Multimed Tools Appl (2007) 32:73–92
DOI 10.1007/s11042-006-0059-6

Capturing high-level image concepts via affinity relationships in image database retrieval

Mei-Ling Shyu · Shu-Ching Chen · Min Chen · Chengcui Zhang · Kanoksri Sarinnapakorn

Published online: 18 October 2006

Abstract In this paper, we present a mechanism called *Markov Model Mediator* (*MMM*) to facilitate the efficient and effective capturing of high-level image concepts in content-based image retrieval (CBIR). *MMM* serves as the retrieval engine of the CBIR system and uses affinity-based similarity measures. This mechanism is effective in capturing subjective user concepts in that it not only takes into consideration the global image features, but also learns the high-level concepts of the images from the history of user access patterns and access frequencies on the images in the image database, which differentiates it from the common methods in CBIR. The advantage of our proposed mechanism is that it exploits the richness in the structured description of visual contents as well as the relative affinity relationships among the images. Consequently, it provides the capability to bridge the gap between the low-level features and the high-level concepts. This mechanism is also efficient in

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

K. Sarinnapakorn
e-mail: ksarin@miami.edu

S.-C. Chen · M. Chen
Distributed Multimedia Information System Laboratory, School of Computer Science,
Florida International University, Miami, FL 33199, USA

S.-C. Chen
e-mail: chens@cs.fiu.edu

M. Chen
e-mail: mchen005@cs.fiu.edu

C. Zhang
Department of Computer and Information Science, University of Alabama at Birmingham,
Birmingham, AL 35294, USA
e-mail: zhang@cis.uab.edu

that it integrates Principal Component Analysis (PCA) to significantly reduce the image search space at a low cost before performing exact similarity matching. An off-line training subsystem for this framework was implemented and integrated into our system. The experimental results demonstrate that *MMM* can effectively capture users' high-level concepts more quickly.

Keywords Content-Based Image Retrieval (CBIR) · Markov Model Mediator (MMM) · Principal Component Analysis (PCA)

1 Introduction

The explosive growth of image databases has made efficient image indexing and retrieval mechanism indispensable. However, the traditional query-by-keyword is not suitable for image retrieval due to the intensive labor for annotating images and the difficulties in choosing the proper keywords for the images. Content-based image retrieval (CBIR) was proposed to solve this problem by retrieving the images based on their visual contents. Extensive research work [2, 7, 12, 28] has served to establish the base for CBIR. Most of the recent efforts have been aimed at meeting a crucial requirement of the CBIR systems: the semantic gap between the high-level concepts and the low-level features.

In the past few years, the Relevance Feedback (RF) approach to image retrieval has been an active research field. This powerful technique has proven successful in many application areas, and various ad hoc parameter estimation techniques have been proposed for the RF approaches. Relevance feedback is an interactive process in which the user judges the quality of the retrieval results returned by the system by marking those images that the user perceives as truly relevant. This information is then used to refine the original query. In addition to the low-level feature-based RF techniques in CBIR [1, 22], some CBIR systems incorporated the semantic content into relevance feedback in addition to the low-level features such as the hidden annotation mechanism in the PicHunter system [7] and semantic propagation in [17]. However, this process should be dealt with in a real-time manner in the loop because the metric dynamically depends upon the user's feedback and the context. The real-time learning of the distance metric or feature space transformations based on the user interactions remains an open issue. Recently, it has been realized that in many cases, image retrieval is ultimately a problem of object recognition and retrieval [2]. Thus, using the global features over the whole image cannot support the region-based queries. Region-based CBIR systems aim to overcome this problem by performing the retrieval at the object/region level. In these approaches, an image is first segmented into a number of regions, where each region roughly corresponds to an object and is represented by a set of local image features. The similarity measurements are then applied at the object/region level. There are a number of systems in this category [5, 9]. Recently, the research in integrating these two major techniques has gained much attention. The representatives are the RF-based region retrieval mechanisms proposed in [11, 29] which integrate RF and single region-based retrieval seamlessly.

In this paper, Markov Model Mediator (MMM), a probability-based mechanism that adopts both the *Markov Model* and the *Mediator* concept [27], is applied to

the dynamic content-based image retrieval process, where the user's perceptions are captured through the training process. The MMM mechanism has been applied in many applications including document management on the World Wide Web [24] and multimedia database management [23]. In addition, in our previous work [25], the empirical studies have been conducted on database clustering and the results demonstrated that the MMM mechanism leads to a better set of clusters in comparison with other well-known clustering methods, such as Single-Link, Complete-Link, Group Average Link, and PAM (Partitioning Around the Medoids) , etc. Some research work has been done to integrate the Markov model into CBIR. For example, Lin et al. [16] used a Markov model to combine the spatial and color information. In their approach, each image in the database is represented by a pseudo two-dimensional hidden Markov model (HMM) [21] in order to adequately capture both the spatial and chromatic information about that image.

In our proposed framework, the training and retrieval processes have the following advantages over other common work using RF:

- In contrast to the common RF approaches, where a heavy burden is placed on one single user to train the system in real-time via many iterations, the MMM mechanism supports accumulative learning. In other words, the MMM mechanism provides the capability to mine the subjective user concepts off-line from the training data set, which is constructed based on the accumulated query history.
- In RF, the user needs to take heavy responsibilities for the correctness of his/her feedbacks. Any mislabeled images will significantly damage the system's performance. Whereas in this work, since we consider the accumulative feedbacks from multiple users in the training process, the small amount of errors from individual users can be compensated in most cases.

Furthermore, the proposed MMM mechanism also provides a solution for reducing the search space by using principal component analysis (PCA). In this study, the original image feature space is transformed and projected into the PCA space where the covariance of any pair of principal components is 0, which means less redundant, less noisy, and more compact representations for the original feature space. Then the pre-filtering process is conducted in the PCA space to generate the desired candidate image pool according to the given query image. In contrast to the approach proposed in [26], the expensive distance computation can be avoided by only applying the distance computation to those images in the candidate image pool. In addition, we developed a training subsystem to collect the user access patterns and access frequencies [4]. Our experimental results show that the proposed mechanism is efficient in terms of the retrieval time and storage without sacrificing much accuracy.

The remainder of this paper is organized as follows. In Section 2, the architecture of the proposed framework is introduced, followed by the detailed discussions for each important component, such as feature extraction process, training process, and retrieval process. Section 3 presents the system implementation and our experimental results in applying the proposed framework to content-based image retrieval. The experimental results demonstrate that our framework can assist in retrieving more accurate results for the user queries. A brief conclusion is given in Section 4.

2 Architecture of the proposed framework

The architecture of the proposed framework is shown in Fig. 1. As can be seen from this figure, our proposed framework is divided into three major components based on their functionalities, namely image feature extraction process, training process, and retrieval process. In our framework, not only the low-level features (e.g., color) and the mid-level features (e.g., object locations), but also the high-level concepts learned from the off-line training process are used in the image retrieval process. Moreover, instead of conducting the exact similarity matching process in the whole database scope, a pre-filtering process using PCA is applied to reduce the search space. Hence, the retrieval process consists of the pre-filtering process and similarity matching process. In the following three subsections, each component and the relationships among the components are presented in detail.

