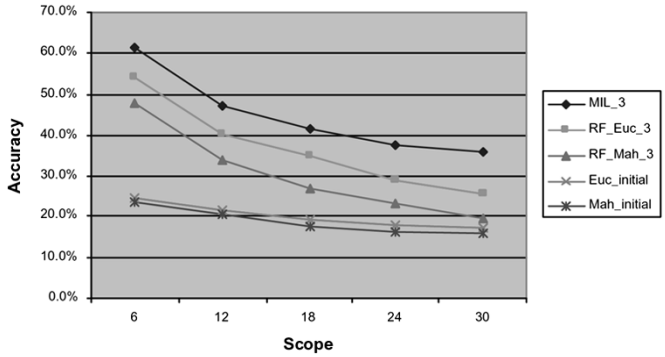


Fig. 10. Accuracy curves of our CBIR system and the histogram-based CBIR system after three RF iterations.

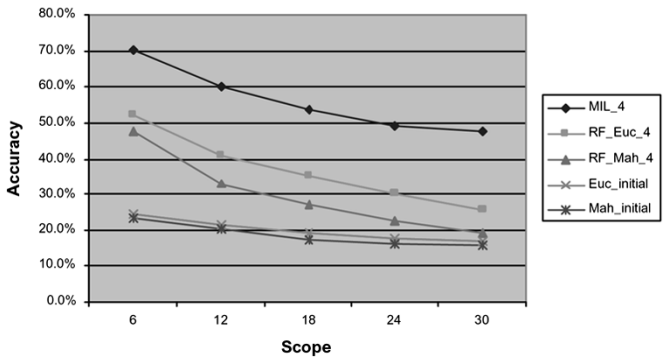


Fig. 11. Accuracy curves of our CBIR system and the histogram-based CBIR system after four RF iterations.

Furthermore, multiple instance learning is applied to discover the user's real interests. Fig. 9 provides the query result after four iterations of relevant feedback and learning. As can be seen in Fig. 9, those images sans green lawn were completely filtered out. All of the retrieved images contain green lawn or grass. It can be concluded, on the basis of this example, that our CBIR system successfully discovered the user's real interests. This, in turn, serves to improve query performance.

Furthermore, a number of experiments have been conducted on our CBIR system. It usually converges after three or four iterations of user feedback. Also, in many cases, the region in the query image associated with the greatest user interest can be discovered. Thus, query performance can indeed be improved.

C. Performance Evaluation

Accuracy serves as the standard by which to measure the retrieval performance of a CBIR system. The term accuracy is defined the same as precision within a certain scope. Recently, in the area of content-based image retrieval (CBIR), accuracy, instead of precision recall, has been widely used for performance evaluation and comparison. Such examples can be easily found in most of the recent works in CBIR [33]. The reasons for using accuracy are twofold: 1) As discussed in [34], image retrieval systems are designed to return only a few relevant images, where the user only browses the top few images; thus, precision is emphasized over recall; and 2) as the size of image database grows, manually separating the collection into relevant and nonrelevant sets becomes infeasible, which in turn prevents the accurate evaluation of recall.

In this experiment, we compare our CBIR system with the histogram-based CBIR RF system. We chose 60 query images in a way that their segmentation results are reasonably good, since the performance of the region-based system largely depends on the image segmentation results. These query images belong to different categories (such as the "flower," "vehicle," "human," "landscape," "animal," etc.) and are used to test the performance of our proposed framework. Figs. 10 and 11 show the curves for the average accuracy values of our CBIR system and the histogram-based CBIR RF system, respectively. In Figs. 10 and 11, MIL_3 and MIL_4 represent accuracy results of the proposed CBIR system after three and four iterations, respectively. RF_Euc_3(4) and RF_Mah_3(4) denote the accuracy results of the histogram-based CBIR RF system after three and four RF iterations, respectively. Here, "Euc" stands for the Euclidean distance, and "Mah" stands for the Manhattan distance. From these figures, we have the following observations: 1) Our CBIR system outperforms the histogram-based RF technique in all cases. 2) The Euclidean metric performs better than the Manhattan metric at each RF iteration—including the initial retrieval results.

