
Multimedia Tools and Applications, 17, 97–120, 2002

Model-Based Video Classification toward Hierarchical Representation, Indexing and Access*

JIANPING FAN†
Department of Computer Science, University of North Carolina at Charlotte, 9201 University City BLVD, Chartlotte, NC 28223, USA

XINGQUAN ZHU
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

MOHAND-SAID HACID
LISI-INSA Lyon, 20 Avenue Albert Einstein, 69621 Villeurbanne, France

AHMED K. ELMAGARMID
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

Abstract. In this paper, we develop a content-based video classification approach to support semantic categorization, high-dimensional indexing and multi-level access. Our contributions are in four points: (a) We first present a hierarchical video database model that captures the structures and semantics of video contents in databases. One advantage of this hierarchical video database model is that it can provide a framework for automatic mapping from high-level concepts to low-level representative features. (b) We second propose a set of useful techniques for exploiting the basic units (e.g., shots or objects) to access the videos in database. (c) We third suggest a learning-based semantic classification technique to exploit the structures and semantics of video contents in database. (d) We further develop a cluster-based indexing structure to both speed-up query-by-example and organize databases for supporting more effective browsing. The applications of this proposed multi-level video database representation and indexing structures for MPEG-7 are also discussed.

Keywords: video analysis, video database, video retrieval

1. Introduction

Owing to the decreasing cost of storage devices, higher transmission rates, and improved compression techniques, digital video is becoming available at an ever increasing rate. To help users find and retrieve relevant information effectively, and to facilitate new and better ways of entertainment, advanced technologies need to be developed for analyzing, representation, indexing and semantic categorizing the vast amount of videos available in database [12].

Content-based video retrieving, navigating, and browsing have emerged as challenging and important areas in computer vision and database management. The present video

*This project is supported by National Sciences Foundation under 9972883-EIA, 9974255-IIS, and 9983249-EIA, and by grants from HP, IBM, Intel, NCR, Telcordia and CERIAS.
†To whom all correspondence should be addressed.

database systems first partition videos in database into a set of accessing units such as shots, objects or regions, and then follow the paradigm of representing videos via a set of feature attributes, such as color, texture, shape, and layout [4, 7, 8, 15, 20, 23, 24]. These representative features are archived along with the videos in database. A retrieval is then performed by matching the feature attributes of the query with those of videos in database that are *nearest* to the query object in high-dimensional spaces. The degree of similarity between the query video and these in database is measured by the Euclidean distances between their representative feature vectors [17, 27].

These query-based video database accessing approaches typically require that users provide an example video or sketch, and database management system is then searched for videos which are more relevant to the query. On the other hand, some approaches to video database management have focused on supporting hierarchical browsing of video contents. For supporting hierarchical video browsing, the video contents are first classified into a set of clusters on the basis of the similarity of their representative features [5, 28]. However, all these feature-based video database systems suffer from the following three major problems:

1. *Efficiency problem*: Since the video contents in databases are represented as independent data points in high-dimensional feature space, the similarity-based query is then equivalent to a *nearest neighbor* (NN) search. Multidimensional indexing structures, that have been investigated in recent years, seem to be a promising solution of this problem [3, 14, 19]. Unfortunately, the efficiency of these present multidimensional indexing structures deteriorates rapidly as the dimensions increase [26].
2. *Semantic problem*: It is not an easy task for a database user to express his or her queries appropriately in terms of the provided features, thus the naive users are interesting in browsing or querying the databases at semantic level. However, the low-level visual features do not correspond in a direct and convenient way to the underlying semantic structure of video contents [16, 21].
3. *Sufficiency problem*: Many data clustering techniques have been proposed in the past, and these data clustering techniques can also be used for video content clustering on the basis of the similarity of their low-level representative features [5, 28]. However, all these low-level feature-based video clustering techniques suffer from the sufficiency problem [25].

Based on the above observations, we propose a novel learning-based video clustering and cluster-based hierarchical indexing technique for solving the efficiency, sufficiency and semantic problems. This paper is organized as follows. Section 2 proposes a hierarchical video database model for supporting multi-level video representation, indexing, retrieving, and browsing. Our works on content-based video analysis are introduced in Section 3. A novel semantic clustering algorithm is proposed in Section 4. By selecting the suitable dimensional weighting coefficients, a learning-based optimization technique is used for finding the "ideal" dimensional features that account for the visual similarity in human concept. Section 5 describes applications of this proposed multi-level video database representation

and indexing structures for MPEG-7. Section 6 discusses applications for access control. We conclude in Section 7.

2. Hierarchical video database model

From the database point of view, a powerful video model is a premise that will enable a good basis of content-based retrieving and browsing of video data. As far as video database modeling is concerned, we can make distinction between two important things that should be modeled: *structure* and *content* of a video database. Unlike a traditional video model, a video database model should include the elements that represent inherent structures of videos in database and the semantics that represent the video contents.

In this section, we introduce the model used in the development of our content-based video database system. In our video database model, the hierarchical structure of video database is exploited by partitioning the video contents into a set of hierarchical manageable units as shown in figures 1 and 2, such as clusters, subclusters, subregions, shots or objects, frames or VOPs (video object planes), and regions, so that more efficient video representation, indexing, and accessing techniques can be supported. Moreover, the semantics of video database are also exploited by an interactive machine learning procedure, so that high-level concept-based querying, browsing and navigating can be supported.

The basic video accessing units such as shots and key objects are first obtained and represented by a set of visual, meta and semantic features. These obtained video shots and video objects, which convey the video contents in database, are then classified into a set of semantic clusters, and each semantic cluster may consist of a set of subclusters. The subclusters can further be partitioned into a set of subregions for obtaining more compact representation. Each subregion consists of a limited number of similar video contents (video

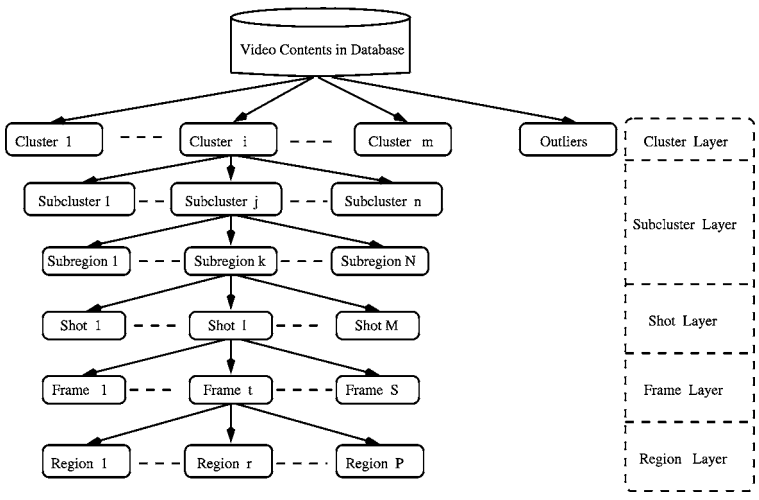


Figure 1. Hierarchical video database model for shot-based video accessing approach.

