Affinity Relation Discovery in 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
In this paper, we propose a unified framework, called *Markov Model Mediator* (MMM), to facilitate image database clustering and to improve the query performance. The structure of the MMM framework consists of two hierarchical levels: local MMMs and integrated MMMs, which model the affinity relations among the images within a single image database and within a set of image databases, respectively, via an effective data mining process. The effectiveness and efficiency of the MMM framework for database clustering and image retrieval are demonstrated over a set of image databases which contain various numbers of images with different dimensions and concept categories.

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, Image Database Clustering, Markov Model Mediators (MMMs).

1. INTRODUCTION
With the proliferation of image data, there is an increasing need to retrieve images efficiently and accurately from image databases. To address such a demand, content-based image retrieval (CBIR) has been actively researched in two major directions, namely retrieval effectiveness and efficiency. For retrieval effectiveness, the global features such as color, texture and shape have been well studied. In addition, in view of the limited power of these low-level features in representing the high-level semantic contents of the images, most of the recent work focuses on bridging such semantic gaps via relevance feedback (RF) [3] and region-based CBIR [2] approaches. Through user feedbacks, RF attempts to refine the query results by estimating the ideal query parameters from the low-level features. Whereas, region-based CBIR performs image retrieval at the object/region level to overcome the deficiencies of global feature matching processes. On the other hand, many indexing techniques and data structures have been proposed with the aim of boosting the retrieval efficiency. For instance, a novel KVA-File (kernel VA-File) that extends VA-File to kernel-based retrieval methods was proposed [1]. Another kind of approaches, which seem to offer better prospects of success, is the use of similarity clustering of images. Clustering techniques allow hierarchical access for retrieval and thus improve the retrieval efficiency [4].

An ideal CBIR system should be both effective and efficient in terms of image retrieval. However, previous studies tend to focus on one aspect of these two requirements without or with little consideration about the other one. Moreover, most of the existing work in data indexing and clustering is conducted at a single database level, which is not sufficient to meet the increasing demand of handling efficient image database retrieval in a distributed environment. In addition, in the traditional database research area, data clustering places related or similar valued records or objects in the same page on disks for performance purposes. However, due to the autonomous nature of each image database, it is not realistic to improve the performance of databases by actually moving around the databases.

In response to these issues, we propose a unified framework called *Markov Model Mediator* (MMM), which is used for both intra-database and inter-database affinity relation discovery with the ultimate goal of improving both the retrieval accuracy and efficiency across multiple image databases. In our previous studies, the MMM mechanism has been applied to content-based image retrieval within a single database [5] and general-purpose database clustering [6]. In this paper, we further extend our previous work by enabling image database clustering and cluster-based image retrieval for efficiency purposes. In particular, we propose the use of MMMs for the construction of probabilistic networks via the affinity mining process, to facilitate the conceptual database clustering and the image retrieval process at both intra-database and inter-database levels. It is a unified framework in the sense that the same mechanism (MMM) is utilized at different hierarchies (local image databases and image database clusters) to build the probabilistic networks which represent the affinity relations among images and databases. The proposed database clustering strategy fully utilizes the information contained in the integrated probabilistic networks,

MM'04, October 10–16, 2004, New York, NY, USA.

and partitions the image databases into a set of conceptual image database clusters without physically moving them. Essentially, since a set of image databases with close relationships are put in the same image database cluster and are required consecutively on some query access path, the number of platter (cluster) switches for data retrieval with respect to the queries can be reduced.

The rest of the paper is organized as follows. Section 2 presents the key components of the MMM-based conceptual clustering framework. Experimental results are presented in Section 3. Finally, Section 4 concludes this paper.

2. MMM-BASED CONCEPTUAL CLUSTERING STRATEGY

The core of the proposed framework is the MMM mechanism that facilitates conceptual database clustering to improve the retrieval accuracy. An MMM-based conceptual clustering strategy consists of two major steps: 1) calculating the similarity measures between every two image databases, and 2) clustering databases using the similarity measures. Here, two image databases are said to be related if they are accessed together frequently or contain similar images. In the first step, a local probabilistic network is built to represent the affinity relationships among all the images within each database, which is modeled by a local MMM and enables accurate image retrieval at the intra-database level. The second step is the proposed conceptual clustering strategy that fully utilizes the parameters of the local MMMs to avoid the extra cost of information summarization, which may be unavoidable in other clustering methods. In our previous work [6], a thorough comparative study has been conducted, in which the MMM mechanism was compared with several clustering algorithms including single-link, complete-link, group-average-link, etc. The experimental results demonstrated that our MMMs produce the best performance in general-purpose database clustering. However, it cannot be directly applied to image database clustering because: 1) image data have special characteristics that are quite different from numerical/textual data; and 2) image database queries are different from traditional database queries in that they may involve users' subjective perceptions in the retrieval process. In this study, we further extend the general MMM-based clustering strategy to handle image database clustering. In Section 3, the effectiveness of the proposed MMM framework for conceptual image database clustering will be examined.

