A Unified Framework for Image Database Clustering and Content-based Retrieval

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
1-305-284-5566
shyu@miami.edu

Shu-Ching Chen, Min Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
1-305-348-3480
{chens, mchen005}@cs.fiu.edu

Chengcui Zhang
Department of Computer & Information Science
University of Alabama at Birmingham
Birmingham, AL 35294, USA
1-205-934-2213
zhang@cis.uab.edu

ABSTRACT
With the proliferation of image data, the need to search and retrieve images efficiently and accurately from a large image database or a collection of image databases has drastically increased. To address such a demand, a unified framework called *Markov Model Mediators* (MMMs) is proposed in this paper to facilitate conceptual database clustering and to improve the query processing performance by analyzing the summarized knowledge. The unique characteristics of MMMs are that it provides the capabilities of exploring the affinity relations among the images at the database level and among the databases at the cluster level respectively, using an effective data mining process. At the database level, each database is modeled by an intra-database MMM which enables accurate image retrieval within the database. Then the conceptual database clustering is performed and cluster-level knowledge summarization is conducted to reduce the cost of retrieving images across the databases. This framework has been tested using a set of image databases, which contain various numbers of images with different dimensions and concept categories. The experimental results demonstrate that our framework achieves better retrieval accuracy via inter-cluster retrieval than that of intra-cluster retrieval with minimal extra effort.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Clustering, Retrieval models.*

General Terms
Algorithms, Experimentation.

Keywords
Content-based Image Retrieval (CBIR), Image Database Clustering, Markov Model Mediators (MMMs).

MMDB'04, November 13, 2004, Washington, DC, USA.

1. INTRODUCTION
At present, there are some commercial image search engines available on the Web such as Google Image Search and AltaVista Image Search. Most of them employ only the keyword based search and hence the retrieval result is not satisfactory. With the advances in image processing, information retrieval, and database management, there have been extensive studies on content-based image retrieval (CBIR) for large image databases [10]. CBIR systems retrieve images based on their visual contents. Earlier efforts in CBIR research have been focused on effective feature representations for images. The visual features of images, such as color [32], texture [5][9], and shape features [18][38] have been extensively explored to represent and index image contents, resulting in a collection of research prototypes and commercial systems [6][7]. There are also some integrated search engines employing both the keyword-based search and content-based image retrieval (e.g., Image Rover [21]).

However, due to the semantic gap between the high-level concepts and low-level image feature representations, it is not easy to select the appropriate feature representations to effectively describe the high-level concepts in a query image. In addition to that, the user's subjective preference for a query image may vary from one user to another user. To solve this problem, more recent studies in CBIR have focused on the approach based on the relevance feedback (RF) technique [10][14][17][35] which is an automatic query refining process and requires the interactive user feedback to help bridge the semantic gap. Furthermore, in order to better capture the user's focus of interest or attention in a query image, object-based retrieval has been proposed to retrieve images at the object/region level, where an image is first segmented into a number of regions/objects and the similarity measures are then applied to individual objects [1][2][9][16]. Though object-based CBIR performs retrieval at a finer level, there is still a gap between the low-level presentation of the images and the high-level subjective semantics. The integrated use of relevance feedback and object-based retrieval can be found in some more recent research efforts [37].

Nowadays, owing to recent advances in high-speed networks and large capacity storage devices, multimedia information, typically image information, is growing rapidly across the Internet and elsewhere. Online image sources and databases, consisting of millions of images, have been used in many applications. The explosive growth in the amount and complexity of the image data

has created an emergent need for efficient and accurate search and retrieval from a large image database or a collection of image databases.

There have been extensive studies on the indexing techniques and data structures to speedup the image search process so that the relevant images can be located quickly. A survey of the techniques and data structures for efficient multimedia retrieval based on similarity, such as A-tree [19] and M-tree [3], was given in [13]. However, most of the existing work on data indexing and data structures are conducted at a single data set level. In contrast, clustering [8] is one of the most useful knowledge discovery techniques for identifying the correlations in large data sets. There are different types of clustering algorithms in the literature such as the partitioning clustering algorithms [15], hierarchical clustering methods where a representative of the data set is computed based on their hierarchical structures [33], and syntactic methods which are based solely on the static structure of the information source [11]. One of the disadvantages of the syntactic method is that it ignores the actual access patterns. Another type of method collects the statistics pertaining to the access patterns and conducts partitioning based on the statistics [25]. In the literature of CBIR, image data clustering is conducted mostly in the form of image classification. An image classification framework based on class keywords and image features was proposed in [36]. In [10], the authors proposed an approach which allows multipoint query in relevance feedback by using the clustering technique. This method clusters the sets of relevance points and chooses the centroids of the clusters as their representatives [10]. Then a multipoint query is constructed by using only a small number of good representatives. As another example, a photo image clustering method was proposed in [4] with the ultimate goal of creating the clusters which minimize the search time for a user to locate a photo of interest.

