INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG97/2290
July 1997

Source:
Status: **Proposal**
Title: **MARS and Its Applications to MPEG-7**
Author: **Yong Rui, Thomas S. Huang, and Sharad Mehrotra**
Beckman Institute and Department of Computer Science
University of Illinois at Urbana-Champaign

Abstract: To address the emerging needs of access to and retrieval of multimedia objects in many applications, we have started a *Multimedia Analysis and Retrieval Systems* project at the University of Illinois. This project addresses three main aspects in Multimedia Information Retrieval, i.e. feature extraction, multimedia object description, and retrieval algorithm. Although MPEG-7 will concentrate only on multimedia object description, such a goal will be better accomplished if its interfaces to feature extraction and retrieval algorithm are appropriate defined. In this proposal, we will first give a brief overview of the MARS system. Then we propose a multimedia object model for MPEG-7's content description interface. The proposed model allows information abstraction at various semantic levels. To better model human perception subjectivity to multimedia data, relevance feedback is integrated into the retrieval process. Our experimental results show that the proposed multimedia object model and retrieval model are general enough for modeling and specific enough to adapt to user's information need.

Keywords: multimedia information retrieval systems, content-based retrieval, multimedia content description interface, multimedia object model

1. INTRODUCTION

Advances in high performance computing, communication, and storage technologies as well as emerging large scale multimedia applications have made Multimedia Information Retrieval (MIR) systems one of the most challenging and important research directions. Such systems will support multimedia data as ``first-class" objects that are capable of being stored and retrieved based on their rich internal contents. Applications of such systems include among others:

- Government and commercial uses of remote sensing images, satellite images, air photos, etc;
- Digital libraries, including digital catalogs, product brochures, training and education, broadcast and entertainment, etc;
- Medical databases, such as X-rays, MRI, etc;
- Special-purpose databases, e.g. face/fingerprint databases for security, business directories, maps, etc.

While current technology allows generation, scanning, transmission, and storage of large numbers of digital images, video and audio, existing practice of indexing, access and retrieval of visual data is still in its infancy. A successful MIR system requires the breakthroughs in the following three aspects:

- Reliable feature extraction.
- Generic multimedia object model (description).
- Effective retrieval algorithms.

The relationship of the three aspects is illustrated in Figure 1.

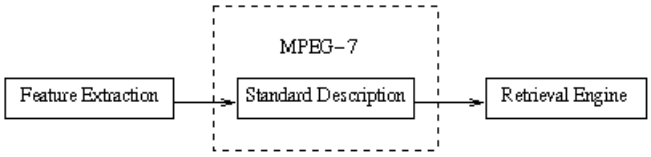

Figure 1. Relationship of the three aspects

Currently, most of the research effort has been focused on feature extraction. Much less effort has been given in retrieval algorithm and even less attention has been given to multimedia object model. To solve this problem, MPEG has started a new work item called MPEG-7, i.e. Multimedia Content Description Interface. Its main focus is the multimedia object model and its interfaces to feature extraction and retrieval algorithm.

While the feature extraction itself is a very important aspect of MIR, it alone can not lead to a successful MIR system. In the past few years, many feature extraction techniques have been proposed in various low-level features such as color, texture, shape, structure, composition, human faces, etc. However, most of them are only suitable in a specific setting or for a particular data set. For a different data set, they may be less effective or even become meaningless. For example, an effective shape feature extraction technique would become meaningless if is applied to a texture image data set.

Therefore, while keep advancing the techniques of feature extraction, we need to develop a multimedia object model such that the low-level features are not only stored in the model, but also they will be invoked at the right time and right place to facilitate the retrieval.

To develop such a model, *human perception subjectivity* needs to be taken into account. That is, for the same multimedia object, different people may perceive it differently. An approach to overcoming this human perception problem is to integrate the relevance feedback technique developed in the traditional text-based Information Retrieval (TIR) into the MIR systems. Relevance feedback is the process of automatically adjusting an existing query using the information fed-back by the user about the relevance of previously retrieved objects. By incorporating relevance feedback into the retrieval process, human perception subjectivity can be better modeled, thus resulting in considerable improvement in retrieval performance [1,2,3].

In MIR, the issue of *human perception subjectivity* is even more important than that in TIR because of the rich multimedia content contained in the multimedia objects. Development of techniques that can incorporate human perception subjectivity into MIR is thus crucial for a successful MIR system. The information abstraction of a multimedia object occurs at various levels, since a multimedia object has

multiple features, each feature has multiple representations, and each representation consists of multiple components. This proposal introduces an integrated relevance feedback architecture in MIR [4,5,6,7], where the relevance feedback is applied at all levels *simultaneously*. The user's information need is distributed by different weights among different features, different feature representations, and different representation components. During the retrieval process, the weights are *dynamically* updated based on the user's relevance feedback. With this integrated relevance feedback architecture, the MIR system can better model various level's of information abstraction; thus better supporting user's information need.

The rest of the proposal is organized as follows. In Section 2, a brief overview of MARS is given. In Section 3, a multimedia object model is proposed for MPEG-7's Multimedia Content Description. The interfaces between the content description and feature extraction and retrieval algorithm are discussed in Sections 4 and 5 respectively. Experimental results and conclusions are in Sections 6 and 7.