For each image database cluster, an inter-database level probabilistic network, represented by an integrated MMM, is constructed to model a set of autonomous and interconnected image databases in it, which serves to reduce the cost of retrieving images across the image databases and to facilitate accurate image retrieval within the conceptual database cluster.

2.1 Calculating the Similarity Measures

Our conceptual image database clustering strategy is to group related image databases in the same cluster such that the intra-cluster similarity is high and the inter-cluster similarity is low. Thus, a similarity measure needs to be calculated for each pair of image databases in the distributed database system. These similarity measures indicate the relationships among the image databases and are used to partition the databases into clusters.

Let d_i and d_j be two image databases, and $X=\{x_1,\ldots,x_{k1}\}$ and $Y=\{y_1,\ldots,y_{k2}\}$ be the set of images in d_i and d_j, where $k1$ and $k2$ are the numbers of the images in X and Y, respectively. Let $N_k=k1+k2$ and $O^k=\{o_1,\ldots,o_{Nk}\}$ be an observation set with the features belonging to d_i and d_j and generated by query q_k, where the features $o_{1,\ldots,}\ o_{k1}$ belong to d_i and $o_{k1+1,\ldots,}\ o_{Nk}$ belong to d_j. Assume that the observation set O^k is conditionally independent given X and Y, and the sets $X\subseteq d_i$ and $Y\subseteq d_j$ are conditionally independent given d_i and d_j. The similarity measure $S(d_i, d_j)$ is defined as follows:

$$S(d_i,d_j)=\left(\sum_{O^k\in OS} P\left(O^k\middle|X,Y;d_i,d_j\right)P\left(X,Y;d_i,d_j\right)\right)F(N_k) \quad (1)$$

where $P(X,Y;d_i,d_j)$ is the joint probability of $X\subseteq d_i$ and $Y\subseteq d_j$, and $P(O^k|X,Y;d_i,d_j)$ is the probability of occurrence of O^k given X in d_i and Y in d_j. They are defined as follows:

$$P\left(O^k\middle|X,Y;d_i,d_j\right)=\prod_{u=1}^{k1}P\left(o_u|x_u\right)\prod_{v=k1+1}^{N_k}P\left(o_v|y_{v-k1}\right) \quad (2)$$

$$P\left(X,Y;d_i,d_j\right)=\prod_{u=2}^{k1}P\left(x_u|x_{u-1}\right)P(x_1)\prod_{v=k1+2}^{N_k}P\left(y_{v-k1}|y_{v-k1-1}\right)P(y_1) \quad (3)$$

In Equation (2), $P(o_u|x_u)$ (or $P(o_v|y_{v-k1})$) represents the probability of observing a feature o_u (or o_v) from an image x_u (or y_{v-k1}). In Equation (3), $P(x_u|x_{u-1})$ (or $P(y_{v-k1}|y_{v-k1-1})$) indicates the probability of retrieving an image x_u (or y_{v-k1}) given the current query image as x_{u-1} (or y_{v-k1-1}), while $P(x_1)$ (or $P(y_1)$) is the initial probability. In order to obtain these probabilities, each image database is modeled by a local MMM. Another level of MMMs (called integrated MMMs) is also constructed in our framework, which is used to model the conceptual image database cluster (discussed in Section 2.3.1). Following gives the formal definition of an MMM.

Definition 1: An MMM is a 5-tuple $\lambda = (\mathcal{S}, \mathcal{F}, \mathcal{A}, \mathcal{B}, \Pi)$, where $\mathcal{S}$ is a set of images called states; $\mathcal{F}$ is a set of distinct features of the images; $\mathcal{A}$ denotes the state transition probability distribution, where each entry (i, j) indicates the affinity relation between images i and j; $\mathcal{B}$ is the observation symbol probability distribution; and Π is the initial state probability distribution.