Most of the above image classification approaches apply clustering on a single database level. However, in a distributed environment, the number of image databases has increased enormously, and the query results may need to access several image databases at different locations. The principles of a distributed database system include executing a query as fast as possible or with as little cost as possible [12] and allowing transparent location-independent accesses to the users in the applications. Hence, there is a strong need to analyze and discover summarized knowledge at the database level, i.e., database clustering [29]. Database clustering is to group related databases in the same cluster such that the intra-cluster similarity is high and inter-cluster similarity is low. Here, two databases are said to be related in the sense that they either are often simultaneously accessed, or contain similar information, which matches the relevance feedback based CBIR perfectly. Intuitively, in relevance feedback, two image databases are said to be related if the set of relevant images contains images from both of the two databases. In other words, these two image databases are accessed simultaneously through the relevant image set because they have some information in common given the query image. In addition, the clustering techniques allow hierarchical accesses for retrieval and improve retrieval efficiency [20].

In brief, an ideal CBIR system should be both effective and efficient. The efficiency requirement is especially critical in a distributed database environment. However, the previous studies tend to focus on only one aspect of these two requirements. As in our earlier effort to meet both the effectiveness and efficiency requirements [23], we have proposed a conceptual image database clustering strategy based on a mechanism called *Markov Model Mediators* (MMMs). The effectiveness and robustness of MMMs have been extensively explored in our previous studies [26] [27][28][30]. More recently, we proposed the use of the MMM mechanism for general-purpose database clustering [31] and content-based image retrieval within a single image database [24]. The work proposed in [23] was an extension to our previous work by enabling conceptual image database clustering and content-based image retrieval at both intra-database and inter-database levels. The MMMs are used in [23] to facilitate conceptual image database clustering and improve query processing performance by analyzing the summarized knowledge at both intra-database and inter-database levels. However, to be more precisely, the inter-database retrieval enabled in [23] actually limited the search scope to a specific cluster to which the query image belongs. However, for those image objects that are not close to the majority or the representatives of the images in that cluster, their most closely related images might not exist in the same cluster. In addition, a predefined cluster size or a predefined number of clusters can also lead to the same situation where two closely related images (belonging to different databases) are put into two clusters. If we restrict the search scope to a single cluster, then there would be no chance that we can adjust the query performance for those images.

To solve this problem, in this paper, we propose an adaptive cluster-based image retrieval framework to achieve a better trade-off between the retrieval accuracy and the searching cost. The key idea here is to perform an additional level of knowledge summarization at the cluster level to further discover the correlations among clusters. Thus, the cluster-based image retrieval can be conducted in an adaptive way of either intra-cluster search or inter-cluster search, depending on the cluster-level knowledge being learned. First, a clustering strategy based on the MMM mechanism is used to partition the image databases into a set of conceptual image database clusters based on the summarized knowledge at the database level. Then the cluster-level knowledge summarization can be conducted. Image retrieval is performed at either the intra-cluster or inter-cluster level, based on the obtained cluster-level knowledge. The strategy here is to maximize the overall retrieval performance without sacrificing too much efficiency (i.e., introducing too many inter-cluster accesses).

The contributions of the proposed work are summarized as follows:

1. The major contribution of this study is that the proposed unified framework takes into consideration both the effectiveness and efficiency requirements in a distributed database environment to address this problem. The conceptual database clustering strategy generates conceptual image database clusters to reduce the search cost. In the meanwhile, a reasonable amount of inter-cluster accesses are allowed in this framework in order to boost the query performance. In addition, the decision process of choosing either intra-cluster retrieval or inter-cluster retrieval is automated, only depending on the summarized knowledge at the cluster level.

2. One of the unique contributions of this study is that the MMM mechanism is used as the base for both database-level and cluster-level knowledge discovery and summarization, which allows us to form a unified, consistent and hierarchical structure. Therefore, it has the advantages of easy-to-maintain and scalability.

3. The conceptual database clustering strategy adopted in this framework has the following advantages over physical database clustering. First, instead of actually moving around the databases, conceptual modeling allows for an abstract representation of the member image databases without physically moving them, which is more realistic due to the autonomous nature of each image database. Second, in conceptual clustering, groups of image databases that show similarities in their data access behavior are conceptually clustered together, which allows us to gain a better understanding of the image databases by revealing their similarity and semantic relationships. Essentially, since a set of image databases with close relationships are put in the same image database cluster and are required to be consecutive on some query access path, the number of platter (cluster) switches for data retrieval with respect to the queries can be reduced. This can significantly improve system response time as well as query performance.

The rest of the paper is organized as follows. The overall architecture of the proposed inter-cluster image retrieval is introduced in Section 2, followed by the discussions of knowledge summarization at both the database level and cluster level, and a brief discussion of conceptual database clustering as well. Experimental studies are presented in Section 3. Section 4 concludes this paper.

2. INTER-CLUSTER IMAGE RETRIEVAL

Currently, many research efforts have been carried out to reduce the query search space via a clustering process. Once the clusters are obtained, the retrieval process is conducted within a certain cluster for a specific query, namely intra-cluster retrieval. However, as the principle of the clustering is to maximize the intra-cluster similarity and minimize the inter-cluster similarity by taking into account all the objects in the clusters (the so-called majority vote), it might not be an optimal solution for some specific objects. In other words, for an object O_i in cluster C_i, its most related object(s) might not be included in C_i. In particular, this issue remains as a challenge for the cluster algorithms with a predefined cluster size (e.g., the single-link clustering method) or a predefined number of clusters (e.g., the k-means algorithm), where two related objects are partitioned into two clusters due to the limited cluster size or the predefined number of clusters. To address this issue, we propose an adaptive inter-cluster image retrieval framework to achieve the best trade-off between the retrieval accuracy and the searching cost.