The current implementation of our proposed framework is based on Matlab. The average execution time per feedback is evaluated based on our current system setting (CPU 2.8 GHz, 512 K RAM). As a result, the histogram-based RF technique takes 4–5 s per feedback, while the proposed CBIR system requires about 20 s for each feedback, since there is an extra cost for learning and the number of image regions is much larger than the number of images (4–5 regions per image). However, our approach is amenable to massive and distributed parallel processing, which serves to partially offset the relative slowness of the backpropagation algorithm as well as the high spatial requirements implied by the power set methodology. As part of our future work, a more comprehensive performance evaluation will be conducted.

V. Conclusion

In this paper, a learning and retrieval framework is presented, which is used to intelligently and efficiently retrieve images for a CBIR system. One of the main goals of our framework is to map the original visual feature space into a space that better describes the user-defined high-level concepts via user relevance feedback and multiple instance learning. Relevant feedback provides for the subjective capture of the user's high-level visual concepts, whereas multiple instance learning enables the automatic and precise learning of the user's high-level concepts. This is accomplished by capturing the user's specific region-of-interest in an image. To achieve this goal, a multiple instance feedback model that accounts for the various concepts/responses of the user is introduced. In our framework, it is assumed that 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 example images from among the displayed images. Multiple instance learning is applied to capture the objects that the user is really interested in subsequent to obtaining user relevance feedback. Low-level features and high-level concepts are simultaneously mapped. Each new query is chosen to more closely achieve user expectation, given previous user responses. Query-by-image-example experiments with accuracy evaluation were conducted to test the performance of our framework. The experimental results demonstrate that our proposed framework can progressively learn the underlying mapping between the local visual feature vector of the specified region-of-interest in the image and the user's high-level concept. Again, this is accomplished through the feedback and learning procedure for effectively retrieving the images.

References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, vol. 28, no. 9, pp. 23–31, Sep. 1995.

[2] The Virage website. [Online]. Available: http://www.virage.com

[3] J. R. Smith and S. F. Chang, "Visualseek: A fully automated content-based image query system," in *Proc. ACM Int. Conf. Multimedia*, Boston, MA, 1996, pp. 87–98.

[4] A. Pentland, R. W. Picard, and S. Sclaroff, "Solving the multiple-instance learning problem: A lazy learning approach," in *Proc. Storage and Retrieval for Image and Video Databases II, SPIE-Int. Soc. Opt. Eng.*, vol. 2185, 1994, pp. 34–47.

[5] I. J. Cox, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, pichunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, Jan. 2000.

[6] Y. Lu, C. H. Hu, X. Q. Zhu, H. J. Zhang, and Q. Yang, "A unified framework for semantics and feature based relevance feedback in image retrieval systems," in *Proc. 8th ACM Int. Conf. Multimedia*, Los Angeles, CA, 2000, pp. 31–37.

[7] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 5, pp. 644–655, Sep. 1998.

[8] Y. Rui and T. S. Huang, "A novel relevance feedback technique in image retrieval," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, 1999, pp. 67–70.

[9] Y. Rui, T. S. Huang, and S. Mehrotra, "Content-based image retrieval with relevance feedback in mars," in *Proc. Int. Conf. Image Process.*, Santa Barbara, CA, 1997, pp. 815–818.

[10] Y. Rui and T. S. Huang, "Optimizing learning in image retrieval," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, Hilton Head Island, SC, 2000, pp. 236–243.

[11] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: query databases through multiple examples," in *Proc. 24th Int. Conf. Very Large Databases*, New York, NY, 1998, pp. 218–227.

[12] T. M. Mitchell, "Version Spaces: An Approach to Concept Learning," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1978.

[13] J. J. Rocchio, "Relevance feedback in information retrieval," in *The Smart System Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1971, pp. 313–323.