Here, $\mathcal{S}$ consists of all the images in an image database (or an image database cluster) and $\mathcal{F}$ includes all the distinct features of the images in $\mathcal{S}$. $\mathcal{A}$ represents the affinity relations among all the images in a database (or an image database cluster) based on the query usage patterns. The relationship of the images are modeled by the sequences of the MMM states connected by transitions. $\mathcal{B}$ consists of the normalized image feature vectors for all the images. Π indicates the probability that an image can be the query image for the incoming queries. According to this definition, $P(x_u|x_{u-1})$, $P(o_u|x_u)$, and $P(x_u)$ (or $P(y_{v-k1}|y_{v-k1-1})$, $P(o_v|y_{v-k1})$ and $P(y_{v-k1})$) correspond to the entries in $\mathcal{A}$, $\mathcal{B}$, and Π for d_i (or d_j), respectively.

2.1.1 Construction of the Local MMMs

For local MMMs, $\mathcal{S}$ contains all the images in an image database, and $\mathcal{A}$, $\mathcal{B}$, and Π are the major parameters for the image database.

2.1.1.1 Relative Affinity Measures

The relative affinity measures are used to indicate how frequently two images are accessed together. Intuitively, the more frequently two images are accessed, the more closely they are related. In order to discover these relative affinity relations, a set of training data is used in the data mining process, which is actually a set of log files consisting of the usage patterns of the queries versus the images in $\mathcal{S}$, and the access frequencies of the queries [5].

Let $Q = \{q_1, q_2, \ldots, q_q\}$ be the set of queries that ran on image databases $d_1, d_2, \ldots, d_d$, $pattern_{m,k}$ be the usage pattern of image m with respect to query q_k and $freq_k$ be the access frequency of q_k during a time period. Here, $pattern_{m,k}$ =1 if image m is accessed by q_k; otherwise it is set to 0. The relative affinity measure between images m and n can be obtained as follows:

$$aff_{m,n} = \sum_{k=1}^{q} pattern_{m,k} \times pattern_{n,k} \times freq_k \quad (4)$$

2.1.1.2 State Transition Probability Distribution

Two images are said to have a higher relative affinity relationship if they are accessed together more frequently. Accordingly, for the state transition probability in the probabilistic network, if two images, m and n, have a higher relative affinity relationship, the chance that a traversal choice to node n given the current node is in m (or vice versa) should be higher. The conditional probability $a_{m,n}$ is the $(m, n)^{th}$ element of $\mathcal{A}$, which indicates the probabilities that go from state m to state n. For the local MMM, we define:

Definition 2: For images $m, n \in d_i$

- $f_{m,n} = aff_{m,n} / \Sigma_m \Sigma_n aff_{m,n}$ = the joint probability;
- $f_m = \Sigma_n f_{m,n}$ = the marginal probability;
- $a_{m,n} = f_{m,n} / f_m$ = the conditional probability which refers to the state transition probability for the local MMM.

2.1.1.3 Observation Symbol Probability Distribution

The observation symbol probability denotes the probability of observing an output symbol (feature) from a state (image). In this study, we use 13 color features and 6 texture features. The HSV color space is used to obtain the color features, including black, gray, white, red, red-yellow, yellow, yellow-green, green, green-blue, blue, blue-purple, purple, and purple-red. For the texture features, one-level wavelet transformation using Daubechies wavelets is used to generate the horizontal detail sub-image, the vertical detail sub-image, and the diagonal detail sub-image. For the wavelet coefficients in each of the above three subbands, the mean and variance values are captured. Let p=19 be the total number of distinct features. $\mathcal{B} = \{b_{i,j}\}$ $(1 \le i \le N, 1 \le j \le p)$ contains all the feature vectors of the images in $\mathcal{S}$, where N is the number of images in $\mathcal{S}$. The features are normalized per row.

2.1.1.4 Initial State Probability Distribution

Based on the information collected from the training data set, the initial states for queries can be obtained as follows:

$$\Pi = \{\pi_m\} = \sum_{k=1}^{q} pattern_{m,k} \Big/ \sum_{l=1}^{N} \sum_{k=1}^{q} pattern_{l,k} \quad (5)$$

Here, N is the number of images in an image database d_i. Once the local MMM is constructed for each image database, the similarity values can be computed for each pair of image databases. Then a probabilistic network is built with each image database represented as a node in it. For nodes d_i and d_j in this probabilistic network, the branch probability $P_{i,j}$ is transformed from the similarity value $S(d_i, d_j)$. Here, the transformation is performed by normalizing the similarity values per row to indicate the branch probabilities from a specific node to all its accessible nodes.