The flowchart of the proposed scheme is demonstrated in Figure 1. As can be seen from this figure, the framework contains four major components, namely database-level knowledge summarization, clustering process, cluster-level knowledge summarization, and inter-cluster image retrieval process. In brief, given a set of image databases and the associated log data, a data mining process is conducted for intra-database knowledge discovery and summarization. Then the similarity measures among the databases are calculated via probabilistic reasoning. With the summarized database-level knowledge, a conceptual database clustering process is carried out. Note that our framework is flexible in the sense that any database clustering strategy can be easily plugged in, as long as it has the capability to partition the databases into a set of database clusters. However, as presented in [23], our conceptual database clustering process is highly effective. Thereafter, cluster-level knowledge summarization is applied to discover the intra-cluster knowledge and explore the inter-cluster relationships. Finally, image retrieval is conducted in the intra-cluster or inter-cluster level based on the obtained cluster-level knowledge. The detailed discussions of these components are presented in the following subsections.

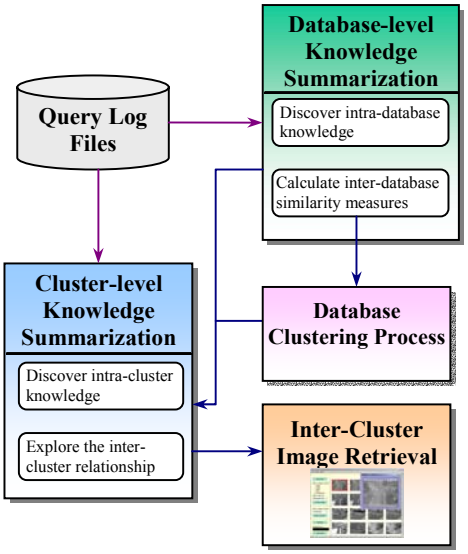

Figure 1. Flowchart of the proposed inter-cluster image retrieval framework

2.1 Database-level Knowledge Summarization

As mentioned above, in this step, the intra-database knowledge is discovered and summarized. In addition, the inter-database similarity values are calculated.

2.1.1 Intra-database Knowledge Discovery

In our framework, probabilistic networks are constructed via the affinity-based data mining process for each database, which is modeled by the intra-database MMMs. The details of the definitions of MMMs and the affinity-based data mining process can be found in our previous work [24]. For the clarification purpose, some of the basic concepts are discussed as follows.

Essentially, an MMM is a stochastic finite state machine with a stochastic output process attached to each state to describe the probability of the occurrences of the output symbols (states). It contains three major parameters as described in Table 1.

Intuitively, the parameter $\mathcal{B}$ contains the low-level feature values for the images in the database, while the parameters $\mathcal{A}$ and Π are two probability distributions obtained by applying the affinity-based data mining process to the query log file, which consists of the information listed in Table 2.

Table 1. Three major parameters in MMM

Parameters	General Definition	Extended Meanings for Image Database
$\mathcal{A}$	The state transition probability distribution	Indicates the affinity relationships among images
$\mathcal{B}$	The observation symbol probability distribution	Represents the image feature values
Π	The initial state probability distribution	Indicates the likelihood of an image being selected as the query image

Table 2. Useful information in the query log file

Name	General Definition	Explanation
use	User access patterns	Denotes the co-occurrence relationship among images given historical user queries
access	User access frequencies	Denotes how often a certain query was issued

The basic idea of the affinity-based data mining process is that the more the two images m and n are accessed together, the higher relative affinity relationship they have, i.e., the probability that a traversal choice to state (image) n given the current state (image) is in m (or vice versa) is higher. Such a probabilistic network is of great importance because image retrieval is actually a process to explore the relationships between the query image and the other images in the database(s).

2.1.2 Inter-database Similarity Measure Calculation

In our framework, the inter-database similarity measure is critical for both the conceptual clustering process and the cluster-level knowledge discovery components (as shown in Figure 1).

The main idea of measuring the inter-database similarity is that two image databases are considered related (or similar) in the sense that they are either accessed together frequently or have similar images. Let $S(d_i, d_j)$ denote the similarity measure between each pair of image databases d_i and d_j in the distributed database system, which is calculated via a probabilistic reasoning procedure using the $\mathcal{A}$, $\mathcal{B}$, and Π parameters [29][31]. The calculated similarity value is then used to measure how well these two image databases together match the observations obtained from the queries in the query log file.

2.2 Conceptual Database Clustering Process

In the traditional databases, data clustering places related or similar records or objects in the same page on disks for performance purposes. A good clustering method ensures that the intra-cluster similarity is high and the inter-cluster similarity is low. However, in a distributed environment where the number of databases increased drastically, the workloads associated with complex queries are quite expensive since they tend to access a huge amount of data records from a set of distributed databases. Therefore, a conceptual database clustering process is necessary to reduce the cost of communication and query processing. Similar to data clustering, database clustering is to group related image databases in the same image database cluster such that the intra-cluster similarity is high and the inter-cluster similarity is low.