2.1 Image feature extraction component

The features of color information and object location information for the images in the image database are both considered in our proposed framework. The HSV color space is used to obtain the color feature for each image due to the following two reasons: (1) it is a perceptual color space particularly amenable to color image analysis [6], and (2) it has been shown in the benchmark results that the color histogram in the HSV color space has the best performance [18]. In order to obtain the object location information, an unsupervised image segmentation method called SPCPE (Simultaneous Partition and Class Parameter Estimation) algorithm [3] is used, where an iterative approach is employed to estimate both the partition and the class parameters. Then each object is covered by a rectangle since the minimal bounding rectangle (MBR) concept in R-tree [10] is adopted. In addition, the centroid point of each object is used for space reasoning so that each object is mapped to a point object.

In this study, our main focus is to evaluate the performance of the MMM mechanism and to reduce the feature space instead of exploring the most appropriate features for image retrieval. Each image has a feature vector of 21 elements, where 12 are for color descriptions and 9 are for location descriptions. Based on the combinations of different ranges of the hue (H), saturation (S), and the intensity values (V), the color features 'black' (bl), 'white' (w), 'red' (r), 'red-yellow' (ry), 'yellow' (y), 'yellow-green' (yg), 'green' (g), 'green-blue' (gb), 'blue' (b), 'blue-purple' (bp), 'purple' (p) and 'purple-red' (pr) are considered. For any color whose number of pixels is less than 5% of the total number of pixels, its corresponding position in the feature vector has the value 0 since we treat it as non-important.

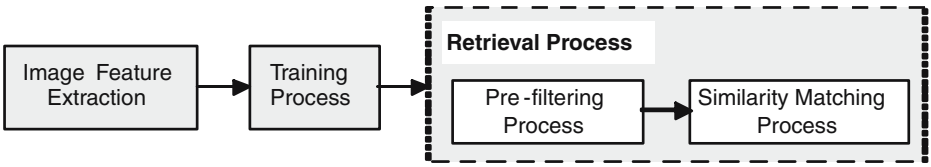


Fig. 1 Architecture of the proposed framework

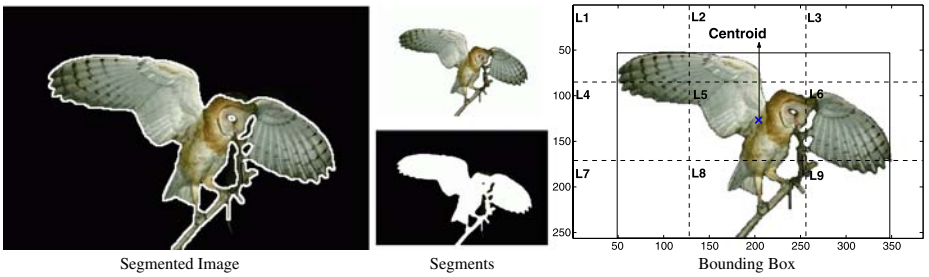


Fig. 2 Object location and the corresponding region in img1

Otherwise, the corresponding percentage of that color component is put in that position. For the location descriptions, each image is divided into 3×3 equal-sized reference regions that are ordered from left to right and top to bottom as L1, L2, L3, L4, L5, L6, L7, L8, and L9 (as shown in Fig. 2). The value 1 is assigned to a region when there is an object in the image whose centroid falls into that reference region, and 0 otherwise. Moreover, an object will be ignored if its area is less than 2% of the total image area. When necessary, each image can be divided into a coarser or finer set of regions.

In order to capture the appearance of a feature in an image, a feature matrix B is defined in the way that its rows are all the distinct images and columns are all the distinct features, where the value in the (p, q) entry is greater than zero if feature q appears in image p, and zero otherwise. We consider that the color and object location information are of equal importance. That is, the color features are normalized and their sum equals 0.5, and the location features are normalized and their sum is 0.5. Hence, each value indicates the likelihood of a feature observed from an image and the sum of the likelihoods of all the features of the image is 1. It is worth mentioning that our framework is flexible in terms of the update on the feature matrix B. Any normalized vector-based image feature sets can be used to construct the B matrix. Table 1 shows the example feature matrix B, where the first row represents the feature vector of the sample image in Fig. 2.

2.2 Training process component

Markov Model Mediator (MMM) is a probability-based mechanism which captures the probabilistic relationships among the states from the training data. In our framework, each image is called a state; while the probabilistic relationships among the states (images) are contained in the relative affinity matrix A, which is obtained from the off-line training process and serves as the indications of users' high-level

Table 1 B matrix—normalized image feature matrix

	bl	w	r	ry	y	yg	g	gb	b	bp	p	pr	L1	L2	L3	L4	L5	L6	L7	L8	L9
Img1	0.40	0.07	0	0.03	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0	0	0	0
Img2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Img3	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

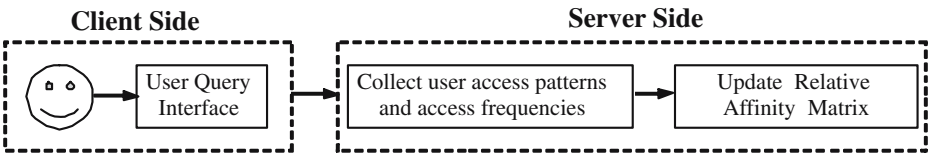


Fig. 3 Flowchart of the training process

concepts for image retrieval. In particular, A is defined to represent the affinity relationships among the images in the database. In order to construct the MMM model, a training subsystem is implemented by using a multi-threaded client/server architecture to collect the user access patterns and access frequencies. Figure 3 shows the flowchart of the training process.

As can be seen from this figure, a user query interface is provided in the client side; while the server side collects the user access patterns and access frequencies from the client and updates the relative affinity matrix A. The detailed implementation and operation of the training subsystem are described in Section 3.2.

2.2.1 Training data set

In our framework, the training subsystem is implemented to record the history of user access patterns and access frequencies on the image database during the training period. User access patterns denote the co-occurrence relationship among images accessed by user queries, while access frequencies denote how often a certain query was issued by the users. This training data set is used to bring in subjective user concepts and to construct the relative affinity matrix A off-line. Definition 1 gives the information available in the training data set.

Definition 1 Assume N is the total number of images in the database. The training data set consists of the following information:

– A set of queries $Q = \{q_1, q_2, ..., q_c\}$ that are issued to the database in the training period. Let $use_{k,m}$ denote the access pattern of image m $(1 \leq m \leq N)$ with respect to query q_k per time period, where $use_{k,m}$ is 1 if image m is accessed by q_k, and 0 otherwise. The value of $access_k$ denotes the access frequency of query q_k per time period.

The pair of user access pattern $(use_{k,m})$ and user access frequency $(access_k)$ provides the capability to capture the user concepts during the training process, as illustrated below.