[14] S. Aksoy and R. Haralick, "A weighted distance approach to relevance feedback," in *Proc. Int. Conf. Pattern Recognition*, Barcelona, Spain, 2000, pp. 812–815.

[15] C.-H. Chang and C.-C. Hsu, "Enabling concept-based relevance feedback for information retrieval on the WWW," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 4, pp. 595–609, Jul./Aug. 1999.

[16] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles," *Artif. Intell. J.*, vol. 89, pp. 31–71, 1997.

[17] O. Maron and T. Lozano-Perez, "A framework for multiple-instance learning," *Advances in Neural Information Processing System 10*, Cambridge, MA: MIT Press, 1998, ch. 3.

[18] C. Yang and T. Lozano-Perez, "Image database retrieval with multiple-instance learning techniques," in *Proc. 16th Int. Conf. Data Engineering*, San Diego, CA, 2000, pp. 233–243.

[19] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts, "Content-based image retrieval using multiple-instance learning," in *Proc. 9th Int. Conf. Machine Learning*, Univ. New South Wales, Sydney, Australia, 2002, pp. 682–689.

[20] P. Auer, "On learning from multi-instance examples: Empirical evaluation of a theoretical approach," in *Proc. 14th Int. Conf. Machine Learning*, San Francisco, CA, USA, 1997, pp. 21–29.

[21] Q. Zhang and S. A. Goldman, "EM-DD: A improved multiple-instance learning technique," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, 2001, pp. 1073–1080.

[22] S. Ray and D. Page, "Multiple-instance regression," in *Proc. 18th Int. Conf. Learning*, Williams College, Williamstown, MA, 2001, pp. 425–432.

[23] J. Wang and J.-D. Zucker, "Learning from user feedback in image retrieval systems," in *Proc. 17th Int. Conf. Machine Learning*, Stanford, CA, 2000, pp. 1119–1125.

[24] J.-D. Zucker and Y. Chevaleyre, "Solving multiple-instance and multiple-part learning problems with decision trees and decision rules. Application to the mutagenesis problem," in *Proc. 14th Biennial Conf. Can. Soc. Comput. Studies of Intelligence*, Ottawa, ON, Canada, 2001, pp. 204–214.

[25] J. Ramon and L. D. Raedt, "Multi-instance neural networks," in *Proc. ICML 2000 Workshop on Attribute-Value and Relational Learning*, Stanford, CA, 2000, pp. 53–60.

[26] J.-H. Lin and J. S. Vitter, "Complexity results on learning by neural nets," *Mach. Learn.*, vol. 6, no. 3, pp. 211–230, 1991.

[27] R. J. Marks, 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 Int. Conf. Neural Networks*, Beijing, China, 1992, pp. 471–477.

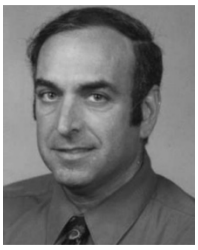
[28] C. Zhang, S.-C. Chen, M.-L. Shyu, and S. Peeta, "Adaptive background learning for vehicle detection and spatio-temporal tracking," in *Proc. 4th IEEE Pacific-Rim Conf. Multimedia*, Singapore, 2004, pp. 1–5.

[29] S.-C. Chen, S. Sista, M.-L. Shyu, and R. L. Kashyap, "An indexing and searching structure for multimedia database systems," in *Proc. IS&T/SPIE-Int. Sec. Opt. Eng. Conf. Storage and Retrieval for Media Databases 2000*, San Jose, CA, Jan.2000, pp. 262–270.
[30] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 24, no. 8, pp. 1026–1038, Aug. 2002.
[31] A. Nagasska and Y. Tanaka, "Automatic video indexing and full video search for object appearance," *IFIP Trans. Visual Database Systems II*, pp. 113–127, 1992.
[32] H. J. Zhang, A. KanKanhalli, and S. W. Smoliar, "Automatic video partitioning and indexing," in *Proc. IFAC 1993 World Congr.*, Sydney, Australia, 1993.
[33] Z. Su, H. Zhang, and S. L. S. Ma, "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 924–937, Aug. 2003.
[34] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: A similarity retrieval algorithm for image databases," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 3, pp. 301–316, Mar. 2004.