In our framework, once the similarity measures are obtained, the probabilistic network of the image databases is constructed, where the branch probability $P_{i,j}$ for the nodes (image databases) i and j is calculated from those similarity measures. More specifically, the calculation is carried out by normalizing the similarity values per row to indicate the branch probabilities from a specific node to all its accessible nodes. Then the stationary probability ϕ_i for each node i of the probabilistic network is computed from $P_{i,j}$, which denotes the relative frequency of accessing node i (the i^{th} image database) in the long run [29][31].

As discussed in [23], our conceptual database clustering strategy is traversal-based and greedy. The basic steps of the proposed conceptual database clustering process are illustrated in Table 3.

Table 3. The proposed conceptual database clustering process

1. Set the value for c, where c is the size of the conceptual image database cluster.
2. Start a new cluster CC with an image database x_i, which has the largest stationary probability in the set of image databases X.
3. $X \leftarrow X - x_i$
4. Check if $|X| > 0$. If false, go to Step 7.
5. Check if the number of image databases in the current cluster is less than c. If false, output the current cluster CC and go to Step 2.
6. Add an image database x_j to CC, which has the largest stationary probability in X and is accessible from the current member database(s) in CC. $x_i \leftarrow x_j$. Go to Step 3.
7. Output the current cluster CC and stop.

2.3 Cluster-level Knowledge Summarization

Once the conceptual image database clusters are obtained, the probabilistic networks at the intra-cluster level should be constructed. In addition, as discussed earlier, our conceptual clustering algorithm is performed with a predefined cluster size. Therefore, though the intra-cluster retrieval produces quite promising results (as presented in [23]) due to the effectiveness of our proposed clustering algorithm, there might be the cases that for some queries, inter-cluster retrieval is a necessity to further boost the retrieval accuracy with minimal extra effort.

2.3.1 Intra-cluster Knowledge Discovery

The construction of the probabilistic network at the intra-cluster level (or intra-cluster MMM) is quite similar to the one at the intra-database level. Especially, the calculations of the $\mathcal{B}$ and Π parameters are very similar except the scope of the images. That is, the only difference is that the scope of an intra-cluster MMM is defined in a cluster instead of a database [23]. For the parameter $\mathcal{A}$, if a conceptual image database cluster CC contains

only one image database, then no re-calculation of $\mathcal{A}$ is needed. However, in most cases, a conceptual image database cluster contains two or more image databases so that an intra-cluster level probabilistic network should merge those intra-database level networks in *CC*. In other words, $\mathcal{A}$ for an intra-cluster MMM needs to be re-calculated in order to satisfy the properties/requirements of an MMM.

$\mathcal{A}$ is re-calculated as follows. For any images $s,t \in CC$, if there exists a link from s to t, the relative affinity measure between s and t is computed. Let λ_i and λ_j denote two intra-database MMMs for image databases d_i and d_j, where $j \neq i$ and $\lambda_i, \lambda_j \in CC$.

- $a_{s,t}$: the state transition probability of an intra-database MMM;
- $a'_{s,t}$: the state transition probability of an intra-cluster MMM;
- $p_{s,t}$: the probability that λ_i goes to λ_j with respect to s and t;
- p_s: the probability that λ_i stays with respect to s.

$$aff_{s,t} = \sum_{k=1}^{q} use_{s,k} \times use_{t,k} \times access_k \quad (1)$$

$$f_{s,t} = \frac{aff_{s,t}}{\sum_{s \in CC} \sum_{t \in CC} aff_{s,t}} \quad (2)$$

$$p_{s,t} = \frac{f_{s,t}}{\sum_{n \in CC} f_{s,n}} \quad (3)$$

$$p_s = 1 - \sum_{t \notin \lambda_i} p_{s,t} \quad (4)$$

The steps for determining $a'_{s,t}$ where $s, t \in CC$ are:

1. If both $s,t \in \lambda_i$, then $a'_{s,t} = p_s a_{s,t}$.
2. If $s \in \lambda_i$ and $t \notin \lambda_i$, then $a'_{s,t} = p_{s,t}$.
3. Repeat the above steps for all intra-database MMMs in *CC*.

Originally, $\Sigma_t a_{s,t} = 1$. Now, we need to check whether the new state transition probability distribution satisfies this requirement, too. For any image $s \in \lambda_i$,

$$\begin{aligned} \sum_t a'_{s,t} &= \sum_{t \in \lambda_i} a'_{s,t} + \sum_{t \notin \lambda_i} a'_{s,t} = \sum_{t \in \lambda_i} p_s a_{s,t} + \sum_{t \notin \lambda_i} p_{s,t} \\ &= p_s \sum_{t \in \lambda_i} a_{s,t} + \sum_{t \notin \lambda_i} p_{s,t} = p_s + \sum_{t \notin \lambda_i} p_{s,t} \\ &= 1 - \sum_{t \notin \lambda_i} p_{s,t} + \sum_{t \notin \lambda_i} p_{s,t} = 1 \end{aligned} \quad (5)$$

The newly calculated probabilities in $\mathcal{A}$ are then attached to the arcs to indicate the probabilities that go from one state (image) to another state (image) within the image database cluster.

2.3.2 Exploring the Inter-cluster Relationships

For a query image q_j in database d_i ($d_i \in CC$), *CC* is used as its main cluster. Generally, the intra-cluster retrieval in the main cluster can produce a reasonably good query result. However, as discussed earlier, in some cases, better retrieval accuracy may be achieved via inter-cluster retrieval than that of intra-cluster retrieval. There is a trade-off between the retrieval accuracy and the search cost. Hence, the motivation of the proposed framework is to utilize inter-cluster retrieval to boost the retrieval accuracy with minimal extra effort.