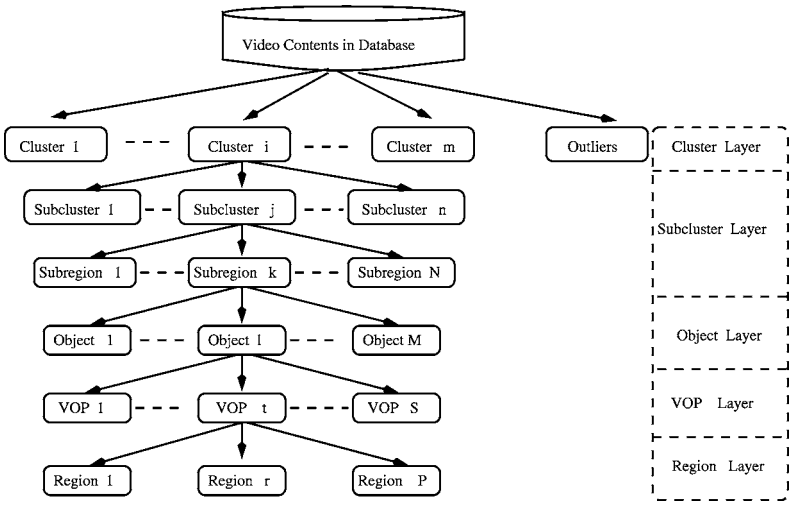


Figure 2. Hierarchical video database model for object-based video accessing approach.

shots or key objects). The semantics on object level or scene level can be exploited by using domain knowledge or unsupervised clustering in the video analysis procedure [9, 22]. The semantics on database level are obtained by a learning-based video clustering procedure as described in Section 4.

The cluster layer is the highest layer in our model, it consists of a set of semantic clusters, which are used for describing the physical structures and semantics of video contents in database. We propose a novel learning-based video clustering technique to exploit this highest layer. The subcluster layer includes the physical structures and compact semantic contents of the clusters. The subcluster layer is obtained by discovering the interesting relationships and characteristics that may exist implicitly in the cluster. We will see that including subcluster layer can provide more efficient video database indexing structure. The video shot or object layer describes the video representation, indexing and accessing units used in our system. The frame or VOP layer represents the visualization of video content at a special time. The region layer describes the spatial components of a visual content and their relationships.

All these video layers are represented by a set of meta, visual, and semantic features as shown in figures 3 and 4. For the cluster layer, each component is characterized by the cluster centroid, radius, feature dimensions, subcluster number, dimensional weighting coefficients, and its node identifier. The cluster centroid and radius are represented by a set of visual features which are also used for describing the video contents in the same cluster. For the subcluster layer, each component is also characterized by the subcluster centroid, radius, feature dimensions, subregion or object number, dimensional weighting coefficients, and its leaf node identifier. The subcluster centroid and radius are also represented by a set of visual features. For the shot or object layer, each component is represented by an indexing identifier, meta features, semantic features, and a set of visual features which are the average and variance of that of the frames or VOPs in the shot. For the frame or VOP layer, each

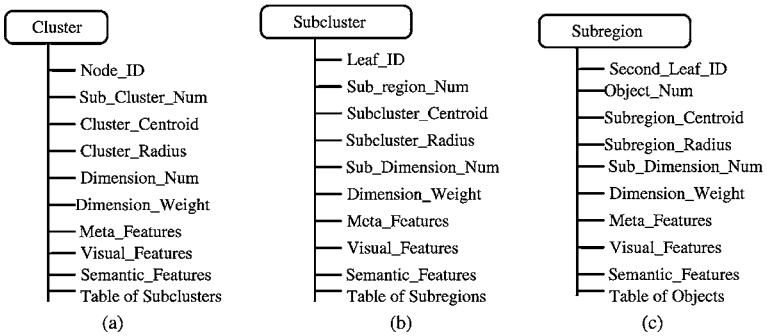


Figure 3. The multi-layer video database representation schemes: (a) semantic cluster; (b) subcluster; (c) subregion.

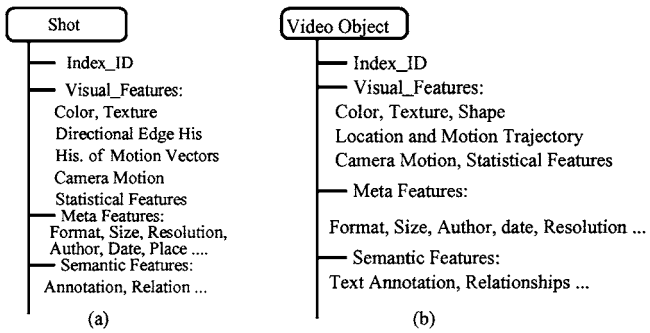


Figure 4. The representative features: (a) video shot; (b) video object.

component is represented by meta features, semantic features, and a set of visual features which can be obtained from the image regions.

Since all these video database representation layers are characterized by a set of same types of visual, meta, and semantic features, this proposed multi-layer video database model can provide a framework for automatic mapping from features to concepts through a learning-based clustering technique. This multi-level abstraction and representation scheme can also provide a scalable method for retrieving and viewing video contents in database.

3. Content-based video analysis

There have two approaches to accessing video source in databases: *shot-based* and *object-based*. The objective of video analysis is to obtain these basic video accessing units (e.g., shots and objects). We have proposed a set of useful techniques for obtaining these units, figure 5 shows the block diagram of this proposed automatic video content analysis and classification scheme.

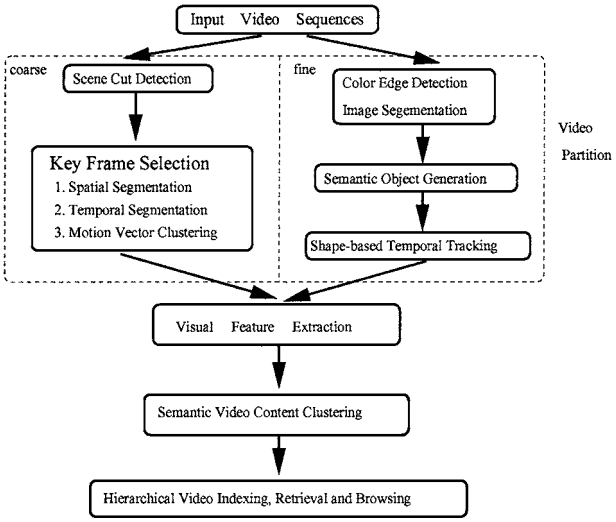


Figure 5. The block diagram of the proposed automatic video partition scheme.

3.1. Video shot detection

Video shots, which are directly related to video structures and contents, are the basic units to be used for accessing video sources. A fundamental task in video analysis is to extract such structures and contents from the video to facilitate user's accessing (retrieving and browsing). Only after such video structure and content information become available can content-based retrieving, browsing and manipulation of video data be facilitated.