Table 2 The user access patterns $(use_{k,m})$ for Img4–Img6

	Img4	Img5	Img6	...
q_1	1	1	0	...
q_2	0	0	1	...
q_3	1	0	1	...
...	...	...	...	...

Fig. 4 Three sample images (Img4–Img6)

Table 2 gives three example queries issued to our image database, where q_1 is a user-issued query related to retrieving some images with parachute jumping scenes, and q_2 and q_3 are two other queries with their interests in natural scenes with sky, mountain and grass. It is worth mentioning that in our training process, the queries issued to each image category (landscape, flower, etc.) are almost the same because the query images are randomly selected by the training subsystem from different image categories. By recording different users' feedbacks during the training process, the corresponding access frequencies $access_k$ for q_1, q_2, and q_3 are 8, 7 and 1, respectively. In Table 2, the entry $(k, m) = 1$ indicates that the m^{th} image is accessed by query q_k. For example, Img4 and Img5 (as shown in Fig. 4) are accessed together in q_1 with their corresponding entries in the user access pattern matrix having value 1, and the access frequency $access_1$ for q_1 is 8. As for queries q_2 and q_3 that focus on retrieving natural scene images, $access_2$ equals to 7 while $access_3$ equals to 1. Notice that Img4 and Img6 are accessed together in q_3 but not in q_2, which is due to the different human perceptions (parachute jumping scene or landscape scene) on Img4. However, since most of the users regard Img4 as a parachute jumping scene rather than a landscape scene, the value of $access_1$ is significantly larger than that of $access_3$. Consequently, after the system training, Img5 is more likely to be retrieved than Img6, given Img4 as the query image. Thus, the subjective user concepts about the images are captured by the pair of user access pattern and user access frequency.

2.2.2 Relative affinity matrix A

Based on the information contained in the training data set, we can capture the affinity relationships among the images in the database. That is, the more frequently two images are accessed together, the more closely they are related to each other. The relative affinity matrix A is constructed in two steps as follows:

– Firstly, a matrix AF is defined to capture the affinity measurements among all the images using the user access patterns and access frequencies as follows:

Definition 2 Each entry $aff_{m,n}$ in matrix AF indicates how frequently two images m and n are accessed together, and consequently how closely these two image are related to each other, where

$$aff_{m,n} = \sum_{k=1}^{c} use_{m,k} \times use_{n,k} \times access_k \tag{1}$$

– Then the matrix A can be obtained via normalizing AF per row and represents the relative affinity relationships among all the images in image database d. Let $a_{m,n}$ be the element in the (m, n) entry in A, where
if $\sum_{n\in d} aff_{m,n} \neq 0$

$$a_{m,n} = \frac{aff_{m,n}}{\sum_{n\in d} aff_{m,n}} \tag{2}$$

else

$$a_{m,n} = \begin{cases} 1, & \text{if } m = n \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The matrix A in MMM looks similar to the term-by-document (TBD) matrix as in Latent Semantic Indexing (LSI) [8]. However, they differ in the following aspects: (1) unlike the term-by-document matrix which captures the relationships between documents and text terms, matrix A is defined to represent the peer-to-peer affinity relationships among the states/images in the database. (2) another difference is that the "global term weighting" used in LSI is not necessary or applicable in MMM since queries are considered independent of each other. (3) in addition, a Singular Value Decomposition of a large term-by-document matrix (as in LSI) is not necessary in MMM mechanism. For the sake of efficiency, instead of updating matrix A on-line, the training subsystem only records all the user access patterns and access frequencies during a training period. Once the number of newly issued-queries reaches a threshold, the update of A matrix is triggered automatically. All the computations are done off-line.

2.3 Retrieval process component

After the training process, the high-level image semantics, which could not be effectively represented by the low-level features, are captured by the relative affinity matrix A, and the system is ready for image retrieval. However, as we mentioned earlier, efficiency is an important issue in CBIR. The proposed content-based image retrieval framework solves this problem by conducting the retrieval process in two steps:

1. A pre-filtering process is performed on the principal component subspace to reduce the search space.
2. The actual retrieval process is performed that computes the similarity functions on the original feature space only for those candidate images.

2.3.1 Pre-filtering process

The principal component analysis (PCA) [15] is integrated into our framework to generate a smaller feature space obtained by applying PCA on the original feature space (matrix B). The resulting principal components will form a more compact feature space for the purpose of pre-filtering. It is worth mentioning that PCA is a well-known technique for dimensionality reduction. In this study, however, we use the principal components to help reduce the image searching space instead of the feature space, as will be detailed in Section 2.3.1.3. Reducing the search space

in this manner reduces the number of images that need to perform exact similarity matching, which in turn reduces the computation time since exact similarity matching is more expensive in terms of computation.

2.3.1.1 Principal Component Analysis (PCA) The main capability of PCA is to reduce the dimensionality of the original feature space without losing the essential information [13]. Principal components are particular linear combinations of p features (X_1, X_2, ..., X_p), which hold three important properties. First, principal components are uncorrelated. Second, the first principal component has the highest variance, while the second principal component has the second highest variance, and so on. Third, the total variation in all principal components combined is equal to the total variation in the original features X_1, X_2, ..., X_p. Principal components can be easily obtained from the eigenanalysis of the covariance matrix or the correlation matrix of the original features. Principal components from these two matrices usually are not the same, and they are not simple functions of the others. It is better to perform PCA on the correlation matrix if the variables are measured on scales with wildly different ranges or if the units of measurement are not commensurate [14].

Although it requires p principal components to reproduce the total system's variability, often a small number k of the principal components are sufficient to explain the variability. In such cases, the initial p features can be replaced by the first k principal components, and the original data set (consisting of n measurements on p features) is reduced to a data set consisting of n measurements on k principal components. There will be almost as much information in the k components as there is in the original p features, if k is chosen appropriately.

2.3.1.2 Obtain principal components As mentioned earlier, the feature matrix B contains 12 color features and 9 spatial location features denoted by X_1, X_2, ..., X_{21} that are normalized to lie in the range of 0.0 to 0.5. Principal components are then obtained from the covariance matrix computed from B. In this case, among the obtained principal components, the first two components account for 50.12% of the total variation in the data and thus are used for image pre-filtering, which can reduce the search space and lower the processing time. Though we sacrifice some of the information in the original feature space in an exchange for search efficiency, these two principal components seem adequate since they provide reasonably good retrieval results (as illustrated in the experimental results in Section 3).

2.3.1.3 Pre-filtering to reduce the search space Since the image retrieval process is usually computationally expensive in particular with a large number of features, we propose the pre-filtering step to reduce the image searching space. The basic idea is that a principal component subspace is constructed based on the above-mentioned two principal component scores (for short, score 1 and score 2, respectively), where only those images close to the query image in this subspace will be included in a candidate image pool for further distance computations in the original feature space. In brief, given a desired candidate pool size c (e.g.,100), the top c images that have score 1 closest to that of the query image will be identified. Similarly, the top c images for score 2 are also identified. The intersection of the two sets of images is the desired candidate image pool. If the size of the intersection set is less than c, then the scan scope in component scores 1 and 2 is increased until the candidate pool is filled. This simpler approach is reasonable because principal components are

uncorrelated, and therefore the nearest neighbors found from individual principal component distributions will not be much different from the joint distribution.