In order to determine which cluster(s) should be accessed during the retrieval process, the inter-cluster relationships between the main cluster and the other clusters, given a specific database, should be explored first. Then a screening scheme is performed to determine the search space.

With the availability of the similarity values among the databases and the clustering results, the relationships between the main cluster and the other clusters, given a specific database, is defined as follows.

Definition 1. Assume CC_m and CC_n are two clusters. Let $R^{d_i}_{CC_m,CC_n}$ denote the relationship between CC_m and CC_n given a database $d_i \in CC_m$, then

$$R^{d_i}_{CC_m,CC_n} = \max_{d_j \in CC_n} S(d_i, d_j)$$

where $S(d_i, d_j)$ represents the similarity value between database d_i and d_j ($d_j \in CC_n$).

Once the relationships between the main cluster and the other clusters are explored, a screening scheme is defined to determine the search space, namely the super cluster SC_{d_i}, as shown in Table 4.

Table 4. The screening scheme

Let $D = \{d_1, d_2, \ldots d_p\}$ be a set of image databases in the distributed environment, $S(d_i, d_j)$ ($1 \le i, j \le p$) be the similarity value between image databases d_i and d_j, $CC = \{CC_1, CC_2, \ldots, CC_C\}$ be the resulted clusters and $R^{d_i}_{CC_m,CC_n}$ ($1 \le i \le p$, $1 \le m, n \le C$) be the inter-cluster relationships between clusters CC_m and CC_n given $d_i \in CC_m$.

1. $SC_{d_i} = CC_m$
2. Define $r^{d_i}_{CC_m,CC_n}$ as the relative relationships between clusters CC_m and CC_n, where
$$r^{d_i}_{CC_m,CC_n} = R^{d_i}_{CC_m,CC_n} \Big/ \sum_{1 \le j \le p} S(d_i, d_j)$$
3. If $r^{d_i}_{CC_m,CC_n} \ge Th_\gamma$, then $SC_{d_i} \leftarrow SC_{d_i} \cup d_j$, where $d_j \in CC_n$ and $S(d_i, d_j)$ is the same as $R^{d_i}_{CC_m,CC_n}$.

Here, Th_γ ($0 \le Th_\gamma \le 1$) is a predefined threshold. As we can see from the table, the smaller Th_γ is, the more search efforts are required. In fact, it indicates the relative importance level between the retrieval efficiency and effectiveness. In our experimental settings, Th_γ is equal to 0.5, i.e., at most two clusters including the main cluster will be accessed and searched. The probabilistic network for the super cluster SC_{d_i} can be easily constructed by changing the image scopes in the equations presented in Section 2.3.1. Note that in many cases, the super cluster is actually the main cluster, thus no re-calculation is required.

2.4 Inter-cluster Image Retrieval

Information utilization concerns how the relevant information can be identified and retrieved from the resource repositories with respect to a user query. In our previous works [23][24], an effective retrieval algorithm is proposed to calculate the similarity score between the query image q and the other images in a single database and a cluster, respectively. In this study, it is further extended to support inter-cluster image retrieval.

As discussed in [24], among the three major parameters of the probabilistic network (or MMM), the parameter $\mathcal{B}$ is applied to represent the low-level image features, whereas the parameters $\mathcal{A}$ and Π contain the high-level user concepts by mining the query log files. Therefore, given a query image q, the edge weights from image q to the other images in the search space (a database or a cluster in the previous studies) are computed by taking into consideration not only the image features (represented by $\mathcal{B}$), but also the high-level user concepts (denoted by $\mathcal{A}$ and Π). Then the similarity scores are directly derived from the edge weights, where a larger score suggests the more similarity between them.

In this study, once the set of conceptual image database clusters is constructed and the cluster-level probability distributions are obtained, the probabilistic network can be constructed for the super cluster as discussed in the previous subsection. Consequently, it is obvious that the previously proposed retrieval algorithm can be applied for inter-cluster image retrieval by expanding the search space to a super cluster. Note that depending on the composition of the super cluster, the cluster-based image retrieval is adaptively conducted in the intra-cluster level or the inter-cluster level.

3. EXPERIMENTS AND EVALUATION

3.1 Experimental Design

In our experiments, three groups of experiments are designed to examine the effectiveness of our framework in cluster-based image retrieval. In addition, in order to demonstrate the efficiency of this framework, the time complexity and search space are also analyzed. In these experiments, the feature set extracted from each image contains 13 color descriptors and 6 texture descriptors. The training system has been implemented [24] to obtain the query log files.

3.2 Performance on Cluster-based Image Retrieval

The first experiment is designed to demonstrate the effectiveness of our proposed framework (for short, MMM) on the intra-database level image retrieval. In fact, the intra-database level image retrieval is a special case of the cluster-based retrieval, where the super cluster contains only one database. An image database with 10,000 color images is used to test the retrieval performance. The training data set contains 1,400 queries issued to the image database in the query log file. For the comparison purpose, the performance of the *nearest-neighbor retrieval* method in [22], which retrieves the most similar images in the feature space with respect to the query image, is measured. This kind of methods includes retrievals using the serial search or index trees such as R-tree and its variants, or using traditional template-based clustered databases. However, since only the effectiveness of the framework is concerned at this point, the nearest-neighbor retrieval using the serial search is implemented in this study. Totally, 80 randomly chosen images belonging to five semantic categories: *landscape, flower, animal, vehicle and human*, with 16 images per category, are used as the query images. The performance comparison is summarized in Table 5.