2. AN OVERVIEW OF MARS

MARS [8,9,10,11,12,13,4,5,6,7,14] is a content-based multimedia information retrieval system developed at University of Illinois at Urbana-Champaign. Currently it supports retrieval of images. We are augmenting the system to support retrieval of video and audio. The major components of MARS are shown in Figure~2 and are discussed below [8,9].

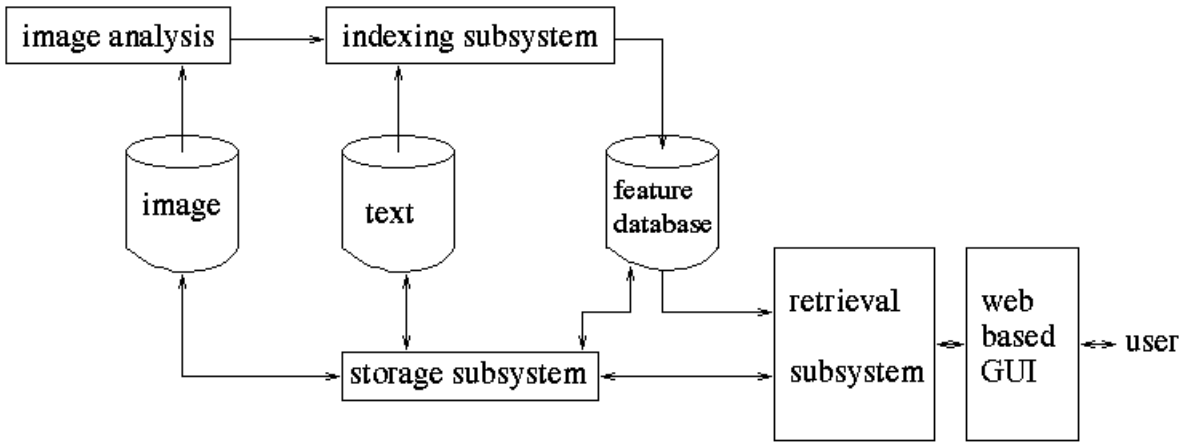

Figure 2. The system architecture

- **User interface**: written using Java applets and accessible over the internet. The user interface allows users to graphically pose content-based queries as well as traditional text-based queries. The URL of MARS is at http://jadzia.ifp.uiuc.edu:8000.

- **Content Indexer (Indexing Subsystem):** The content indexer takes as input an image as well as its text annotation. With the help of the image analyzer it extracts both low-level multimedia features (e.g. color, texture, shape), and salient textual descriptions (e.g. the author's name, category or subject of the image, etc.) and stores these contents into the feature database.

- **Image Analyzer:** The image analyzer extracts salient low-level image features, such as color, texture, shape, and layout.

- **Feature Database:** An image in the feature database is represented using its low-level features as well as high-level textual descriptions.

- **Query Processor (Retrieval Subsystem):** The query processor is written on top of POSTGRES in C. It takes the query specified at the user interface, evaluates the query using the feature database, and returns to the user images that are best matches to the input query. The query language supported allows users to pose complex queries that are composed using low-level image features as well as textual descriptions.

In MARS project, we are addressing many challenging research issues involved in MIR, including automatic feature extraction, compression techniques suitable for browsing and retrieval, indexing and content-based retrieval, efficient query processing, support for information abstraction at various semantic levels, and techniques for seamless integration of the multimedia databases into the organizations' information infrastructure.

3. MULTIMEDIA CONTENT DESCRIPTION INTERFACE

In this section, we will address two main issues related to MPEG-7, i.e. the multimedia object model and the indexing techniques.

3.1. The Multimedia Object Model

As mentioned in Section 1, although many useful features have been identified and feature extraction techniques developed, how they should be organized is still an open problem. There is an urgent need of a multimedia object model to organize all the extracted features in such a way that the appropriate features will be invoked at the right place and right time to answer user's information need. Since user's information need may be at different semantic levels, such a model should also support information abstraction at various semantic levels. To meet the above requirements, we propose the following multimedia object model *O* [7]:

$$O = O(D, T, F, R) \tag{1}$$

- *D* is the raw data of the object, e.g. a JPEG image, or an MPEG video, etc.
  - T is the textual description of the multimedia object.
  - fixed descriptors like title, author, year, etc. that are associated with the object.
- free-text description of the object.

- $F = \{f_i\}$ is the set of low-level multimedia features associated with the object, e.g. color, texture, and shape for images; motion parameters for video;

- $R = \{r_{ij}\}$ is the set of representations for a given feature f_i, e.g. both color histogram and color moments are representations for color feature [15]. Note that, each representation r_{ij} itself is a vector consisting of multiple components, i.e.

$$r_{ij} = [r_{ij1}, ..., r_{ijk}, ..., r_{ijN}] \tag{2}$$

where N is the length of the vector.

The proposed model supports both multiple features and multiple representations to accommodate the rich content in the multimedia objects. Different weights, $W_{q,f}$, $W_{f,r}$, and $W_{r,r}$, are associated with features f_i, representations r_{ij}, and components r_{ijk} respectively, to precisely capture the user's perception subjectivity (for simplicity, we drop the i, j, k indices for the weights). Relevance feedback is used to find the appropriate values for the weights, as will be discussed in Section 5.