The relationships among successive frames in a video sequence can be classified into two opposite categories: *scene cut* versus *non-scene cut*. In this sub-section, we propose a novel scene cut detection algorithm that is based on two-class data classification and the optimal threshold can be automatically determined by a fast entropic thresholding technique [10]. One advantage of our scene cut detection technique is that the threshold for scene cut detection can be adapted to the activities of variant video sequences. Our scene cut detection algorithm first calculates the color histogram differences among successive frames:

$$HD(j, j-1) = \sum_{k=0}^{M} \frac{[H_{j-1}(k) - H_j(k)]^2}{[H_{j-1}(k) + H_j(k)]^2}$$

where $H_j(k)$ denotes the color histogram of the jth frame, $H_{j-1}(k)$ indicates the color histogram of its previous $(j-1)$th frame, k is one of the M potential color components. If $HD(j, j-1)$ is above an optimal threshold $\bar{T}_c$, the jth frame is detected as a scene cut.

The probability $P_{nsc}(i)$ for *non-scene cut* is defined as:

$$P_{nsc}(i) = \frac{f_i}{\sum_{h=0}^{T} f_h}, \quad 0 \le i \le T$$

where f_i denotes the number of frames for which the color histogram differences with previous frames are equal to i, $\sum_{h=0}^{T} f_h$ represents the total number of frames for which the color histogram differences with previous frames are in the range $0 \le i \le T$.

The probability $P_{sc}(i)$ for *scene cut* is defined as:

$$P_{sc}(i) = \frac{f_i}{\sum_{h=T+1}^{M} f_h}, \quad T+1 \le i \le M$$

The entropies for non-scene cut and scene cut frames are defined as:

$$H_{nsc}(T) = -\sum_{i=0}^{T} P_{nsc}(i) \log P_{nsc}(i), \quad non\_scene\_cut$$

$$H_{sc}(T) = -\sum_{i=T+1}^{M} P_{sc}(i) \log P_{sc}(i), \quad scene\_cut$$

The optimal threshold $\bar{T}_c$ is determined automatically by maximizing the following criteria function:

$$H(\bar{T}_c) = \max_{T=0,1,\ldots,M} \{H_{nsc}(T) + H_{sc}(T)\}$$

A fast searching technique has been developed for reducing the computation burden to $O(M)$ [10].

The *temporal relationships* among successive frames in a video sequence are then classified into two opposite classes on the basis of their color histogram differences and the obtained optimal threshold $\bar{T}_c$: *scene cut* versus *non-scene cut*.

$$\begin{cases} HD(j, j-1) > \bar{T}_c, & scene\_cut \\ HD(j, j-1) \le \bar{T}_c, & non\_scene\_cut \end{cases}$$

The video frames between two successive scene cuts are taken as one video shot. If the length of a video shot (number of frames) is less than a pre-defined threshold, the corresponding scene cut may be induced by flash light (e.g., scene cuts induced by flash light are widely distributed in video News), and thus a false elimination procedure is included for removing these false alarms.

The key frames, that are used for representing the content abstract for fast browsing, can be selected according to the following criteria [10]:

Figure 6. The detected video shot boundaries from a CCTV News.

1. *Shot-based criteria*: Given a video shot, the scene cut frame should be taken as a key frame, whether more than one key frame need to be chosen in a shot depends on the following two criteria.
2. *Camera-based criteria*: Global motion of camera is one important source pf video content changes and should be taken as a critical feature for key frame selection.
3. *Activity-based criteria*: Another important source of video content changes among frames is the active moving objects and it should be taken as a critical feature for key frame selection.

The experimental results for two CCTV video News are given in figures 6 and 7. Some semantic scenes have been obtained by exploiting the domain knowledge and time constraint [22].

3.2. Semantic object generation

The previous shot-based video accessing and representation technique does not capture the underlying semantic structure of video sources. Extracting semantic structure of video sources is very important for providing more effective video retrieving and browsing because

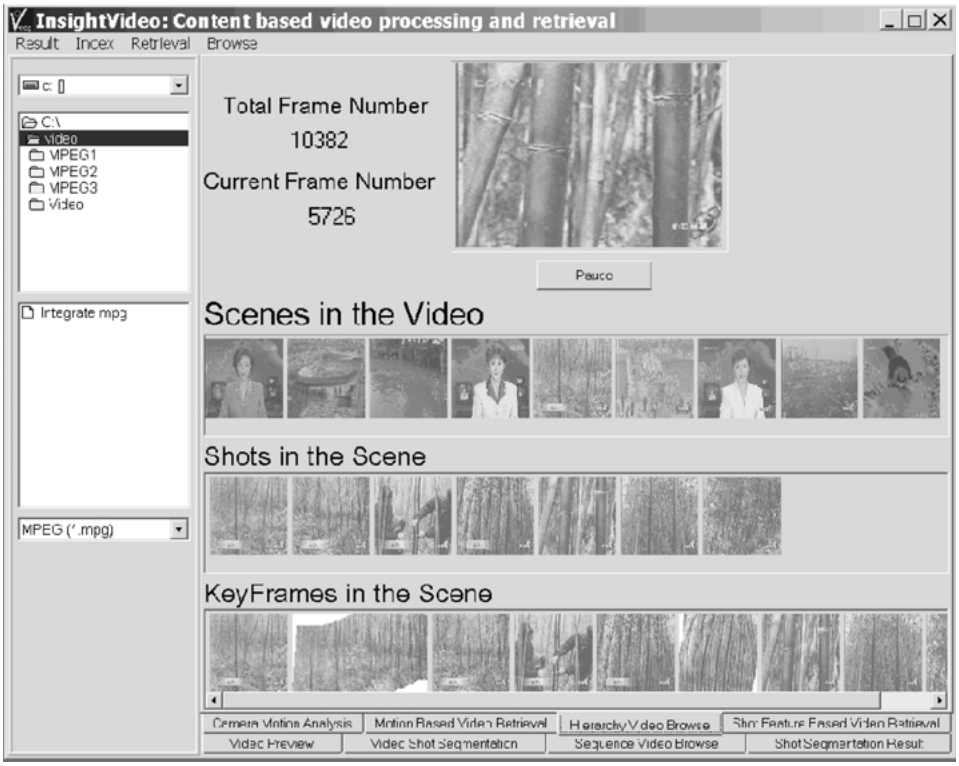


Figure 7. The detected semantic scenes, video shots and key frames from a CCTV News.

people watch the video based on its semantic contents not on its physical shots or key frames. Due to their inherent content-dependence, video objects are especially suitable for representing semantic video contents.

It is very difficult to design an universal semantic object generation technique, which can provide variant semantic objects by using the same function [2, 6, 11, 13]. However, semantic object generation for content-based video database application becomes possible because the videos can be indexed by some semantic objects of interest for the users, such as human being, cars, airplanes. This interest-based video indexing approach is reasonable because the users do not focus on all the objects presented in the videos [9]. Hence, the difficulties of automatic semantic object generation for video database applications will be reduced.