As shown in Table 5, 'MMM' represents the results by using our proposed mechanism and 'Nearest-Neighbor' denotes the results by using nearest neighbor retrieval. In addition, 'scope' specifies the number of images returned and the average retrieval accuracies for 80 query images are compared in different scopes. Here, the retrieval accuracy is defined as the ratio of the number of the relevant images within the top s images, which is a commonly used measure of retrieval effectiveness in CBIR [34]. As can be seen from this table, our framework outperforms the nearest-neighbor approach in all the 5 image categories. In addition, it should be noted that though with only 19 global-level features (color and texture features) in the experiment, the average accuracy of the first 10 retrieved images can reach above 87%, which demonstrates the effectiveness of MMM for the intra-database level image retrieval.

Table 5. Accuracy comparison between MMM and the nearest-neighbor retrieval method in the intra-database level

	Scope			
	10	20	30	40
MMM	87%	79%	70%	64%
Nearest-Neighbor	44%	30%	24%	21%

The second experiment is to demonstrate the effectiveness of our framework in the intra-cluster level. In other words, the search space can be reduced dramatically without decreasing the accuracy significantly. A set of image databases is tested in this experiment, where totally 18,700 images are stored in 12 image databases, numbered from 1 to 12, with the number of images in each database ranging from 1,350 to 2,250. The training data set is also obtained from the 1,400 queries in the log file.

Table 6: Cluster information and the issued queries

Cluster No.	Member Databases #s	Number of Images	Number of Queries
1	4, 12	2650	15
2	11,3	3250	26
3	9,8	3650	38
4	5,1	3050	44
5	2,7	2550	17
6	10,6	3550	32
Total	**12 databases**	**18,700 *images***	**172 queries**

Without loss of generosity, during the clustering process, the size of the conceptual image database cluster, denoted as c, is set to 2. Table 6 shows the number of images contained in each cluster and the number of randomly issued testing queries to each cluster. As can be seen from this table, image database No. 4 has the largest stationary probability of all the 12 databases at the beginning. Therefore, it starts the first cluster. Then within the image databases which are accessible from image database No. 4 in the probabilistic network, image database No. 12 has the largest stationary probability. It should be pointed out that the

sequence numbers of the member image databases in the clusters shown in Table 6 do not follow the orders (for instance, image cluster 2 contains image databases No. 11, 3 instead of databases No. 3, 11). It indicates that image database No. 11 has a larger stationary probability than those of image database No. 3 and the other remaining image databases. Therefore, once the number of member image databases in cluster 1 reaches 2, image database No. 11 is selected to start a new cluster. The process continues until all the image databases are assigned to a certain image database cluster conceptually. Note that the number of member image databases in each image database cluster may not reach *c* if the number of reachable image databases from the member image databases in the cluster is less than *c*. The image retrieval process in the intra-cluster level can be conducted once the clustering process finishes and the intra-cluster MMMs are constructed.

As the results shown in the first experiment, our framework is effective in terms of retrieving images from one database. Therefore, in order to demonstrate the its effectiveness for intra-cluster retrieval, a single database, namely *DB_ALL*, is constructed to have all the 18,700 images and accessed by the same set of queries listed in Table 6 (totally 172 queries), for the comparison purposes. Table 7 shows the comparison results.

Table 7. Accuracy comparison between 'Intra_Cluster' and 'DB_ALL'

	Scope			
	10	20	30	40
Intra_Cluster	76%	73%	70%	68%
DB_ALL	84%	81%	79%	77%

Here, 'Intra_Cluster' represents the average retrieval accuracy achieved by issuing the queries to each of the clusters, while 'DB_ALL' denotes the results obtained by conducting the retrieval in *DB_ALL*. As can be seen from this table, the results are quite promising considering that the search space can be reduced dramatically (about 1/6 of the whole search space) without the significant decreases in the accuracy (with the average decrease of about 8% compared to that of 'DB_ALL').

In the third experiment, the inter-cluster access is applied by setting Th_γ to 0.5. This experiment is conducted using the same set of image databases and query log files as the ones used in the second experiment. Table 8 lists the super cluster SC_{d_i} for a database d_i.

Table 8. The super clusters for each database

Database #	SC_{d_i}	Database #	SC_{d_i}
db_1	clu_4 ∪ db_4	db_7	clu_5 ∪ db_9
db_2	clu_5 ∪ db_12	db_8	clu_3 ∪ db_4
db_3	clu_2	db_9	clu_3 ∪ db_4
db_4	clu_1	db_10	clu_6 ∪ db_4
db_5	clu_4 ∪ db_4	db_11	clu_2
db_6	clu_6 ∪ db_2	db_12	clu_1

In Table 8, 'db_1' denotes image database No. 1, and 'clu_1' represents the image database cluster No. 1. As can be seen from this table, for some databases (i.e., db_3, db_4, db_11, and db_12), the super clusters are the same as their main clusters, whereas for the other databases, the inter-cluster accesses are required. Therefore, based on the query image issued in the retrieval process, the inter-cluster or intra-cluster retrieval is performance based on the compositions of the super clusters. Table 9 shows the average retrieval accuracy obtained for the same 172 queries listed in Table 6.