2.3.2 Similarity matching process

After the pre-filtering process, the size of the candidate image pool for a certain query can be reduced dramatically. Then the exact retrieval process can be carried out to extract the most matched images based on both the image features and the relative affinities among images learned from the training process. Let C be the total number of images in the candidate image pool, p be the total number of distinct features of the images in the database, and the non-zero features of the query image q be denoted as $\{o_1, o_2, ..., o_T\}$, where T is the total number of non-zero features in the query ($1 \leq T \leq p$).

Definition 3 $W_t(i)$ is defined as the edge weight from the image i to the query image q at the evaluation of the t^{th} feature (o_t) in the query, where $1 \leq i \leq C$ and $1 \leq t \leq T$.

Based on the definition, the retrieval algorithm is given as follows.
At $t = 1$,

$$W_1(i) = a_{q,i}(1 - |b_i(o_1) - b_q(o_1)|/b_q(o_1)) \tag{4}$$

The values of $W_{t+1}(i)$, where $1 \leq t \leq T - 1$, are calculated by using the values of $W_t(i)$.

$$W_{t+1}(i) = W_t(i)(1 - |b_i(o_{t+1}) - b_q(o_{t+1})|/b_q(o_{t+1})) \tag{5}$$

Then the similarity function is defined as:

$$S(i) = \sum_{t=1}^{T} W_t(i) \tag{6}$$

Here, $a_{q,i}$ ($a_{q,i} \in A$) is the relative affinity relationship to indicate how closely the query image q is related to image i, and $b_i(o_k)$ ($b_i(o_k) \in B$) is the value of the k^{th} feature extracted from image i. The $S(i)$ value for image i in the candidate image pool represents the similarity score between images q and i, where a larger score suggests a higher similarity between images q and i. Note that for those images $\{i_1, i_2, ..., i_K\}$ whose affinity relationships with query image q have not been explored during the training process, i.e., $a_{q,i_k} = 0$ ($1 \leq k \leq K$), the corresponding $S(i_k)$ are equal to zeros. In this case, the Euclidean distance function can be used to calculate the similarity scores based on their low-level image features to break the tie.

Figure 5 gives the flowchart of our proposed image retrieval process. Unlike other methods that either have difficulties in capturing the high-level concepts or trying to

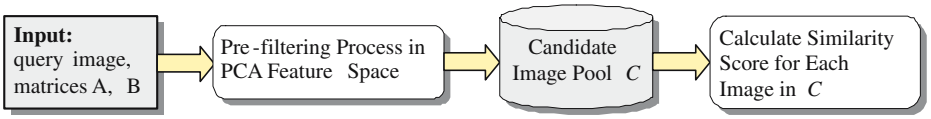


Fig. 5 Flowchart of our proposed image retrieval process

learn the concepts in real-time, our proposed framework captures the user concepts off-line from the training data set and achieves high efficiency in terms of storage and retrieval due to the following reasons: (1) Before the exact matching process, the pre-filtering process generates a small set of candidate images at low cost, which normally accounts for 4–8% of the total number of images in the database, and (2) Given a query image q issued by a user, in the retrieval process, only the data in the row q of matrix A are used. In addition, normally the features contained in one query image are no more than six, which enables us to load less than half of the whole feature matrix. Thus, we can retrieve the results more accurately and efficiently.

3 Experiments

3.1 Image database

Our image database consists of 10,000 color images of 72 semantic categories with various dimensions. In our experiments, the color information and object location information of the images are considered and the query-by-example strategy is used. For the purpose of supporting high-level meaning in the queries, a training subsystem is implemented to collect the user access patterns and query access frequencies information for the training data set.

The training subsystem for this framework is implemented and integrated into our system [4], a prototype multimedia management system developed by our research group aimed at supporting a comprehensive set of functionalities and components for multimedia database management systems. Figure 6 shows the interface of the training subsystem.

3.2 Implementation of training subsystem

The detailed training process is described as follows. First, the user selects one query image. After clicking the *Query* button, a query message is sent to the server through UDP (User Datagram Protocol). The query results will be sent back after the server fulfills the query process. It is worth mentioning that for training purposes, any available image retrieval methods can be implemented on the server side. Upon receiving the results, the user selects the images that he/she thinks are related to the query image by right-clicking on the image canvases, and clicks the *Feedback* button to send the feedback back to the server. When the server receives and identifies this feedback message, it updates the user access patterns and access frequencies accordingly. Then the user can continue the training process or exit.

In order to avoid bias and to capture the general users' perceptions, we asked ten university graduate students, who were not involved in the design or the development of our framework and had no knowledge of the image content in the database, to conduct the training process. Currently, there are 1,400 queries issued to the database which cover nearly 50% images in the database. As shown in our experimental results, this training data set can improve the query results dramatically. The more images covered in the training process, the higher retrieval accuracy that can be achieved.

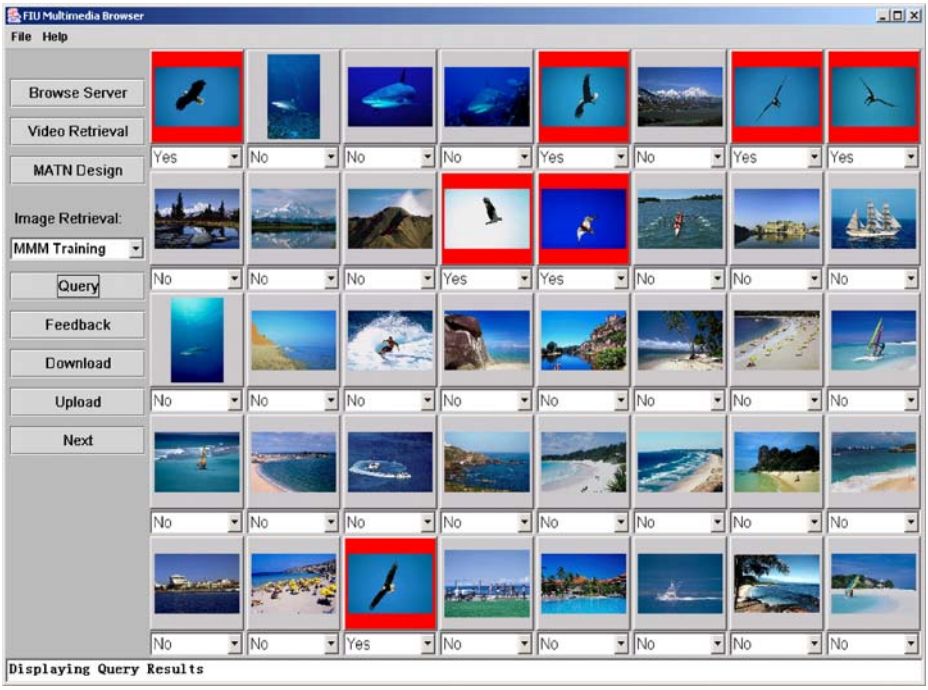


Fig. 6 The interface of the training subsystem

3.3 Experimental results and analysis

We use 80 randomly chosen query images that belong to the "landscape," "flower," "animal," "vehicle," and "human" categories (16 images per category) to test the performance and efficiency of our proposed framework. For each query image, our proposed retrieval process is executed to find a set of similar images. The degrees of matching between the query image and other images in the database are determined by the similarity values according to the S function (as shown in Eq. 6). In the following subsections, a query-by-image example is first given to demonstrate the effectiveness of our proposed approach, and then the complete performance analysis and comparison are conducted to show the performance improvement brought by the training process as well as the scalability of this framework.