This model also supports information abstractions at various semantic levels. The highest level is *T* level, where the textual description is annotated, either purely by human or with the aid from the annotation system [16,17]. The lowest level is the representation (*R*) level. Since the normal user does not have the knowledge of the characteristics of the representations, the information abstraction at this level is transparent to the user. However, the user is still able to access this level's information by using relevance feedback, as will be discussed in Section 5. The middle level is the feature (*F*) level, where all the low-level features are stored. The information abstraction at this level is in between the other two levels. That is, the user can directly access this level's information, but he can access it more effectively with the system's help, as will also be discussed in Section 5.

The proposed multimedia object model provides a structure of organizing different multimedia contents. In Section 5, base on this multimedia object model, we will describe how appropriate contents will be invoked at the right place and right time by using relevance feedback.

Although the proposed object model is aimed at images, it is readily extensible to other audiovisual objects. For example, if we incorporate temporal features into *F*, this model can then support video objects. Of course, a simple extension like the above will not best capture the characteristics of video objects. We are now investigating the techniques of explicitly incorporating temporal features, such as motion parameters and temporal structures, into this model.

3.2 Efficient Feature Indexing

In addition to the multimedia object model, to support effective and efficient retrieval of multimedia data, the indexing techniques need to be explored.

The feature space in MIR normally is very high dimensional and, therefore, usage of conventional multidimensional and spatial indexing methods (e.g., R-trees, quad trees, grid files) is not feasible. Existing multidimensional index method are only useful when the number of dimensions are reasonably small. For example, the R-Tree based methods, which are among the most robust multidimensional indexing mechanisms, work well only for multidimensional spaces with dimensionality around 20. Other methods do not even scale to 20 dimensions.

An approach used by the QBIC to overcome the dimensionality curse of the feature space is to transform the high dimensional feature space to a lower dimensional space using, for example, a K-L transform [18,19]. An R* tree is then used for indexing and retrieval in a lower dimensional space. The retrieval over the index provides a superset of the answers which can then be further refined in the higher dimensional space. While the approach is attractive and the QBIC authors report good retrieval efficiency over small image databases, it is not clear whether it will scale to large databases and complex feature spaces that are very highly multidimensional. In such situations, the large number of false hits in the lower dimensional space might make the approach unusable.

We have developed a dynamic hierarchical clustering technique which is scalable to high dimensionality required by MIR system [14]. The problem of clustering consists of partitioning *N* points in a metric space *M* into *k* clusters based on some criterion. By storing similar objects together, retrieval can be processed more efficiently by reducing the number of objects to be accessed to return the results.

The clustering algorithms used for information retrieval can be classified into two categories, i.e. *static clustering* and *dynamic clustering*. In static clustering, all the objects are first clustered (indexed) before any search can be performed. Whenever a new object is presented, the indexing structure needs a total reorganization to support later searches. In dynamic clustering, search can be *interlaced* with indexing. When a new object is presented, the indexing structure will grow dynamically, no periodic reorganization is needed. Due to the high frequency of data update in MIR systems [20], static clustering is unacceptable.

We have developed a dynamic hierarchical clustering technique in MARS. We developed a graph-theorem based merging strategy and an *ideal* cluster centroid based retrieval algorithm to adapt the clustering technique into multimedia database applications. We further studied the *relationship* between the indexing technique and retrieval algorithm, an area seldom addressed in the literature. We proposed a retrieval algorithm which is parameterized so as to provide the user with the ability to control the tradeoff between the retrieval quality and speed. We integrated the retrieval algorithm with the clustering technique to achieve optimal retrieval performance. Extensive experiments over large scale high dimensional datasets demonstrate the high retrieval performance of our approach [14].

3.3 The Relation to MPEG-7

Based on the discussions in the above two subsections, we propose the following framework for MPEG-7's multimedia content description interface.

For each of the multimedia object, its low-level features are extracted possibly with multiple representations to support human perception subjectivity. The level-level features, together with their representations, are stored within the multimedia object model. This process is done automatically or semi-automatically. To support information abstraction at various levels, high-level textual descriptions are also extracted, by either human or human-aided algorithms, and stored in the multimedia object model.

To support efficient retrieval, both the high-level and low-level features are indexed by using dynamic hierarchical clustering, as described in Section 3.2.

The final information *O(D,T,F,R)*, i.e. the part defined by MPEG-7, is transmitted. How this MPEG-7 specified information can support efficient and powerful retrieval will be discussed in Section 5. In the next section, we will briefly summarize the input to MPEG-7, i.e. feature extraction.

4. FEATURE EXTRACTION

Based on the proposed multimedia object model, in this section we will briefly summarize what are the useful information at each level.

4.1. D

As we mentioned before, *D* is the raw data of the multimedia object. It can either be compressed or uncompressed. It can be of any format as long as it can be perceived by human.

4.2 T

T is a set of high-level textual descriptions associated with the multimedia object. It may contain the following information:

- Fixed description
  - Format: The coding scheme used. This information helps determining which ``viewer'' should be invoked for the user to perceive the multimedia content.
  - Conditions for accessing the material: This could include copyright information, price, etc.
  - Links to other relevant material
  - Date of production
  - Category or subject: For video, for example, they can be classified as comedy, action, fiction, etc.

- Free-text description
  Any textual description that can help retrieving the multimedia object.