Table 9. Accuracy comparison between 'Intra_Cluster' and 'Inter-Cluster'

	Scope			
	10	20	30	40
Intra_Cluster	76%	73%	70%	68%
Inter_Cluster	81%	78%	76%	74%

As shown in Table 9, 'Inter-Cluster' indicates that the search space is not limited to the main cluster during the cluster-based retrieval process. The 'Intra-cluster' results are the same as the ones shown in Table 7 and they are listed here to make the comparison clearer. As we can see from this table, by constructing the super clusters, the cluster-based retrieval process can perform the intra-cluster or inter-cluster retrieval accordingly to greatly improve the retrieval accuracy.

3.3 Efficiency of the Proposed Framework

To evaluate the efficiency, we divide the proposed framework into two parts: off-line processes and on-line retrieval process. All the computationally intensive steps in our framework are conducted in the off-line processes, which include database-level knowledge summarization, clustering process, and cluster-level knowledge summarization. Since the off-line process can be carried out on an annual or semi-annual base, its performance will not have much effect on the on-line retrieval process. Hence, only on-line retrieval efficiency is discussed.

The on-line retrieval of this framework is efficient because of the following reasons. First, with the conceptual image database clustering process, the search space in a distributed database system is reduced dramatically. Second, for a query image *q*, only a single vector related to this query image in $\mathcal{A}$ (at the intra-database, intra-cluster, or super cluster level) needs to be fetched into the memory. In addition, due to the effectiveness of our framework, we can achieve reasonably good retrieval results using a small-size feature set (e.g., 19 global image features in this paper) compared to dozens of features normally used in other CBIR systems. This could further reduce the storage cost and alleviate the search complexity. Moreover, in order to further improve the efficiency, we can use the existing data indexing schemas, such as R-tree and M-tree, to further expedite the searching process.

4. CONCLUSIONS

In this paper, we propose a unified framework to facilitate conceptual database clustering and content-based image retrieval. The proposed framework is built upon the Markov Model Mediators (MMMs) mechanism and addresses both the retrieval efficiency and effectiveness issues that remain to be the problems in the CBIR society. A clustering strategy based on the MMM mechanism is used to partition the image databases into a set of

conceptual image database clusters. The cluster-level knowledge summarization is then conducted to enable the intra-cluster and inter-cluster retrieval and to achieve a better trade-off between the retrieval accuracy and the search cost. Our proposed framework performs cluster-based image retrieval adaptively in either the intra-cluster level or the inter-cluster level. Several experiments were conducted on a large collection of images to demonstrate the effectiveness and efficiency of the proposed framework, and the results of the comparative studies were reported. The experimental results exemplify that our proposed framework achieves better retrieval accuracy via inter-cluster retrieval than that of intra-cluster retrieval.

5. ACKNOWLEDGEMENT

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 HRD-0317692. For Chengcui Zhang, this research was supported in part by SBE-0245090 and the UAB ADVANCE program of the Office for the Advancement of Women in Science and Engineering.

6. REFERENCES

[1] Carson, C., Belongie, S., Greenspan, H., and Malik, J. Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24,8 (2002), 1026–1038.

[2] Chen, Y. and Wang, J. Z. A Region-based Fuzzy Feature Matching Approach to Content-based Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 9 (2002), 1252-1267.

[3] Ciaccia, P., Patella, M., and Zezula, P. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *Proceedings of the 23rd VLDB conference*, 1997, 426-435.

[4] Cooper, M., Foote, J., and Girgensohn, A. Temporal Event Clustering for Digital Photo Collections. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2003, 364-373.

[5] Do, M. N. and Vetterli, M. Rotation Invariant Texture Characterization and Retrieval Using Steerable Wavelet-Domain Hidden Markov Models. *IEEE Transactions on Multimedia*, 4, 4 (Dec. 2002), 517-527.

[6] Flickner, M., et al. Query By Image and Video Content: The QBIC System. *IEEE Computer*, 28, 9 (1995), 23-32.

[7] Gupta, A. and Jain, R. Visual Information Retrieval. *Communications of the ACM*, 40, 5 (1997), 71-79.

[8] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2-3), (December 2001), 107-145.

[9] Jing, F., Li, M., Zhang, H. J., and Zhang, B. An Effective Region-based Image Retrieval Framework. In *Proceedings of the Tenth ACM international conference on Multimedia*, 2002, 456-465.

[10] Kim, D.-H. and Chung, C.-W. QCluster: Relevance Feedback Using Adaptive Clustering for Content-based Image Retrieval. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, 2003, 599-610.

[11] Kim, S. J., Baberjee, J., Kim, W., and Garza, J. F. Clustering a Dag for Cad Databases. *IEEE Transactions on Software Engineering*, 14, 11 (Nov. 1988), 1684–1699.

[12] Kossmann, D. The State of the Art in Distribute Query Processing. *ACM Computing Surveys*, 32, 4 (Dec. 2000), 422-469.