Based on the above observations, several independent object generation functions can be designed and each function can provide one type of semantic object. Each function is designed by using the *object seed* and *region constraint graph* (perceptual model) of the corresponding semantic object. The selected object seed should represent the distinguished characters of the corresponding semantic object, and the region constraint graph can guide how the connected regions of the object seed should be put together for generating the

corresponding semantic object. Our semantic object generation technique takes the following steps:

1. An efficient color edge detection procedure is first performed on the three color components for exploiting potential edges among variant image components.
2. A region growing algorithm is performed on the luminance component for providing homogeneous image regions with closed boundaries.
3. The obtained color edges and region growing results (e.g., region boundaries) are then integrated for providing more reasonable homogeneous image regions with accurate boundaries.
4. The semantic object generation function then tries to find the corresponding object seed from these obtained homogeneous regions.
5. If the object seed is detected, a *seeded region aggregation* procedure is used for merging the adjacent regions of the object seed as the semantic object. For example, the human face can be taken as the object seed for semantic human object generation, and the region constraint graph (perceptual model) of human being, can be used for managing the seeded region aggregation procedure. The perceptual model of human being can guide the way the adjacent regions corresponding to face (or head, taken as object seed), body, arms, legs should be put togehter. A novel semantic human being generation scheme is proposed and its major steps are shown in figure 8, where the detected face is taken as the object seed and the perceptual model is used for managing the region aggregation procedure.
6. The generated semantic objects can then be tracked along the time axis for exploiting their correspondences among frames.

A set of experimental results are given in figures 9 and 10. One can find that our seeded semantic object generation technique is very attractive for multiple object extraction, because different semantic objects have different seeds.

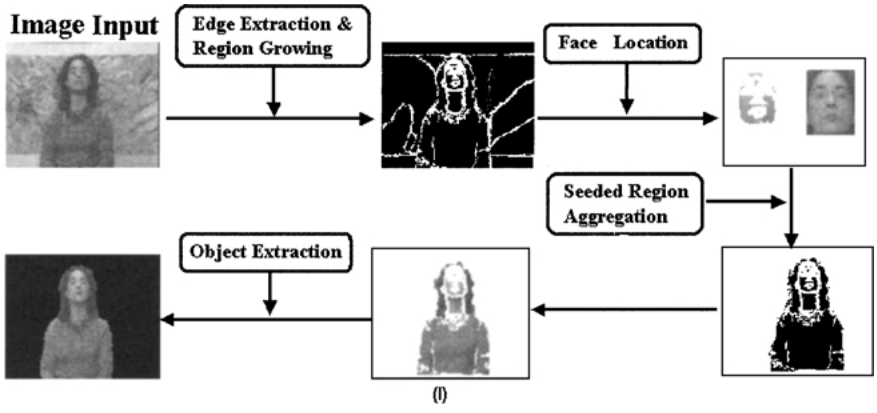


Figure 8. The major steps for semantic human object generation.

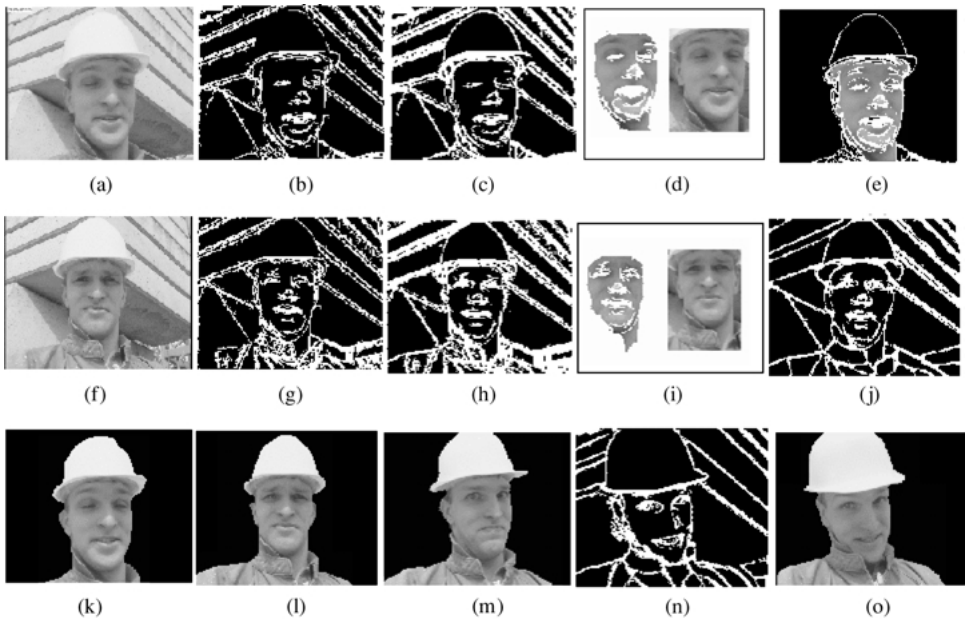


Figure 9. (a) The reference frame of "Foreman"; (b) the intensity edges; (c) the color edges; (d) the human face and its rectangular region; (e) the connective object edges; (k) the object in reference frame; (f) the original 15th frame; (g) the intensity edges; (h) the color edges; (i) the human face and its rectangular region; (j) the connective object edges; (l) the tracked semantic object in 15th frame; (m) the tracked semantic object in 99th frame; (n) the region boundaries for 149th frame; (o) the tracked semantic object in 149th frame.

4. Semantic video content clustering

Classifying variant video contents into a set of semantic clusters is very important for managing large scale video databases:

1. *Querying requirement*: To achieve the scalability of searching video database, it must be ensured that the search time does not increase linearly with the database size. Our solution of this problem is to create a novel semantic clustering and hierarchical indexing scheme, so that at the time of query, only the relevant clusters need to be examined.
2. *Indexing requirement*: Since there has high overlapping in high-dimensional feature space, the performance of NN-based query techniques rapidly deteriorates when the number of feature dimensions increases. Our solution of this problem is to develop a cluster-based hierarchical video indexing structure which performance is not largely depended on the number of dimensions.
3. *Browsing requirement*: Classifying variant video contents into a set of semantic clusters in human concept is also very important for fast browsing of video databases, because users may just want to browse the video databases through the semantic categories. Our solution of this problem is to develop a learning-based clustering technique to exploit the semantics of video contents in databases.