Generally, fixed description has to be entered by human. For free-text description, it can either be entered by human or by some automatic texture annotation tools. For example, if an image is downloaded from the web, most likely the text surrounding the image is a related description of the image and can be automatically annotated to the image.

At *T* level, the information abstraction level is high. Normally, events, things, objects can be identified and later be retrieved at this level.

4.3. F

Because of the rich content in the multimedia object, various low-level features are extracted to support potential queries. Some of the useful low-level features are color, texture, shape, appearance, layout, etc.

At *F* level, the information abstraction level is a lower level compared with that in *T* level. In everyday life, human exchange information at high level, such as a beach, forest, yellow flowers, a sunset, buildings, etc. However, in order to query at the *F* level, the user has to map his high-level query to low-level features. For example, the high-level concept *sunset* can be mapped to low-level color layout feature. Subsequent retrieval will be based on the low-level feature. This high-level concept to low-level feature mapping sometime is easy and obvious, but sometimes is not. For the not-so-easy mappings, relevance feedback is used to facilitate the retrieval process, as will be discussed in Section 5.

4.4. R

To accommodate human perception subjectivity, the proposed multimedia model supports multiple representations for a given feature. At *R* level, the information abstraction level is the lowest. Human can directly access the information at this level, only if he knows the characteristics of all the representations. For normal user, this requirement is not true. Therefore, this level's information is transparent to the user. Instead of directly interacting with the information, the user accesses the information at this level implicitly by relevance feedback. This is a very effective way of accessing information, as will be discussed in Section 5.

In the remaining of this subsection, we will briefly describe some most popular representations for each low-level feature. For detailed characteristics of the representations, please refer to the references.

4.4.1. **Color**

Color is one of the most important low-level multimedia features in MIR. While color features could be represented in many color spaces, HSV color space has the best trade-off between closeness to human perception and low computation cost. Some of the most popular color representations are:

- Color Histogram [21,15]
- Cumulative Histogram [15]
- Color Moments [15]
- Color Correlogram [22]

4.4.2. **Texture**

Texture is another important feature of images. Texture feature representations fall into two main categories, Statistics-based and Transform-based.

- CCD (Contrast-Coarseness-Directionality) [18,19]
- Markov Random Filed Model [23]
- Co-occurrence Matrix [23]
- Wold Decomposition [24]
- Shift-invariant Eigenvector [17]
- Gabor Filter [25]
- DCT Transform [26]
- Wavelet Transform [27]

The first five representations are Statistics-based while the last three are Transform-based representations.

4.4.3. **Shape**

Although shape is a very important feature that human can easily extract from an image, reliable automatic extraction and representation of shapes is a challenging open problem in computer vision. The following are some most popular shape feature representations:

- Geometry Features [28]
  The perimeter, area, number of holes, eccentricity, symmetry, etc. of the shape.
- Moment-Invariants [29]
- Turning Angle [30]
- Chamfer [31,32,4]
- Fourier Descriptor [33,13]
- Wavelet Descriptor [34]

4.4.4. **Appearance**

Appearance is the feature describing the appearance of an object, such as human face, fingerprint, etc. It differs from texture feature in that it has multiple texture regions. It differs from shape feature in that it concerns not only the outer boundary of the object but also the object's interior characteristics. One representation of appearance is the eigenimage representation [24].

4.4.5. **Layout**

While the color feature is useful for queries on the relative amount of each color in an image; it is not useful for queries on the spatial location of colors. For example, it is not possible to retrieve all images that contain a red region above and to the right of a large blue region based solely on the color feature. Such queries can be answered correctly only if an image can be segmented into different color regions. The salient color regions are then indexed into the database to support later retrieval [12,35].

4.5. Summary

As we can see from the above descriptions that many high-level and low-level features have been explored by various researchers. For each feature, there exist multiple representations, which model the human perception subjectivity from different angles.

What features and representations should be included in the model is application dependent. This is why we propose the multimedia object model without specifying its fixed features and representations. Instead, features and representations are included based on the characteristics of the application. Once features and representations are determined, the information embedded in them can be accessed by the retrieval model, as will be described in the following section.

5. THE RETRIEVAL MODEL

Effective and efficient retrieval of multimedia data is the final goal of MARS and that of MPEG-7. Equipped with the feature extraction techniques and the proposed multimedia object model for MPEG-7, we will describe the retrieval model in this section.

5.1. Relevance Feedback Enhanced Retrieval

A multimedia object model *O(D,T,F,R)*, together with a set of similarity measures $M=\{m_{ij}\}$, specifies a MIR model *(D,T,F,R,M)*. The similarity measures are used to determine how similar or dissimilar two objects are. Different similarity measures are used for different feature representations. For example, Euclidean is used for comparing vector-based representations while Histogram Intersection is used for comparing color histogram representations.

An important consideration in the design of MIR system is its integration with the organization's existing databases. This requires integration of the query language developed for the multimedia database (which allows content-based and similarity retrieval) with SQL (a popular database query language). Such an integration will allow users to develop complex queries based on both the high-level textual description and low-level content. To allow this high-level and low-level integration, *T* is considered a subset of *F*. This is valid because both of them are features. The only difference is that the former is high-level text-based feature while the latter is low-level content-based feature. After merge *T* to *F*, we have a high-level low-level combined retrieval model *(D,F,R,M)*. Based on the above retrieval model, the retrieval process is described below and also illustrated in Figure 3.