[13] Lu, G. Techniques and Data Structures for Efficient Multimedia Retrieval Based on Similarity. *IEEE Transactions on Multimedia*, 4, 3 (2002), 372-384.

[14] Lu, Y., Zhang, H., Liu, W., and Hu, C. Joint Semantics and Feature Based Image Retrieval Using Relevance Feedback. *IEEE Transactions on Multimedia*, 5, 3 (2003), 339-347.

[15] Mao, J. and Jain, A. K. A Self-organizing Network for Hyperellipsoidal Clustering (hec). *IEEE Transactions on Neural Networks*, 7, 1 (1996), 16-29.

[16] Natsev, A., Rastogi, R., and Shim, K. WALRUS: A Similarity Retrieval Algorithm for Image Databases. *IEEE Transactions on Knowledge and Data Engineering*, 16, 3 (2004), 301-316.

[17] Rui, Y., Huang, T., Ortega, M., and Mehrotra, S. Relevance Feedback: A Power Tool for Interactive Content-based Image Retrieval. *IEEE Transactions on Circuit and Video Technology*, 8, 5 (1998), 644-655.

[18] Safar, M., Shahabi, C. and Sun, X. Image Retrieval by Shape: A Comparative Study. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'00)*, 2000, 141-144.

[19] Sakurai, Y., Yoshikawa, M., Uemura, S., and Kojima, H. The A-tree: An Index Structure for High-Dimensional Spaces Using Relative Approximation. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, 2000, 516-526.

[20] Saux B. L. and Boujemaa N. Image Database Clustering with SVM-based Class Personalization. *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia*, 2004.

[21] Sclaro, S., LaCascia, M., Sethi, S., and Taycher, L. Unifying Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. *Computer Vision and Image Understanding*, 75 (1/2), (1999), 86–98.

[22] Sheikholeslami, G., Chang, W., and Zhang, A. SemQuery: Semantic Clustering and Querying on Heterogeneous Features for Visual Data. *IEEE Transactions on Knowledge and Data Engineering*, 14, 5 (2002), 988-1002.

[23] Shyu, M.-L., Chen, S.-C., Chen, M., and Zhang, C. Affinity Relation Discovery in Image Database Clustering and Content-based Retrieval. Accepted for publication (short paper), *ACM International Conference on Multimedia*, October 10-16, 2004.

[24] Shyu, M.-L., Chen, S.-C., Chen, M., Zhang, C. and Sarinnapakorn, K. Image Database Retrieval Utilizing Affinity Relationships. In *Proceedings of the 1st ACM International Workshop on Multimedia Databases*, 2003, 78-85.

[25] Shyu, M.-L., Chen, S.-C., and Haruechaiyasak, C. Mining User Access Behavior on the WWW. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 2001, 1717–1722.

[26] Shyu, M.-L., Chen, S.-C., Haruechaiyasak, C., Shu, C.-M., and Li, S.-T. Disjoint Web Document Clustering and Management in Electronic Commerce. In *Proceedings of the Seventh International Conference on Distributed Multimedia Systems (DMS'2001)*, 2001, 494-497.

[27] Shyu, M.-L., Chen, S.-C., and Kashyap, R.L. Organizing a Network of Databases Using Probabilistic Reasoning. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, 2000, 1990-1995.

[28] Shyu, M.-L., Chen, S.-C., and Kashyap, R. L. A Probabilistic-Based Mechanism for Video Database Management Systems. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2000)*, 2000, 467-470.

[29] Shyu, M.-L., Chen, S.-C., and Kashyap, R. L. Database Clustering and Data Warehousing. In *Proceedings of the 1998 ICS Workshop on Software Engineering and Database Systems*, 1998, 30-37.

[30] Shyu, M.-L., Chen, S.-C., and Shu, C.-M. Affinity-Based Probabilistic Reasoning and Document Clustering on the WWW. In *Proceedings of the 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC)*, 2000, 149-154.

[31] Shyu, M.-L., Chen, S.-C., and Rubin, S. H. Stochastic Clustering for Organizing Distributed Information Source. Accepted for publication, *IEEE Transactions on Systems, Man and Cybernetics*, Part B, 2004.

[32] Stehling, R. O., Nascimento, M. A., and Falcao, A. X. On Shapes of Colors for Content-based Image Retrieval. In *ACM International Workshop on Multimedia Information Retrieval (ACM MIR'00)*, 2000, 171-174.

[33] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. Computing Iceberg Concept Lattices with Titanic. *Data & Knowledge Engineering*, 42, 2 (2002), 189-222.

[34] Tong, S. and Chang, E. Support Vector Machine Active Learning for Image Retrieval. In *Proceedings of the Ninth ACM International Conference on Multimedia*, 2001, 107-118.

[35] Wu, H., Lu, H., and Ma, S. A Practical SVM-based Algorithm for Ordinal Regression in Image Retrieval. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2003, 612-621.

[36] Yanai, K. Generic Image Classification Using Visual Knowledge on the Web. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, 2003, 167-176.

[37] Zhang, C., Chen, S.-C., and Shyu, M.-L. Multiple Object Retrieval for Image Databases Using Multiple Instance Learning and Relevance Feedback. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'04)*, 2004.

[38] Zhang, D. S. and Lu, G. Generic Fourier Descriptors for Shape-based Image Retrieval. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME'02)*, 1 (2002), 425-428.