2.2 Clustering Image Databases

Based on the probability distributions for the local MMMs and the probabilities $P_{i,j}$ for the probabilistic network, 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, or d_i) in the long run.

$$\Sigma_i \phi_i = 1 \qquad \phi_j = \Sigma_i \phi_i P_{i,j} \qquad j = 1, 2, \ldots \quad (6)$$

The conceptual image database clustering strategy is traversal based and greedy. Conceptual image database clusters are created according to the order of the stationary probabilities of the image databases. The image database that has the largest stationary probability is selected to start a new image database cluster. While there is room in the current cluster, all image databases accessible in the probabilistic network from the current member image databases of the cluster are considered. The image database with the next largest stationary probability is selected and the process continues until the cluster fills up. At this point, the next un-partitioned image database from the sorted list starts a new image database cluster, and the whole process is repeated until no un-partitioned image databases remain. The time complexity for this conceptual database clustering strategy is $O(p\log p)$ while the cost of calculating the similarity matrix is $O(p^2)$, where p is the number of image databases. The size of each image database cluster is predefined and is the same for all image database clusters.

2.3 Retrieval Algorithms

Once the image database clusters are obtained via the proposed conceptual database clustering strategy, for each cluster, an integrated MMM is constructed to model a set of autonomous and interconnected image databases within it, which serves to reduce the cost of retrieving images across image databases and to facilitate accurate image retrieval. The cluster-based image retrieval is then supported by using the integrated MMM. In the following two subsections, the construction of the integrated MMMs for the database clusters will first be introduced, followed by the proposed cluster-based image retrieval algorithms.

2.3.1 Construction of the Integrated MMMs

For any images s and t in a conceptual image database cluster CC, the formulas to calculate $\mathcal{A}$ are defined in **Definition 3**. Here, we assume CC contains two or more image databases; otherwise, $\mathcal{A}$ is the same as the one for a single image database.

Definition 3: Let λ_i and λ_j denote two local MMMs for image databases d_i and d_j, where $j \ne i$ and $\lambda_i, \lambda_j \in CC$.

- $f_{s,t}$ are defined similarly as in **Definition 2**, except that they are calculated in CC instead of a single image database;
- $p_{s,t} = f_{s,t} / \Sigma_{n \in CC} f_{s,n}$ = the probability that λ_i goes to λ_j with respect to s and t;
- $p_s = 1 - \Sigma_{t \notin \lambda_i} p_{s,t}$ = the probability that λ_i stays with respect to s;
- $a_{s,t}$= the conditional probability of a local MMM;
- $a'_{s,t}$=the state transition probability of an integrated MMM, where if $s, t \in \lambda_i \Rightarrow a'_{s,t} = p_s a_{s,t}$, and if $s \in \lambda_i \wedge t \notin \lambda_i \Rightarrow a'_{s,t} = p_{s,t}$;

$\mathcal{A}$ is obtained by repeating the above steps for all local MMMs in CC. As for $\mathcal{B}$ and Π in the integrated MMM, the construction methods are similar to those presented in Sections 2.1.1.3 and 2.1.1.4, except that the image scope is defined in the cluster CC.

2.3.2 Image Retrieval over Image Database Clusters

Once the integrated MMMs are obtained, content-based retrieval can be conducted at the image database cluster level as follows. $W_t(i)$ is defined as the edge weight from image i to query image q at the evaluation of the t^{th} non-zero feature (o_t) in the query, where $1 \le i \le |CC|$, $1 \le t \le T$, and T is the total number of features within a conceptual image database cluster (or a single database). The retrieval algorithm is given as follows:

At $t = 1$, $W_1(q,i) = a_{q,i}(1-|b(i,o_1)-b(q,o_1)|/b(q,o_1))$ (7)

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

$$W_{t+1}(q,i) = W_t(q,i)(1-|b(i,o_{t+1})-b(q,o_{t+1})|/b(q,o_{t+1})) \quad (8)$$

Then the similarity function is defined as:

$$SS(q,i) = \sum_{t=1}^{T} W_t(q,i) \quad (9)$$

Here, $a_{q,i}$ ($a_{q,i} \in \mathcal{A}$) is the relative affinity relation indicating how closely query image q is related to image i, and $b(i,o_k)$ ($b(i,o_k) \in \mathcal{B}$) is the value of the k^{th} feature in image i. $SS(q, i)$ represents the similarity score between images q and i, where a larger score suggests higher similarity. Note that the same retrieval algorithms can be applied to image retrieval at both local database and database cluster levels by using local or integrated MMMs. We have demonstrated its effectiveness in image retrieval at the local database level in [5]. In this study, the effectiveness of the proposed framework in conceptual image database clustering and inter-database level image retrieval is examined.