1. The user's information need, represented by the query object *Q*, is distributed among different features f_i, according to the weights $W_{q,f}$.

2. Within each feature f_i, the information need is further distributed among different feature representations r_{ij}, according to the weights $W_{f,r}$.

3. The objects' similarity to the query, in terms of r_{ij}, is calculated according to the corresponding similarity measure m_{ij} and the weights $W_{r,r}$:

$$S(r_{ij}) = m_{ij}(r_{ij}, W_{r,r}) \tag{3}$$

4. Each representation's similarity values are then combined into a feature's similarity value:

$$S(f_i) = \sum_j W_{f,r} S(r_{ij}) \tag{4}$$

5. The overall similarity S is obtained by combining individual $S(f_i)$'s:

6. The objects in the database are ordered by their overall similarity to Q. The N_{RT} most similar ones are

$$S = \sum_i W_{q,f} S(f_i) \tag{5}$$

returned to the user, where N_{RT} is the number of objects the user wants to retrieve.

7. For each of the retrieved objects, the user marks it as *relevant*, *non-relevant*, or *no-opinion*, according to his information need and perception subjectivity.

8. The system updates the weights according to the user's feedback and goes to Step 1.

In Figure 3, the information need embedded in Q flows up while the multimedia content of O's flows down. They meet at the dashed line, where the similarity measures m_{ij} are applied to calculate the similarity values between Q and O's.

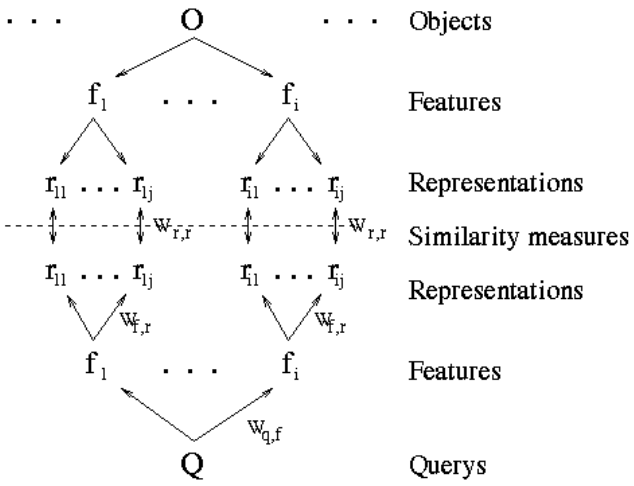

Figure 3. The retrieval process

During the retrieval process, the weights associated with the query can be *dynamically* updated via relevance feedback to best reflect the user's information need and perception subjectivity. This enables the retrieval system to invoke appropriate features and representations at the right time and right place.

Let RT be the set of the most similar N_{RT} objects, according to the overall similarity value S:

$$RT = [RT_1, ..., RT_l, ..., RT_{N_{RT}}] \tag{6}$$

Let *Score* be the set containing the relevance scores fed-back by the user for RT_l's:

$$\text{Score}_l = \begin{cases} 1, \text{ if relevant} \\ 0, \text{ if no-opinion} \\ -1, \text{ if non-relevant} \end{cases}$$

5.2 Update of $W_{q,f}$ and $W_{f,r}$

By examing the retrieval process described earlier, we can see that both S and $S(f_i)$ are linear combinations of their corresponding lower level similarities. Because of the nature of linearity, these two levels can be combined into one, i.e.:

$$S = \sum_i \sum_j W_{q,r} S(r_{ij}) \tag{7}$$

where $W_{q,r}$ are the weights by which the information need in Q is distributed directly into r_{ij}'s.

For each r_{ij}, let RT^{ij} be the set containing the most similar N_{RT} objects to the query Q, according to the similarity values $S(r_{ij})$:

$$RT^{ij} = [RT_1^{ij}, ..., RT_l^{ij}, ..., RT_{N_{RT}}^{ij}] \tag{8}$$

To calculate the weight for r_{ij}, first initialize $W_{q,r} = 0$, and then use the following procedure:

$$W_{q,r} = W_{q,r} + Score_l, \quad if \quad RT^{ij} \quad is \quad in \quad RT$$

$$l = 0, ..., N_{RT}$$

After this procedure, if $W_{q,r} < 0$, set it to 0.

5.3. Update of $W_{r,r}$

In contrast to the linear similarity value calculation at the feature and the query levels (Equations (4) and (5)), the similarity calculation at the representation level (Equation (3)) can be any arbitrary *non-linear* function, such as Euclidean, Cosine, Histogram Intersection, etc. Because of the non-linearity, this level's similarity calculation can not be combined with the other two levels'. In the following, we describe how to update the weights, $W_{r,r}$, for the feature components r_{ijk}, $k = 1, ..., N$.