Shu-Ching Chen (M'95–SM'04) received the M.S. degree in computer science, electrical engineering, and civil engineering and the Ph.D. degree from Purdue University, West Lafayette, IN, in 1998.

He has been an Associate Professor in the School of Computer Science (SCS), Florida International University (FIU), Miami, since August 2004. Prior to that, he was an Assistant Professor in SCS at FIU dating from August 1999. His main research interests include distributed multimedia database systems, data mining, and multimedia networking. He has authored and coauthored more than 130 research papers in journals, refereed conference/symposium/workshop proceedings, and has writen book chapters.

Dr. Chen was awarded the University Outstanding Faculty Research Award from FIU in 2004. He also received the Outstanding Faculty Research Award from SCS at FIU in 2002. He is the general Co-Chair of the IEEE International Conference on Information Reuse and Integration and Program Chair of several conferences.



Stuart H. Rubin (M'88–SM'00) received the B.S. degree in business from the University of Rhode Island, Kingston, in 1975, the M.S. degree in industrial and systems engineering from Ohio University, Athens, in 1977, the M.S. degree in computer science from Rutgers University, Piscataway, NJ, in 1980, and the Ph.D. degree in computer and information science from Lehigh University, Bethlehem, PA, in 1988.

He is currently a Senior Scientist with the Space and Naval Warfare Systems Center (SSC), San Diego, CA. He has authored over 150 refereed conference and journal papers as well as several patent applications on behalf of SSC, San Diego. His professional interests center on heuristic methodologies for data mining, multisensor fusion, associative memory, and knowledge discovery. He is an Associate Editor for the *International Journal of Modeling and Simulation* and the *Journal of Systemics, Cybernetics, and Informatics*.

Dr. Rubin is a member of AFCEA, the American Association for the Advancement of Science, the New York Academy of Sciences, the North American Fuzzy Information Processing Society, and several other scientific societies. He chairs the IEEE Systems, Man, and Cybernetics Technical Committee on knowledge acquisition in intelligent systems. He is also an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART C. He is the Founder and General Co-Chair of the IEEE International Conference on Information Reuse and Integration (IRI) and has delivered several keynote lectures at international conferences. He also currently serves on the IEEE Board of Governors (BoG). In 2003, he was awarded the IEEE Systems, Man, and Cybernetics Society's Outstanding Service Award.

Mei-Ling Shyu (M'95–SM'03) received the M.S. degrees in computer science, electrical engineering, and restaurant, hotel, institutional, and tourism management and the Ph.D. degree from Purdue University, West Lafayette, IN, in 1999.

She has been an Associate Professor in the Department of Electrical and Computer Engineering (ECE), University of Miami (UM), Coral Gables, FL, since June 2005. Prior to that, she was an Assistant Professor in ECE at UM dating from January 2000. Her research interests include data mining, multimedia database systems, multimedia networking, and database systems. She has authored and coauthored more than 110 technical papers published in various prestigious journals, referred conference/symposium/workshop proceedings, and book chapters.

Chengcui Zhang (M'00) received the B.S. and M.S. degrees in computer science from Zhejiang University, Zhejiang, China, and the Ph.D. degree from Florida International University (FIU), Miami, in 2004.

She has been an Assistant Professor of computer and information science at the University of Alabama at Birmingham (UAB) since August 2004. Her research interests include multimedia databases, multimedia data mining, image and video database retrieval, and GIS data filtering.

Dr. Zhang is the recipient of several awards, including the UAB ADVANCE Junior Faculty Research Award from the National Science Foundation and the Presidential Fellowship and the Best Graduate Student Research Award at FIU.