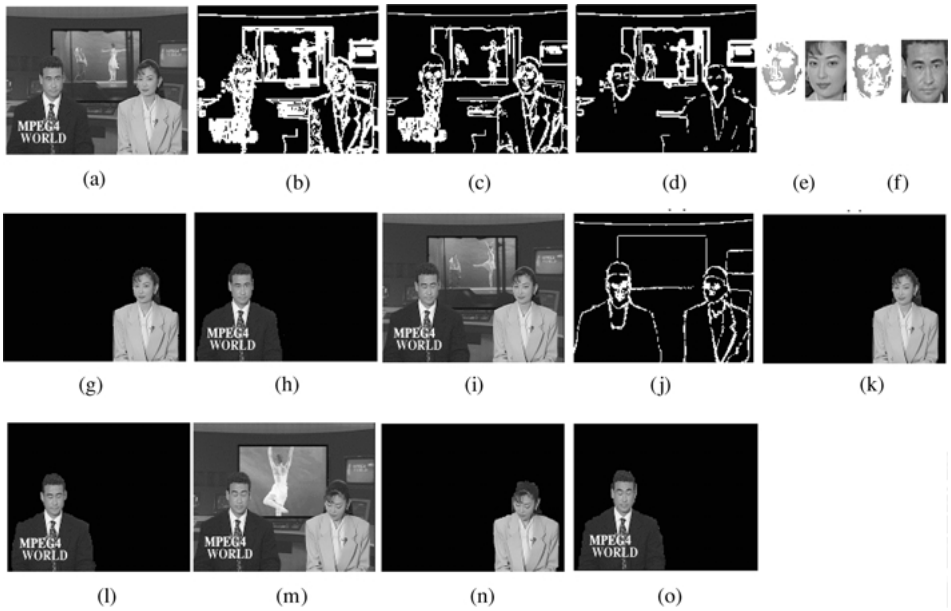


Figure 10. The object extraction results from "News". First frame: (a) original image; (b) color edges; (c) luminance edges; (d) chrominance edges; (e) human face of object1; (f) human face of object 2; (g) object 1; (h) object 2; 10th frame: (i) original image; (j) region boundaries; (k) tracked object 1; (l) tracked object 2; 260th frame: (m) original image; (n) tracked object 1; (o) tracked object 2.

4.1. Dimensional weight detection

After the automatic video analysis procedure, video contents (shots and objects) are characterized by a set of representative features as shown in figure 4(a) and (b). However, not all n dimensions contribute equally to defining the visual similarity in human concept, so we need to give a weight to each dimension. Furthermore, among n feature dimensions, only some of them may be relevant to defining the visual similarity in human concept, while the remaining ones should be reduced. We suggest a novel learning-based optimization technique to select the suitable dimensional weighting coefficients for providing semantic clustering.

There are two different similarity measures for comparing two video contents with semantic labels s and t, the *weighted feature-based similarity distance* $d_F(O_s, O_t)$ and the *semantic similarity distance* $d_S(O_s, O_t)$ in human concept.

$$d_F(O_s, O_t) = \sum_{i=1}^{n} \frac{1}{a_i} d_F^i\left(O_{s_i}, O_{t_i}\right) \tag{1}$$

$$d_S(O_s, O_t) = \sum_{i=1}^{n} d_S^i\left(O_{s_i}, O_{t_i}\right) \tag{2}$$

where a_i is the ith dimensional weighting coefficient, $d_F^i(O_{s_i}, O_{t_i})$ is the ith dimensional feature-based similarity distance between objects O_s and O_t, $d_S^i(O_{s_i}, O_{t_i})$ is the ith dimensional semantic distance between objects O_s and O_t in human concept such as shape, color, texture.

$$d_F^i(O_{s_i}, O_{t_i}) = \sum_{j=1}^{N}\sum_{k=1}^{N} m_{jk}\left(f_{s,j}^i - f_{t,j}^i\right)\left(f_{s,k}^i - f_{t,k}^i\right) \tag{3}$$

$$d_S^i(O_{s_i}, O_{t_i}) = \begin{cases} 0, & \text{if } s_i = t_i \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where $f_{s,j}^i$ is the ith dimensional visual feature of the jth training sample, an $N \times N$ matrix $\mathbf{W_i} = [m_{jk}]$ defines a *generalized ellipsoid distance*, N is the total number of training samples, s_i and t_i indicate the ith dimensional semantic lable in human concept for objects O_s and O_t.

We first assume that the video contents in database should be partitioned into a set of semantic clusters according to some well-known human concepts such as *sport*, *news*, *fashion*, *film*, *landscape etc*, and the primitive *icon video contents* (semantic templates) for each semantic cluster are selected by human being (expert). These primitive icon video contents (or semantic templates) for each semantic cluster should be intuitive, understandable and representative video examples or animated sketches in human concept. These icon video contents can be taken as the *initial seeds* for semantic video content clustering and they are also described by a set of representative features. Note that a icon video content in high-level human concept is characterized by a set of representative features.

We do not know which feature dimension is more important for accounting the visual similarity in human concept, a machine-learning procedure takes care of finding the "ideal" dimensions for us. A set of *training* video contents, whose semantic labels in human concept and their representative features have been given, are used to determine the dimensional weighting coefficients for semantic clustering. Since the concept-based similarity distances among these labeled training video contents are given, the system then learns from these training video contents, and makes the weighted feature-based similarity among video contents correspond directly to their concept-based similarity by selecting the suitable dimensional weighting coefficients. The weighting coefficient for each dimension is determined by a learning-based optimization procedure:

$$\begin{aligned} \text{Positive Examples:} \quad \min \quad & \left\{ d_F(O_s, O_t) = \sum_{i=1}^{n} \frac{1}{a_i} d_F^i(O_{s_i}, O_{t_i}) \right\} \\ \text{subject to:} \quad & \sum_{i=1}^{n} d_S^i(O_{s_i}, O_{t_i}) \approx 0 \\ & \sum_{i=1}^{n} \frac{1}{a_i} = 1 \\ & det(W_i) = 1 \end{aligned} \tag{5}$$

$$\text{Negative Examples:} \quad \max \quad \left\{ d_F(O_s, O_t) = \sum_{i=1}^{n} \frac{1}{a_i} d_F^i\left(O_{s_i}, O_{t_i}\right) \right\}$$
$$\text{subject to:} \quad \sum_{i=1}^{n} d_S^i\left(O_{s_i}, O_{t_i}\right) \approx n \tag{6}$$
$$\sum_{i=1}^{n} \frac{1}{a_i} = 1$$
$$det(W_i) = 1$$

The *Lagrangian* optimization technique can be used for selecting the suitable weighting coefficients [16, 21]. The "ideal" dimensional features, which have big weighting coefficients, are more important on making the training video objects close to their similar objects (positive examples) but far from their dissimilar video objects (negative examples).