A standard deviation based weight updating approach has been proposed in [5]. For all the objects that are marked with *relevant* by the user, stack their r_{ijk}'s to form a $M \times N$ matrix, where M is the number of objects marked with *relevant*. In this way, each column of the matrix is a length-M sequence of r_{ijk}'s. If all the relevant objects have similar values for the component r_{ijk}, it means that the component r_{ijk} captures the user's perception subjectivity. On the other hand, if the values for the component r_{ijk} are very different among the relevant objects, then r_{ijk} does not capture the user's perception subjectivity. Therefore, the inverse of the standard deviation of the r_{ijk} sequence is a good estimation of the weight $W_{r,r}$ for r_{ijk}. That is, the smaller the variance, the larger the weight and vice versa. By incorporating relevance feedback to *dynamically* update $W_{r,r}$, the MIR system's retrieval performance is improved considerably [5].

5.3. Comparison Between Existing and Proposed Approaches

Most existing approaches [18,36] to MIR are based on the Pattern Recognition techniques developed in Computer Vision. The corresponding retrieval process can be summarized as follows:

1. Low-level multimedia features are extracted from the multimedia objects. Those features include, for example, color, texture, shape features for images; motion parameters for video; etc.

2. The multimedia objects are then represented by the set of feature vectors in the database. The user explicitly maps his information need into one or more of the low-level features supported by the retrieval system, possibly with different weights. Such a set of features with the associated weights is submitted as the query.

3. The query is considered as the matching pattern, and Pattern Recognition techniques are used to retrieve similar objects from the database.

While this Pattern Recognition based approach successfully establishes the basis of MIR, their performance is not satisfactory. This is because that the Pattern Recognition based approach requires the user to precisely map his perception subjectivity to low-level features with precisely specified weights. This mapping requires the user to have a comprehensive knowledge of the characteristics of all the low-level features, which is normally not the case.

In this section, based on the proposed multimedia object model for MPEG-7 (Section 3), we have further proposed a relevance feedback enhanced retrieval model. With the proposed retrieval model, the user is no longer required to decompose his high-level information need into low-level features and representations, instead, the user can submit a coarse initial query and continuously refine his information need via relevance feedback. This enables the retrieval system to invoke appropriate features and representations at the right time and right place. The proposed approach greatly reduces the user's effort of composing a query and captures the user's information need more precisely. The effectiveness of the proposed approach is demonstrated in [4,5,6,7].

6. EXPERIMENTAL RESULTS

In the experiments reported here, the image database is provided by Fowler Museum of Cultural History at the University of California-Los Angeles. The image database is part of the Museum Educational Site Licensing Project (MESL), sponsored by the Getty Information Institute.

In the current system, the multimedia features used include color, texture and shape of the objects in the image. To validate the proposed approach, multiple representations are used for each feature, e.g. color histogram and color moments [15] are used for color feature; coarseness-contrast-directionality [18] and co-occurrence matrix [23] texture representations are used for texture feature; Fourier descriptor and Chamfer shape descriptor [4] are used for shape feature. The proposed relevance feedback architecture is an *open* retrieval architecture. Other multimedia features or feature representations can be easily incorporated, if necessary.

Extensive experiments have been done in evaluating the system's retrieval performance. Users from various disciplines, such as Computer Vision, Art, Library Science, etc., were asked to compare the retrieval performance between the proposed approach and the Pattern Recognition based approach. The users rated the proposed approach much higher in terms of capturing their perception subjectivity and information need, which leads to a better retrieval performance. A typical retrieval process is given in Figures 4 and 5.

The user can browse through the image database. Once he finds an image of interest, that image is submitted as a query. Alternate to this query-by-example mode, the user can also submit images outside the database as queries. In Figure 4, the query image is displayed at the upper-left corner and the best 11 retrieved images are displayed in the order from top to bottom and from left to right. The retrieved results are obtained based on their overall similarities to the query image, which are computed from all the features and all the representations. Some retrieved images are similar to the query image in terms of shape feature

while other are similar to the query image in terms of color or texture feature. The user is no longer required to explicitly maps his information need to low-level features, but rather he can express his intended information need by marking the scores of relevance of the returned images. In this example, the user's information need is ``retrieve similar images based on their shapes''. Images 247, 218, 228 and 164 are marked relevant, while images 191, 168, 165, and 78 are marked non-relevant. From this feedback information, the system *dynamically* adjusts the weights, putting more emphasis on the *shape feature*. The improved retrieval results are displayed in Figure 5. Note that our shape representations are invariant to translation, rotation, and scaling. Therefore, images 164 and 96 are relevant to the query image.