3.3.1 Query-by-image example

As shown in Fig. 7, the retrieved images are ranked and displayed in descending order of their similarity scores from the top left to the bottom right, where the upper leftmost image is the query image marked with a red box. In this example, the query image belongs to the 'vehicle' category and contains complicated scenes. Though it contains an object (boat) with clear semantics, it is difficult to extract the foreground objects from the uneven background scenes for most of the existing region-based image retrieval systems due to the inaccuracy of the object segmentation. Figure 7 shows a snapshot of the most qualified twenty images retrieved for this query

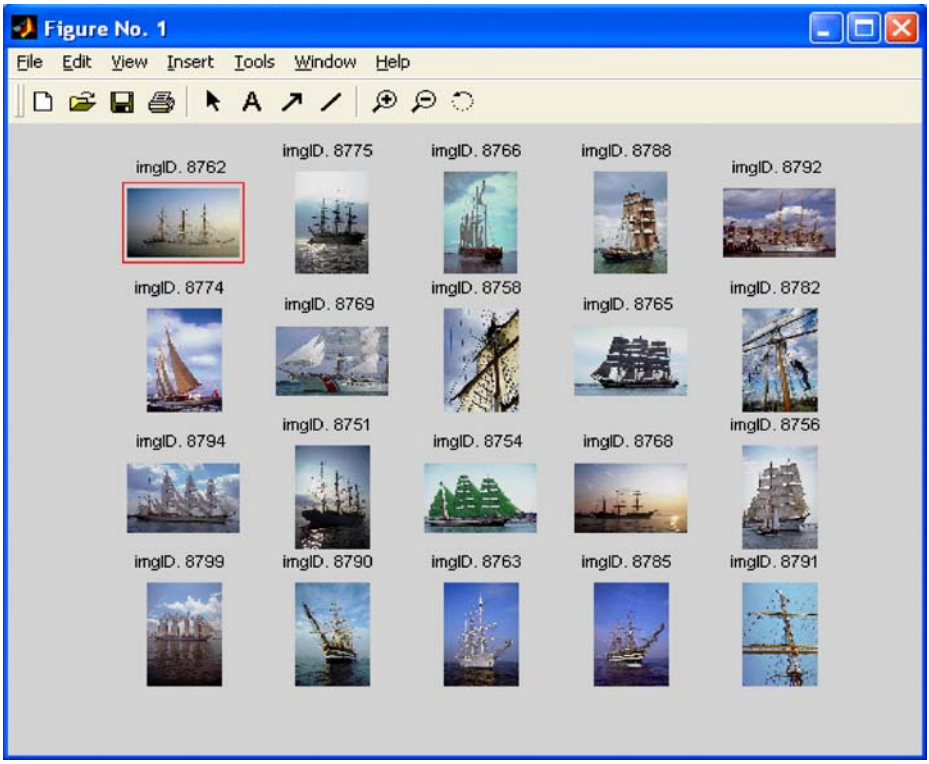


Fig. 7 A snapshot of the retrieval results

image. As can be seen from this figure, the perceptions contained in these returned images are quite similar and the ranking is reasonably good. This query example demonstrates the advantages and potentials brought by utilizing user access patterns and access frequencies.

3.3.2 Performance comparison

Off-line training processes can improve the query results dramatically by capturing users' perceptions. The more images that are covered in the training process, the more accurate the retrieval results would be. In this section, we use the accuracy-scope curve to analyze the performance of our proposed framework. In the accuracy-scope curve, the scope specifies the number of images returned to the users and the accuracy (or called *precision* in some work) is defined as the percentage of the retrieved images that are semantically related to the query image.

In our first experiment, two different training data sets are used. The first training data set includes 500 queries which cover 2,034 images in our database (denoted as MMM_500); while the other set is called MMM_1400 which contains 1,400 queries with 4,890 images being covered. We compare the overall performance of our proposed MMM framework with the "noPattern" method that does not integrate the information of user access patterns and access frequencies, and performs the

full sequential search through the image database based on the feature matrix B. Euclidean distance is used as the similarity measure in "noPattern" method. For the pre-filtering step, we choose the size of candidate pool as 400 after PCA search space reduction, which means there are only 400 images that need to do exact similarity matching. Thus, the candidate image pool constitutes only 4% of the total amount of images in database.

Figure 8 shows the results of the performance evaluation. In Figs. 8a–f, 'noPattern,' 'MMM_500' and 'MMM_1400' represent the accuracy results of the noPattern method, MMM_500, and MMM_1400, respectively. From this figure, we have the following two observations: (1) our proposed framework (MMM_500 or MMM_1400) outperforms the noPattern method in all cases, which proves that the user access patterns and access frequencies obtained from the off-line training process can capture the subjective aspects of the user concepts; and (2) with more queries issued and more images covered in the training process, more accurate retrieval results can be achieved. The reason is that the relative affinity relationships among images can be revealed more completely, and the bias caused by individual users can be corrected (more or less) as more user feedback is received. It also should be pointed out that the subjective user concepts obtained through the training process reflect the subjective perceptions of the majority of users, which might not be able to reflect the subjective perception of a particular user if such a perception deviates significantly from that of the majority. However, since the proposed framework is intended to complement the RF technique rather than replace it, the personalization feature of the on-line RF can still be preserved when it is in combined use with MMM.