4.2. *Seeded video clustering*

Given the initial seeds of semantic clusters $\Psi = \{s_1, s_2, \ldots, s_q\}$ and their dimensional weighting coefficients, *seeded-region-growing* (SRG) technique [1] is then used to classify variant video contents into a set of semantic clusters. Let S be the set of all unallocated video contents and $e = [x_1, x_2, \ldots, x_n]$ be a random unallocated video content represented by n-dimensional features. For $e \in S$, the weighted feature-based similarity distance $d_F(e, s_i^j)$ between the testing video content e and the jth video seed s_i^j of cluster A_i is:

$$d_F\left(e, s_i^j\right) = \sum_{r=1}^{n} \frac{1}{a_r} d_F^r\left(e_r, s_{ir}^j\right) \tag{7}$$

where a_r is the rth dimensional weighting coefficient, $d_F^r(e_r, s_{ir}^j)$ denotes the similarity distance between e and s_i^j on the basis of their rth dimensional features. Since the cluster A_i may have a set of video seeds, the final similarity distance $d(e, A_i)$ between e and A_i is determined as:

$$d(e, A_i) = \min_{j=1,2,\ldots,p} \left\{ d_F\left(e, s_i^j\right) \right\} \tag{8}$$

where p is the total number of the given icon videos for the cluster A_i. The semantic cluster A_k, which has the smallest weighted feature-based similarity distance with e, is determined by:

$$d_k(e) = \min_{i=1,2,\ldots,q} \{ d(e, A_i) \} \tag{9}$$

If $d_k(e)$ is less than a pre-defined threshold $\bar{T}$, e is involved into the cluster A_k, otherwise, e is taken as a new outlier.

$$\begin{cases} d_k(e) \le \bar{T}, & \text{merge } e \text{ into } A_k \\ d_k(e) > \bar{T}, & \text{take } e \text{ as outlier} \end{cases} \tag{10}$$

One can find that our SRG-based semantic clustering technique can handle new data to be added efficiently.

4.3. Dimension reduction

Given the dimensional weighting coefficients $\{a_1, a_2, \ldots, a_n\}$ for a semantic cluster, the degree of importance of its n-dimensional representative features is also given. Bigger dimensional weighting coefficients mean that the associated representative features are more important in predicting the judgment of similarity by humans. For a cluster A_i, its *centroid* $\bar{x}_c^i = \{\bar{x}_{1,c}^i, \bar{x}_{2,c}^i, \ldots, \bar{x}_{n,c}^i\}$ can be defined as:

$$\begin{cases} \bar{x}_{1,c}^i = \dfrac{\sum_{h=1}^{M} x_{1,h}}{M} \\ \quad \vdots \\ \bar{x}_{j,c}^i = \dfrac{\sum_{h=1}^{M} x_{j,h}}{M}, \quad (x_{1,h}, \ldots, x_{j,h}, \ldots, x_{n,h}) \in A_i \\ \quad \vdots \\ \bar{x}_{n,c}^i = \dfrac{\sum_{h=1}^{M} x_{n,h}}{M} \end{cases} \tag{11}$$

where M is the total number of video contents in the cluster A_i, $\bar{x}_{j,c}^i$ is its *projected centroid* on the jth dimension, and $\{x_{1,h}, \ldots, x_{j,h}, \ldots, x_{n,h}\}$ indicates the *dimensional attributes* of the video content in the cluster A_i. The *radius* of the cluster $\varphi_c^i = \{\varphi_{1,c}^i, \ldots, \varphi_{j,c}^i, \ldots, \varphi_{n,c}^i\}$, which represents the average generalized ellipsoid distance between the objects and the cluster centroid, can be defined as:

$$\begin{cases} \varphi_{1,c}^i = \dfrac{\sum_{h=1}^{M} \sum_{k=1}^{M} m_{hk}\left(x_{1,h} - \bar{x}_{1,c}^i\right)\left(x_{1,k} - \bar{x}_{1,c}^i\right)}{M} \\ \quad \vdots \\ \varphi_{j,c}^i = \dfrac{\sum_{h=1}^{M} \sum_{k=1}^{M} m_{hk}\left(x_{j,h} - \bar{x}_{j,c}^i\right)\left(x_{j,k} - \bar{x}_{j,c}^i\right)}{M} \\ \quad \vdots \\ \varphi_{n,c}^i = \dfrac{\sum_{h=1}^{M} \sum_{k=1}^{M} m_{hk}\left(x_{n,h} - \bar{x}_{n,c}^i\right)\left(x_{n,k} - \bar{x}_{n,c}^i\right)}{M} \end{cases} \tag{12}$$

Radius is a good quality measure of a clustering technique. Small value of radius means that all the similar objects in the same cluster are distributed more densely. Large value of

radius indicates that the similar objects in the same cluster are distributed sparsely. From the clustering quality point of view, we hope the radius of a cluster is small enough so that the cluster only consists of the similar objects. From the indexing point of view, we hope that the similar objects in the same cluster should be distributed sparsely so that they can be separated efficiently. A good trade-off between these two problems should be found. We first make the radius of the cluster is below a threshold so that the cluster only consists of the similar objects, and then we select the "principal" dimensions with large radius so that the similar objects in the same cluster can be separated efficiently to support more effective indexing. Therefore, the dimensional features, which have smaller weighting coefficients (less important) and smaller dimensional radius (data points are distributed more densely on them), can be reduced to support more effective multidimensional video indexing.

5. Applications for MPEG-7

To define exchangeable formats, MPEG has initiated a new work item, formally called "Multimedia Content Description Interface", better known as MPEG-7. MPEG-7 aims to create a multimedia content description standard in order to facilitate multimedia searching and filtering application. In the context of MPEG-7, a description of an audiovisual (AV) document includes descriptors (termed Ds), which specify the syntax and semantics of a representation entity for a feature of the AV data, and description schemes (termed DSs) which specify the structures and semantics of a set of Ds and DSs. Descriptions are expressed in a common description definition language (DDL) to allow their exchange and access.

The video database is represented as the multiple hierarchies of the semantic clusters that classify individual contents in database based on their representative features. Moreover, general relations among the semantic clusters can be exploited and represented by a constraint graph. In our work, the hierarchical property of video database has been exploited for building the indexing hierarchies and generating multiple levels of abstraction. This multi-level abstraction scheme provides a scalable method for retrieving and browsing video contents in database, one can find this multi-level indexing structure and abstraction are very attractive for MPEG-7 applications. In this section, we develop a database-level description scheme on the basis of our multi-level video model and hierarchical indexing structure.

The hierarchical video database DS as shown in figure 11 describes the physical orgnization (structure) of the database and its semantic properties. The physical structure involves the descriptions of the ontology of the semantic clusters to be used in the description, the multi-level representation, abstraction and indexing structures of database and the spatio-temporal orgnization of videos in database. The ontological structures of the semantic clusters are obtained by the learning-based clustering procedure. The partition tree of video database as described in Section 2 allows the creation of multi-level DSs of the hierarchical representation, abstraction and indeixng structures. The spatio-temporal organization structures of videos have been exploited by the video analysis procedure. The semantic relationship, which is represented in this hierarchy as shown in figure 11, is of the type "is-made-of " to address the high-level description of database. One can find that the higher level DS is an aggregation of a set of lower level DSs.