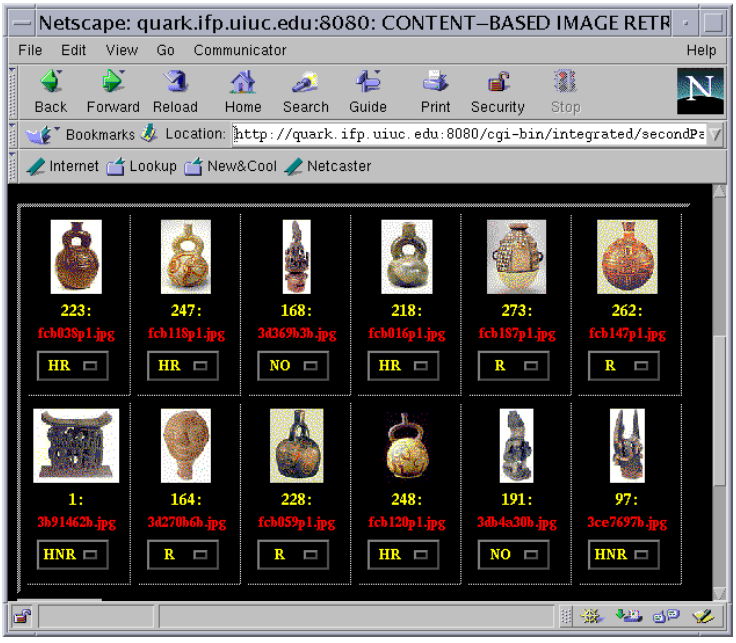

Figure 4. The initial retrieval results

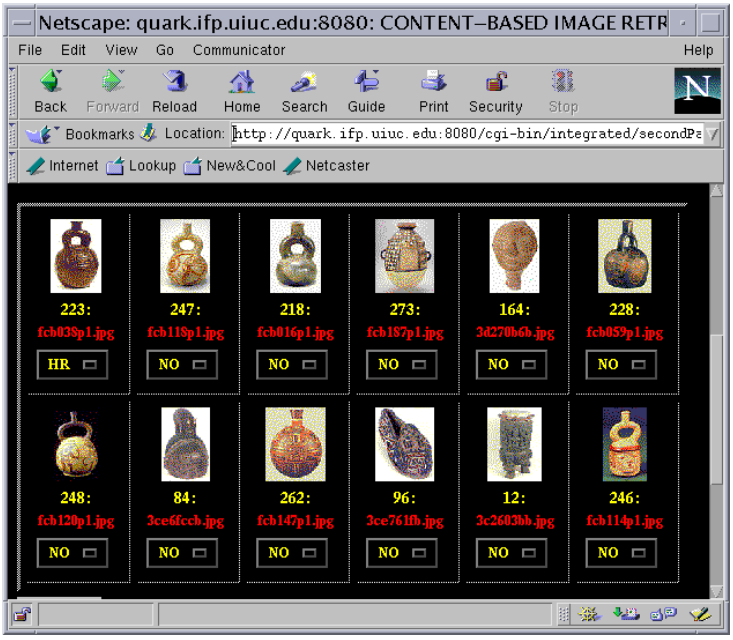

Figure 5. The retrieval results after relevance feedback

Unlike the Pattern Recognition based approach, where the user has to precisely decompose his information need into different features and representations and specify all the weights associated with them, the proposed approach allows the user to submit a coarse initial query and continuously refine his information need via relevance feedback. This approach greatly reduces the user's effort of composing a query and captures the user's information need more precisely.

7. CONCLUSIONS

In this proposal, we propose a multi-level multimedia object model for MPEG-7's multimedia content description. The proposed model supports information abstraction at various semantic levels. We also describe the interfaces between the multimedia object model and feature extraction and retrieval algorithm. Useful features are identified and their representations described. A relevance feedback enhanced retrieval algorithm is proposed, where the user's information need and perception subjectivity is better supported. The relevance feedback technique enables the retrieval system to invoke appropriate features and representations at the right time and right place. The system framework can be summarized in Figure 6.

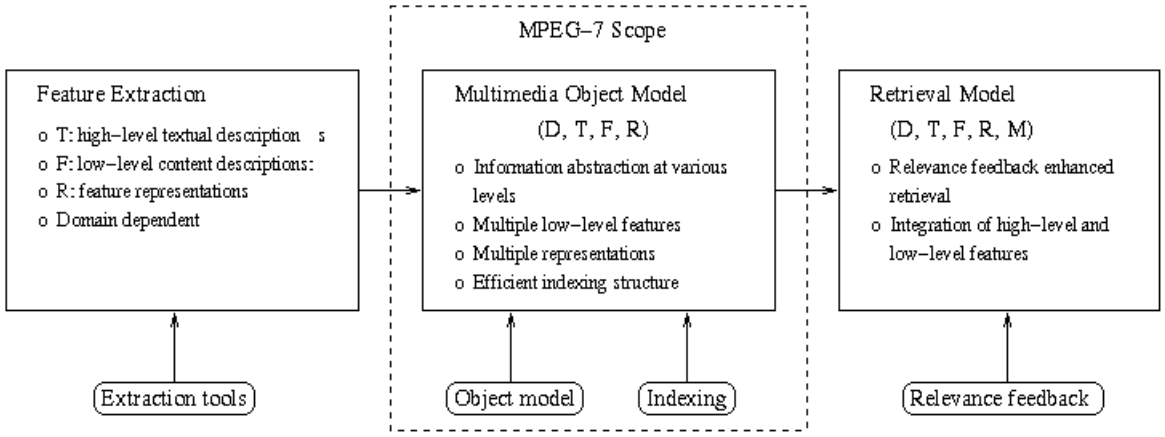

Figure 6. The system framework

MPEG-7 only specifies the multimedia object model and indexing scheme. What are the useful features for a particular application and how they are extracted is the responsibility of the sending end. Similarly, how to use the extracted
features, multimedia object model and indexing structure to support MIR is the responsibility of the receiving end.

8. ACKNOWLEGEMENT

The authors would like to thank Dr. Homer Chen, Rockwell Science Center for his valuable discussions.

9. REFERENCES

[1] W. M. Shaw, "Term-relevance computations and perfect retrieval performance", Information processing and Management.