In the second experiment, with the aim to demonstrate the performance of the PCA-based pre-filtering approach, we select three different sizes for the candidate

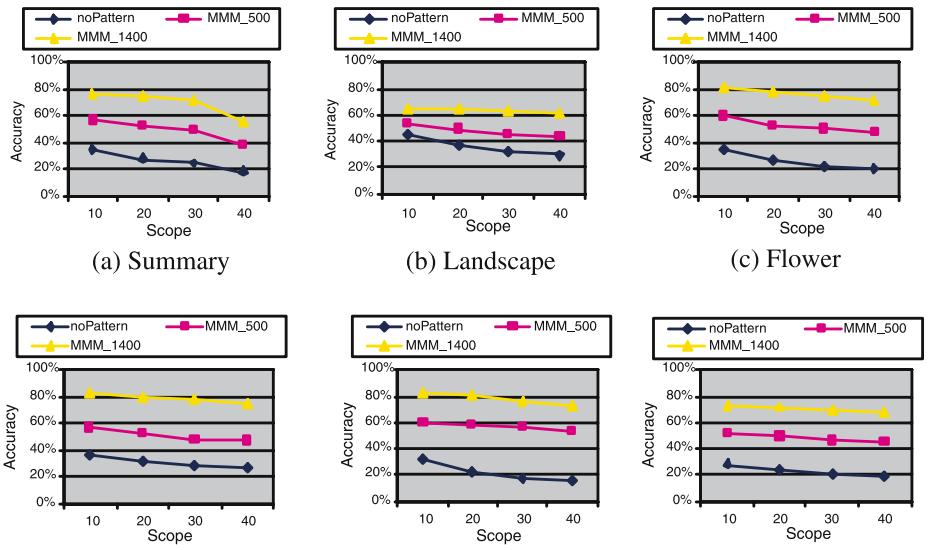


Fig. 8 Accuracy comparison between the proposed framework and the noPattern method

pool as 200, 400 and 800, called PCA_200, PCA_400, and PCA_800, respectively. In addition, the retrieval performance without PCA reduction, namely noPCA, is presented for comparison. All the results are achieved by using the training data set MMM_1400. It can be easily seen from Fig. 9 that more accurate retrieval results

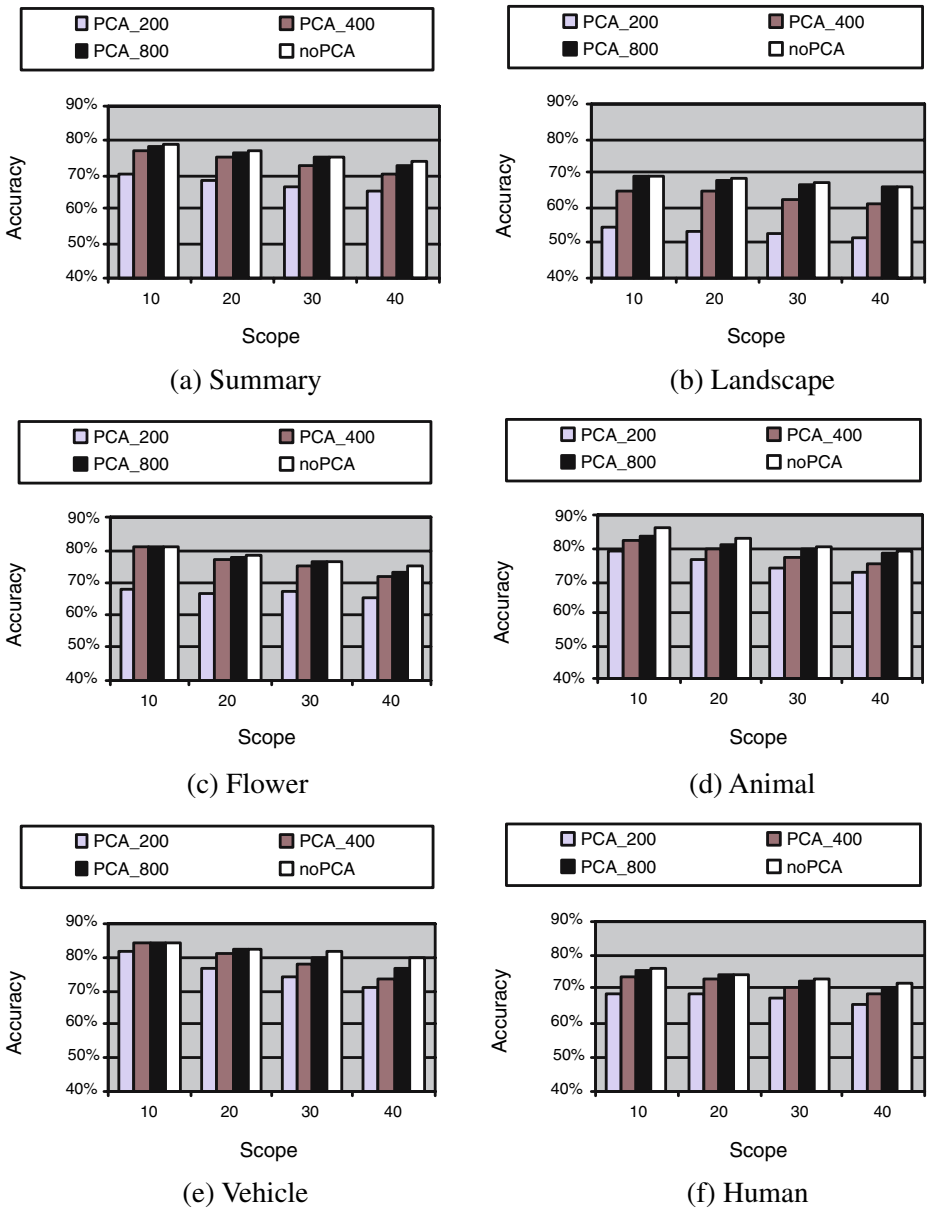


Fig. 9 Accuracy comparison of the proposed framework against three different PCA candidate pool sizes (200, 400, and 800)

can be achieved with a larger size of the PCA candidate pool, and the most accurate results are achieved when no pre-filtering process is conducted. On the other hand, the size 400 is a reasonably good choice in terms of time and space because, using only 4% of the total images in the database, the accuracy in the top 20 retrieved