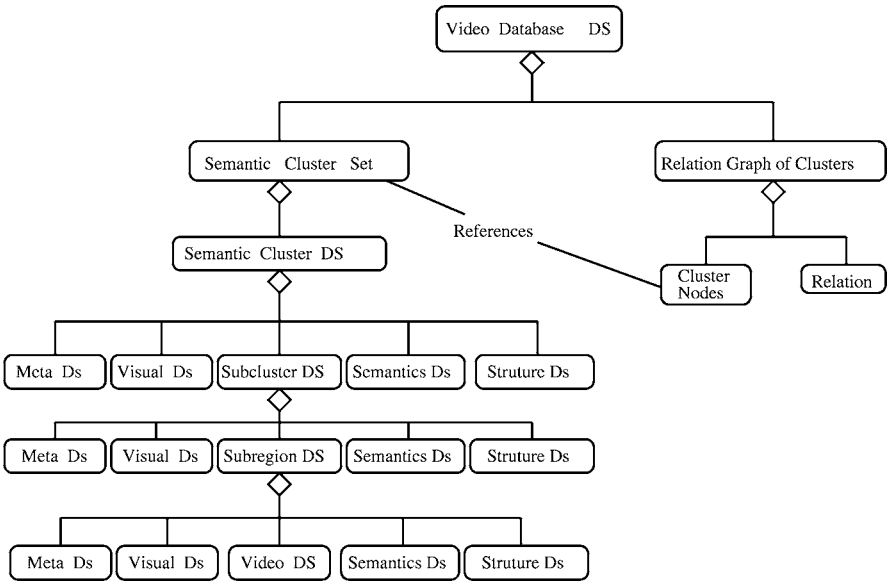


Figure 11. The video database description scheme, where symbol ◇ represents the aggregation of sub-DSs, **Ds** indicates the descriptors and **DS** denotes description scheme.

Our system supports two approaches to accessing the video contents in database: shot-based and object-based. The shot-based video DS as shown in figure 12 is to define the shot-based temporal and spatial organization strutures of a video and to describe its visual properties. The object-based video DS as shown in figure 13 is to describe the object-based spatio-temporal organization structure of a video and to represent its visual properties. The image DS as shown in figure 14(a) is used to define the spatial organization structure of an image and to represent the relationships among the regions. The object DS as shown in figure 14(b) is to describe the hierarchical organization structure of object components and to represent the visual properties of object. These DSs consist of a set of descriptors (Ds) which are characterized by meta, visual, and semantic representative features.

6. Applications for access control

All the present video database systems try to provide *free* and *equal* accessing to information, but the truth of the matter is that information was never free to begin with. We also know that not all information is intended for every person, this is especially true in business world. Video database is also used in variant environments with very different objectives, it is often the case that different classes of users must receive different authorizations for the same set of video data. Therefore, content-based accessing control is also becoming one of emerging problems, because network users have different permission to access different types and qualities of videos in database.

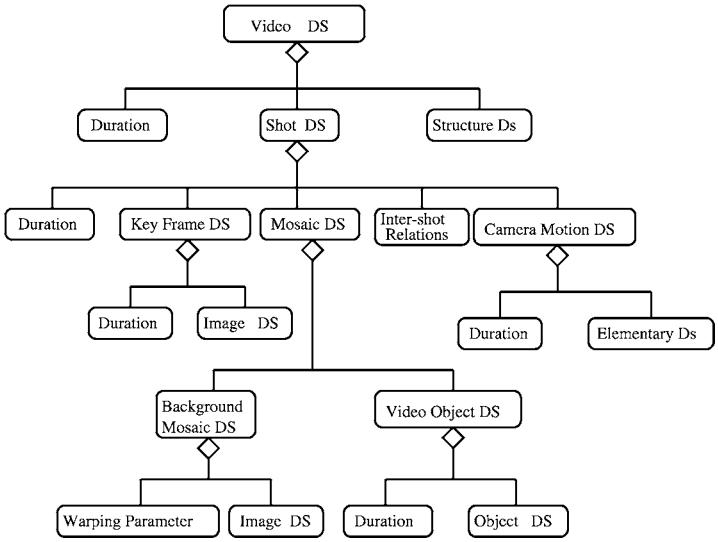


Figure 12. The shot-based video description scheme.

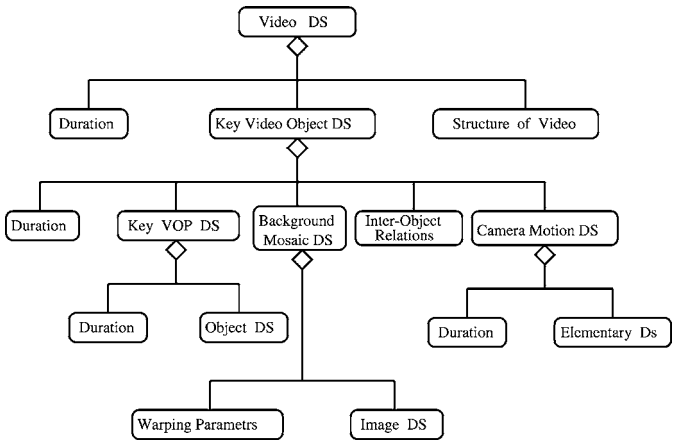


Figure 13. The object-based video description scheme.

The practical video database accessing control techniques should have the following properties:

1. It should be user-adaptive because different classes of users have different permission to access different types of video contents or even different levels of the same video content in database.
2. It should be content-adaptive because the video contents in database are used for very different objectives.

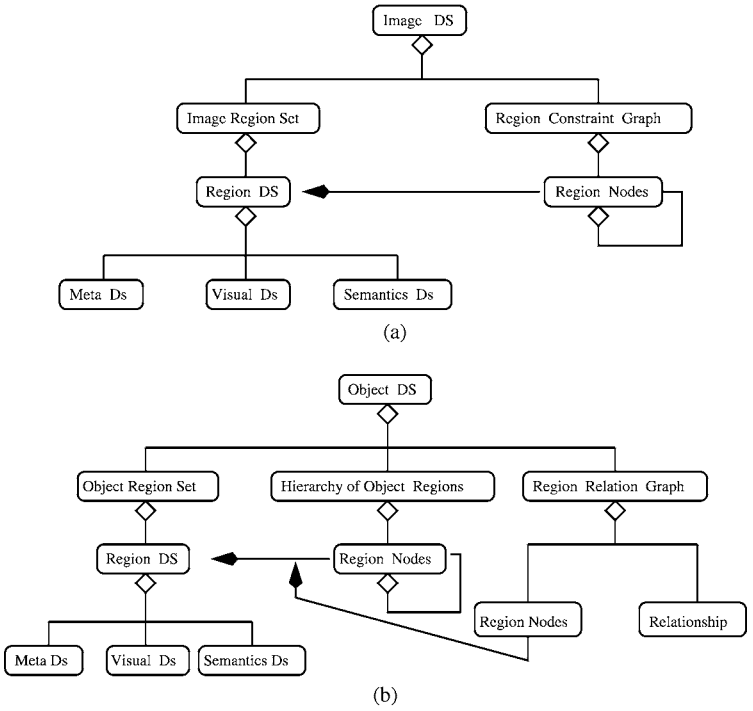


Figure 14. The image description scheme: (a) frame-based approach; (b) object-based approach.

3. It should be hierarchical or multi-level because different users may have different permission to access different levels of the same video data.