[2] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983

[3] Buckley and G. Salton, "Optimization of relevance feedback weights", in Proc. Of SIGIR'95

[4] Y. Rui, T. S. Huang, S. Mehrotra, and M. Ortega, "Automatic matching tool selection using relevance feedback in MARS", in Proc. Of 2nd Int. Conf. On Visual Information Systems, 1997

[5] Y. Rui, T. S. Huang, and S. Merotra, "Content-based image retrieval with relevance feedback in MARS", in Proc. of IEEE Int. Conf. On Image Processing, 1997

[6] Y. Rui, T. S. Huang, S. Mehrotra, and M. Ortega, "A relevance feedback architecture in content-based multimedia information retrieval systems", in Proc. of IEEE workshop on Content-based Access of Image and Video Libraries, in conjunction with IEEE CVPR'97, 1997

[7] Y. Rui, T. S. Huang, and S. Mehrotra, "Human perception subjectivity and relevance feedback in multimedia information retrieval", submitted to SPIE Storage and Retrieval of Image/Video Databases VI, 1998

[8] T. S. Huang, S. Mehrotra, K. Ramchandran, "Multimedia analysis and retrieval system (MARS) project", in Proc. of 33rd Annual Clinic on Library Application of Data Processing – Digital Image Access and Retrieval, 1996

[9] S. Mehrotra, Y. Rui, K. Chakrabarti, M. Ortega, and T. S. Huang, "Multimedia analysis and retrieval system", submitted to 3rd Int. workshop on Information Retrieval Systems, 1997

[10] S. Mehrotra, Y. Rui, M. Ortega, and T. S. Huang, "Supporting content-based queries over images in MARS", in Proc. of IEEE Int. Conf. On Multimedia Computing and Systems, 1997

[11] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T. S. Huang, "Supporting similarity queries in MARS", in Proc. of ACM Conf. On Multimedia, 1997

[12] Y. Rui, A. C. She, and T. S. Huang, "Automated shape segmentation using attraction-based grouping in spatial-color-texture space", in Proc. of IEEE Int. Conf. On Image Processing, 1996

[13] Y. Rui, A. C. She, and T. S. Huang, "Modified Fourier descriptors for shape representation – a practical approach", in Proc. of 1st Int. Workshop on Image Databases and Multi Media Search, 1996

[14] Y. Rui, K. Chakrabarti, S. Mehrotra, Y. Zhao, and T. S. Huang, "Dynamic clustering for optimal retrieval in high dimensional multimedia databases", TR-MARS-10-97, 1997

[15] M. Stricker and M. Orengo, "Similarity of color images", In Proc SPIE 1995, No. 0-8194-1767-X/95

[16] R. Picard and T. P. Minka, "Vision texture for annotation", Multimedia Systems: Special Issue on Content-based Retreival, 1996

[17] T. P. Minka and R. W. Picard, "Interactive learning using a "Society of models"", In Proc. IEEE CVPR, 1996

[18] M. Flickner, et al. "Query by image and video content: the QBIC system", IEEE Computer, 1995

[19] C. Faloutsos, et al. "Efficient and effective querying by image content", IBM TR, 1993

[20] M. Charikar, C. Chekur, T. Feder, and R. Motwani, "Incremental clustering and dynamic information retrieval", in Proc. of the 29th Annual ACM Symposium on Theory and Computing, 1997

[21] M. J. Swain, "Interactive indexing into image databases", in Proc SPIE 1993, No. 0-8194-1141-8/93

[22] J. Huang, et al. "Image indexing using color correlogram", In Proc. of IEEE Conf. On CVPR, 1997

[23] P. P. Ohanian and R. C. Dubes, "performance evaluation for four classes of texture features", Pattern Recognition, Vol. 25, No. 8, 1992

[24] Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Tools for content-based manipulation of image databases", in Proc. SPIE Storage and Retrieval for Image and Video Databases II, vol 2, 1994

[25] S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data", TR, UCSB, 1995

[26] H. Wang, "Compressed-domain image search and applications", TR Columbia Univ., 1996

[27] J. R. Smith and S.-F. Chang, "Automated image retrieval using color and texture" TR Columbia Univ. 1996

[28] K. Jain, Fundamentals of Digital Image Processing, Prentice Hall

[29] M. K. Hu, Visual Pattern Recognition by Moment Invariants, Computer Methods in Image Analysis. IEEE Computer Society.

[30] E. M. Arkin, "An efficiently computable metric for comparing polygonal shapes" IEEE Trans. On PAMI, 1991

[31] H. G. Barrow, "Parametric correspondence and chamfer matching: Two new techniques for image matching", in Proc of 5th Int. Joint Conf. Artificial Intelligence.

[32] G. Borgefors, "Hierarchical chamfer matching: a parametric edge matching algorithm", IEEE Trans. PAMI, 1988

[33] T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves", IEEE Trans Computers, 1972

[34] G. C.-H. Chuang and C.-C. J. Kuo, "Wavelet descriptor of planar curves: Theory and applications", IEEE Trans Image Processing, 1996

[35] J. R. Smith and S.-F. Chang, "Querying by color regions using the VisualSEEk content-based visual query system", TR Columbia Univ

[36] J. R. Bach, et al. "The Virage image search engine: An open framework for image management", in SPIE Storage and Retrieval for Still Image and Video Databases IV.