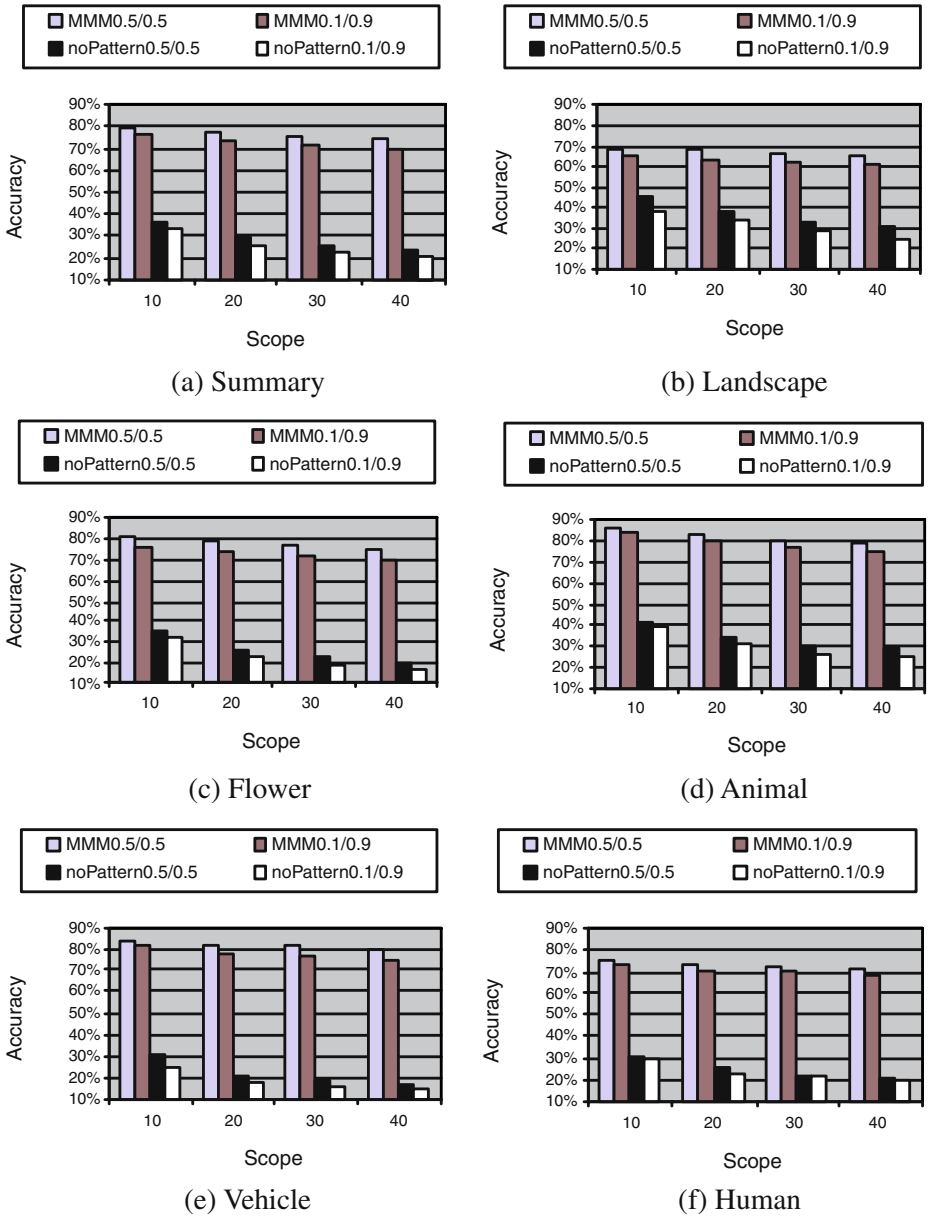


Fig. 10 Accuracy comparison between the proposed framework and the noPattern method against different feature vectors

images can reach 80%. Also, in most cases, the accuracies of PCA_400 are quite close to those of noPCA. Moreover, enlarging the pool size from 400 to 800 does not improve the performance significantly, which indicates the effectiveness of the proposed PCA-based pre-filtering method.

Then in our third experiment, we alter the feature matrix B by assigning different weights, 0.1 and 0.9, respectively, to the color features and location features instead of 0.5 each. It is worth mentioning that the weights are chosen arbitrarily and in fact any normalized vector-based image feature set can be plugged into B. Figure 10 shows the results of the performance evaluation. In Figs. 10a–f, 'MMM0.5/0.5' and 'noPattern0.5/0.5' indicate the accuracies of the MMM mechanism (without PCA pre-filtering) and the noPattern method, respectively, with the original weight assignment (0.5 each). 'MMM0.1/0.9' and 'noPattern0.1/0.9' indicate the accuracies of 'MMM' and 'noPattern' methods, respectively, with the new weight assignment (0.1 for color and 0.9 for location). From Fig. 10, we have the following observations. First, our model is flexible enough to use different feature vector sets. Second, the MMM mechanism and the noPattern method share almost the same trend, which implies that the more effective feature vectors extracted from the images, the higher accuracy our proposed framework can achieve.

Note that usually the measure *recall* is also used to evaluate the overall system performance, which is defined as the fraction of relevant images retrieved over the total number of relevant images in the database. However, as discussed in [19], the image retrieval systems are designed to return only a few relevant images and the user only browses the top few images. Thus, accuracy or *precision* is emphasized over *recall*. In addition, as the size of an image database becomes larger, manually separating the collection into relevant and non-relevant set becomes infeasible [20], which in turn prevents the accurate evaluation of *recall*. Therefore, in this work, accuracy-scope instead of *precision-recall* is used to evaluate the system performance.

It is also worth mentioning that, in addition to the training process, the users can also provide feedback during the actual retrieval process, which enables accumulative learning and can further boost the system performance.

4 Conclusions

In this paper, a Markov Model Mediator (MMM) mechanism is proposed and applied to Content-Based Image Retrieval (CBIR) to bridge the semantic gap between the low-level features and the high-level concepts. Our proposed mechanism provides the capability to learn high-level concepts and affinities among the images off-line based on the training data set, such as access patterns and access frequencies. In addition, for retrieval efficiency, a pre-filtering step enabled by the principal component analysis (PCA) is conducted before the exact similarity matching and retrieval process. Then a similarity matching process is performed to measure the similarity between the query image and each candidate image in the candidate pool, based on not only the low-level features (such as color and location features), but also the concepts obtained by training in the previous steps. The experimental results demonstrate the high accuracy and high scalability of the proposed mechanism. Our future work includes utilizing the MMM mechanism to facilitate the distributed image retrieval where the image databases are located in a distributed environment. Moreover, the proposed work can be integrated with Relevance Feedback (RF) to

explore more complete affinity relationships among images and to further refine the query results.

Acknowledgements For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260. For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562 and NSF HRD-0317692.