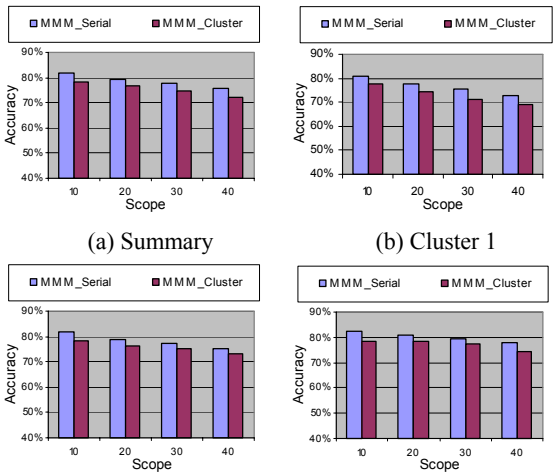
3. EXPERIMENTAL RESULTS

To show the effectiveness of image retrieval in conceptual image database clusters, 12 image databases with totally 18,700 images (the number of images in each image database ranges from 1,350 to 2,250) are used. The affinity-based data mining process is conducted utilizing the training data set, which contains the query trace generated by 1,400 queries issued to the image databases. The proposed conceptual image database clustering strategy is employed to partition these 12 image databases into a set of image database clusters. Here, the size of the conceptual image database cluster is set to 4, which represents the maximal number of member image databases a cluster can have. As a result, 3 clusters are generated with 6,450, 5,900 and 6,350 images, respectively. The performance is tested by issuing 160 test queries to these 3 clusters (51, 54 and 55 queries, respectively). For comparison, an image database (namely *DB_whole*) with all the 18,700 images is constructed and tested by the same set of the queries.

Figure 1 shows the comparison results, where the scope specifies the number of images returned and the accuracy at a scope s is defined as the ratio of the number of the relevant images within the top s images. In this Figure, 'MMM_Cluster' represents the retrieval accuracy achieved by issuing queries to each of the database clusters, while 'MMM_Serial' denotes the results of carrying out the search throughout the *DB_whole* image database. For instance, 'MMM_Cluster' and 'MMM_Serial' in Figure 1(b) represent the results obtained by issuing 51 queries to cluster 1 and *DB_whole*, respectively. As shown in this figure, the accuracy of 'MMM_Cluster' is slightly worse than 'MMM_Serial', which is reasonable because 'MMM_Cluster' carries out the search in a subspace of *DB_whole*. Considering that the search space is reduced at about one third of the whole space and the image retrieval is conducted at the inter-database level, the effectiveness of our framework in both conceptual image database clustering and content-based image retrieval is obvious. In other words, by using the conceptual image database clusters, the query cost can be reduced dramatically (almost 1/3) without significant decreases in accuracy (averagely 3%).

4. CONCLUSIONS

In this paper, we propose the use of a mathematically sound framework, Markov Model Mediators (MMMs), to facilitate both the conceptual image database clustering and the cluster-based content-based image retrieval. The proposed framework takes into consideration both the efficiency and effectiveness requirements in content-based image retrieval. More specifically, an effective database clustering strategy is employed in our framework to partition the image databases into a set of conceptual image database clusters, which reduces the query cost dramatically without decreasing the accuracy significantly. In addition, the affinity relations among the images in the databases are explored through the data mining process, which capture the users' concepts in the retrieval process and significantly improve the retrieval accuracy.



(a) Summary (b) Cluster 1

(c) Cluster 2 (d) Cluster 3

Figure 1. Accuracy comparison

5. ACKNOWLEDGMENTS

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] Heisterkamp, R. and Peng, J. Kernel VA-Files for Relevance Feedback Retrieval. In *Proc. of the First ACM International Workshop on Multimedia Databases*, 2003, 48-54.

[2] Natsev, A., et al. WALRUS: A Similarity Retrieval Algorithm for Image Databases. *IEEE Trans. on Knowledge and Data Engineering*, 16, 3 (2004), 301-316.

[3] Rui, Y., et al. Relevance Feedback: A Power Tool for Interactive Content-based Image Retrieval. *IEEE Trans. on Circuit and Video Technology*, 8, 5 (1998), 644-655.

[4] Saux B. L. et al. Image Database Clustering with SVM based Class Personalization. In *IS&T/SPIE Conf. on Storage and Retrieval Methods & Applications for Multimedia*, (2004).

[5] Shyu, M.-L., et al. Image Database Retrieval Utilizing Affinity Relationships. In *Proc. of the 1st ACM International Workshop on Multimedia Database*, (2003), 78-85.

[6] Shyu, M.-L., et al. Stochastic Clustering for Organizing Distributed Information Source. Accepted for publication, *IEEE Trans. on Systems, Man and Cybernetics*, Part B, (2004).