References

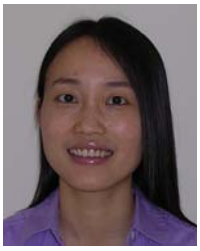
1. Aksoy S, Haralick RM (2000) A weighted distance approach to relevance feedback. In: Proceedings of the international conference on pattern recognition. Elsevier, New York, pp 812–815
2. Carson C et al (2002) Blobworld: image segmentation using expectation–maximization and its application to image querying. IEEE Trans Pattern Anal Mach Intell 24(8):1026–1038
3. Chen S-C, Sista S, Shyu M-L, Kashyap RL (2000) An indexing and searching structure for multimedia database systems. In: Proceedings of IS&T/SPIE conference on storage and retrieval for media databases, San Jose, California, pp 262–270
4. Chen S-C, Shyu M-L, Zhao N, Zhang C (2003) An affinity-based image retrieval system for multimedia authoring and presentation. In: Proceedings of 11th ACM international conference on multimedia (MM'03), Berkeley, California, pp 446–447, 2–8 November 2003
5. Chen Y, Wang JZ (2002) A region-based fuzzy feature matching approach to content-based image retrieval. IEEE Trans Pattern Anal Mach Intell 24:1252–1267
6. Cheng HD, Sun Y (2001 December) A hierarchical approach to color image segmentation using homogeneity. IEEE Trans Image Process 9(12):2071–2082
7. Cox IJ, Minka TP, Papathomas TV, Yianilos PN (2000) The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. IEEE Trans Image Process 9(1):20–27
8. Deerwester S, Dumais ST, Landauer TK, Furnas GW, Harshman RA (1990) Indexing by latent semantic analysis. J Soc Inf Sci 41(6):391–407
9. Fauqueur J, Boujemaa N (2002) Image retrieval by regions: Coarse segmentation and fine color description. Visual Information and Information Systems, pp 24–35
10. Guttman A (1984 June) R-tree: A dynamic index structure for spatial search. In: Proceedings of ACM SIGMOD, Boston, Massachusetts, pp 47–57
11. Huang X, Chen S-C, Shyu M-L (2003) Incorporating real-valued multiple instance learning into relevance feedback for image retrieval. In: Proceedings of the IEEE international conference on multimedia & expo (ICME), vol. 1, Baltimore, Maryland, pp 321–324, 6–9 July 2003
12. Jing F, Li M, Zhang H-J, Zhang B (2002) An effective region-based image retrieval framework. In: Proceedings of the 2002 ACM workshops on multimedia, Juan-les-Pins, France, pp 456–465
13. Jobson JD (1992) Applied multivariate data analysis volume II: Categorical and multivariate methods. Springer, Berlin Heidelberg New York
14. Johnson RA, Wichern DW (1998) Applied multivariate statistical analysis, 4th edn. Prentice-Hall, New Jersey
15. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, Berlin Heidelberg New York
16. Lin HC, Wang LL, Yang SN (1997) Color image retrieval based on hidden Markov models. IEEE Trans Image Process 6(2):332–339
17. Lu Y, Hu CH, Zhu XQ, Zhang HJ, Yang Q (2000) A unified framework for semantics and feature based relevance feedback in image retrieval systems. In: Proceedings of the 8th ACM international conference on multimedia, Los Angeles, California, pp 31–37
18. Ma W-Y, Zhang HJ (1999) Content-based Image Indexing and Retrieval. Handbook of multimedia computing, Chapter 13, CRC Press, Boca Raton, Florida
19. Natsev A, Rastogi R, Shim K (2004) WALRUS: A similarity retrieval algorithm for image databases. IEEE Trans Knowl Data Eng 16(3):301–316
20. Ortega M, Rui Y et al (1998) Supporting ranked boolean similarity queries in MARS. IEEE Trans Knowl Data Eng 10(6):905–925
21. Rabiner LR, Huang BH (1986 January) An introduction to hidden Markov models. IEEE ASSP Mag 3(1):4–16
22. Rui Y, Huang TS (1999) A novel relevance feedback technique in image retrieval. In: Proceedings of the 7th ACM international conference on multimedia, Part 2, Orlando, Florida, pp 67–70

23. Shyu M-L, Chen S-C, Kashyap RL (2000) A probabilistic-based mechanism for video database management systems. In: Proceedings of IEEE international conference on multimedia and expo (ICME'00), New York, USA, pp 467–470
24. Shyu M-L, Chen S-C, Haruechaiyasak C, Shu C-M, Li S-T (2001) Disjoint web document clustering and management in electronic commerce. In: Proceedings of the 7th international conference on distributed multimedia systems (DMS'01), Tamkang University, Taipei, Taiwan, pp 494–497
25. Shyu M-L, Chen S-C, Rubin SH (2004) Stochastic clustering for organizing distributed information sources. IEEE Trans Syst Man Cybern, Part B, Cybern 34(5):2035–2047
26. Su Z, Li S, Zhang H (2001) Extraction of feature subspaces for content-based retrieval using relevance feedback. In: Proceedings of the 9th ACM international conference on multimedia (MM'01), Ottawa, Canada, pp 98–106
27. Wiederhold G (March 1992) Mediators in the architecture of future information systems. IEEE Computers 25:38–49
28. Zhang DS, Lu G (2002 August) Generic fourier descriptors for shape-based image retrieval. In: Proceedings of IEEE international conference on multimedia and expo (ICME'02), vol. 1. Lausanne, Switzerland, pp 425–428
29. Zhang Q, Goldman SA, Yu W, Fritts J (2002 July) Content-based image retrieval using multiple-instance learning. In: Proceedings of the 19th international conference on machine learning, Sydney, Australia, pp 682–689

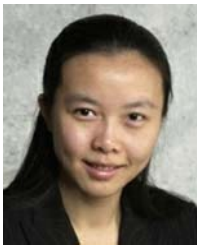
Dr. Mei-Ling Shyu received her Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN in 1999, and three Master's degrees from Computer Science, Electrical Engineering, and Restaurant, Hotel, Institutional, and Tourism Management from Purdue University. She has been an Associate Professor in the Department of Electrical and Computer Engineering (ECE) at the 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 co-authored more than 130 technical papers published in various prestigious journals, refereed conference/symposium/workshop proceedings, and book chapters. She is/was the guest editor of several journal special issues.

Dr. Shu-Ching Chen received his Ph.D. from the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA in December 1998. He also received Master's degrees

in Computer Science, Electrical Engineering, and Civil Engineering from Purdue University. He has been an Associate Professor in the School of Computing and Information Sciences (SCIS), Florida International University (FIU) since August, 2004. Prior to that, he was an Assistant Professor in SCIS at FIU dating from August 1999. His main research interests include distributed multimedia database systems and multimedia data mining. Dr. Chen has authored and co-authored more than 160 research papers in journals, refereed conference/symposium/workshop proceedings, and book chapters. In 2005, he was awarded the IEEE Systems, Man, and Cybernetics Society's Outstanding Contribution Award. He was also awarded Excellence in Graduate Mentorship Award from FIU in 2006, University Outstanding Faculty Research Award from FIU in 2004, Outstanding Faculty Service Award from SCIS in 2004 and Outstanding Faculty Research Award from SCIS in 2002.



Min Chen received her bachelor's degree in Electrical Engineering from Zhejiang University in China. She is currently a Ph.D. candidate in the School of Computing and Information Sciences (SCIS) at Florida International University (FIU), Miami, FL, USA. Her research interests include distributed multimedia database systems, image and video database retrieval, and multimedia data mining. She has authored and co-authored 20 technical papers published in various prestigious journals, refereed conference/workshop proceedings and book chapters. She is the recipient of several awards, including a Presidential Fellowship and the Best Graduate Student Research Award from SCIS at FIU.



Dr. Chengcui Zhang is an Assistant Professor of Computer and Information Science at University of Alabama at Birmingham (UAB) since August 2004. She received her Ph.D. from the School of Computer Science at Florida International University (FIU), Miami, FL, USA in August 2004. She also received her bachelor and master degrees in Computer Science from Zhejiang University in China. Her research interests include multimedia databases, multimedia data mining, Bioinformatics, and GIS data filtering. She is the recipient of several awards, including the IBM UIMA Award, the UAB ADVANCE Junior Faculty Research Award from the National Science Foundation, UAB Faculty Development Award, and the Presidential Fellowship and the Best Graduate Student Research Award from FIU.

Kanoksri Sarinnapakorn is currently a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Miami, Coral Gables, FL, USA.