6.1. Access object specification

In our hierarchical video database model, a video element can be represented as either a semantic cluster, a subcluster with special contents, a video stream, a video segment (e.g., video shot or video object), a video frame, or even a region of interesting. The specification of an *authorization object* is based on these video elements, and these video elements are characterized by a set of visual, meta, and semantic features as shown in figure 2.

The authorized objects, that are used for video database accessing, consist of three major components:

1. The first component is the representative features and their dimensional weighting coefficients which are used for characterizing the semantic video contents in databases as shown in figures 3 and 4. This component is also used in the traditional *free* and *equal* video database systems. Moreover, the semantics of video elements have been exploited by a learning-based clustering technique, thus content-based video database accessing control can be based on not only the semantics of the video elements but also the weighted attributes characterizing them.

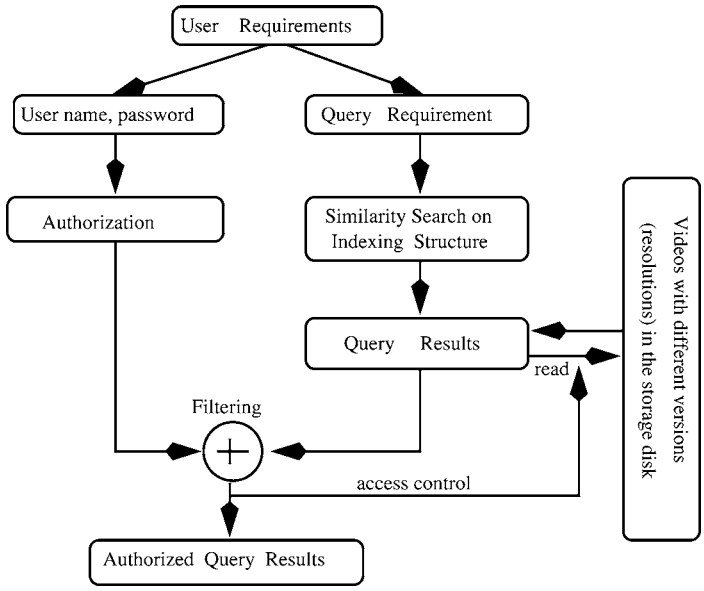


Figure 15. The video access control architecture for an authorized query-by-example procedure.

2. The second component specifies the censored rules, the censored rules describe what kind of the potential users are allowed to access the corresponding video element. This component is used for filtering the results which are obtained by using only the first component for video database accessing as shown in figures 15 and 16. Based on this component, the video database system decides what kind of video elements or what kind of versions of the video elements should be delivered to the corresponding user.
3. The third component can be defined for specifying the mode of video database accessing, e.g., reading or editing. In our current work, we did not include this component, thus only the reading operation is permitted for all users, and only the database manager has the right to edit (e.g., insert and delete) the video elements in database.

6.2. Authorized video access

Different users may have different permissions to obtain different details or layers of the same video element. Therefore, multi-level specification of authorization should be provided for content-based video database accessing control. One can find that our hierarchical video database modeling, representation, and indexing techniques are very suitable for managing the potential multi-level accessing control procedure.

Integrating accessing control into video database systems is achieved by specifying a set of authorization rules and control procedures besides the tradisional *free* and *equal* video accessing procedures. Authorization rules describe *who* is allowed to access *what* in the video database. Therefore, the authorized video database accessing procedures (e.g., querying and browsing) can be partitioned into two parts as shown in figures 15 and 16:

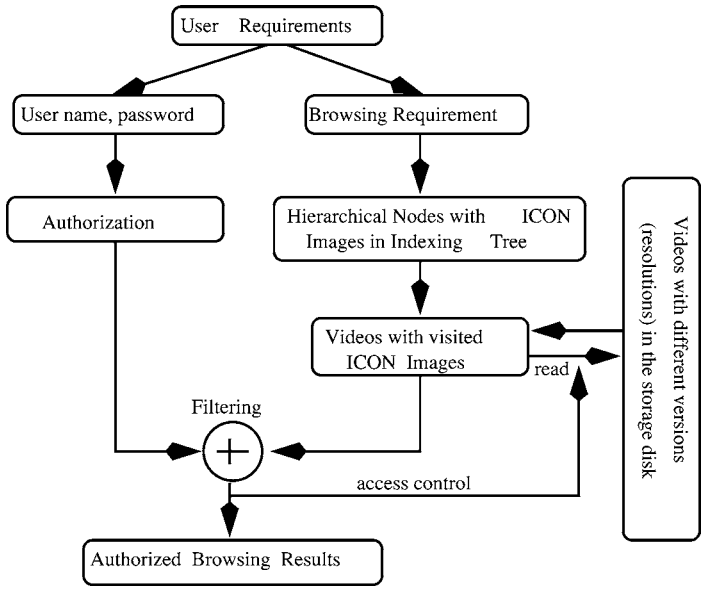


Figure 16. The video access control architecture for an authorized browsing procedure.

authorization and traditional *free* and *equal* video accessing part (querying and browsing). The authorization information is used for controlling what kind of querying or browsing results can be sent to the users according to their authorization. The filtering operator as shown in figures 15 and 16 is used to select the suitable videos with suitable versions according to the user authorization. This often leads to the generation of multiple copies of the same video element stored in the disk and censored at different levels for different classes of users.

7. Conclusions

We have proposed a hierarchical video database model for supporting multi-level video retrieval and browsing. We focus on the use of cluster-based hierarchical indexing structure to both speed-up query-by-example and organize databases for providing more effective browsing. The major contribution of this paper is that a novel cluster-based video indexing structure is provided. Moreover, our works are also very attractive for MPEG-7 applications. Multi-level access control technique can also be supported by our hierarchical video database modeling, representation and indexing structures.

References

1. R. Adams and L. Bischof, "Seeded region growing," IEEE Trans. on PAMI, Vol. 16, pp. 641–647, 1994.
2. A. Alatan et al., "Image sequence analysis for emerging interactive multimedia services—The European COST 211 framework," IEEE Trans. on CSVT, Vol. 8, pp. 802–813, 1998.

3. S. Berchtold, D.A. Keim, and H.P. Kriegel, "The X-tree: An index structure for high-dimensional data," in Proc. of VLDB'96, Bombay, India, 1996, pp. 28–39.
4. S.F. Chang, W. Chen, H.J. Meng, H. Sundaram, and D. Zhong, "A fully automatic content-based video search engine supporting spatiotemporal queries," IEEE Trans. on CSVT, Vol. 8, pp. 602–615, 1998.
5. J.-Y. Chen, C. Taskiran, A. Albiol, E.J. Delp, and C.A. Bouman, "ViBE: A compressed video database structured for active browsing and search," in Proc. SPIE: Multimedia Storage and Archiving Systems IV, Sept. 1999, Boston, Vol. 3846, pp. 148–164.
6. J.D. Courtney, "Automatic video indexing via object motion analysis," Pattern Recognition, Vol. 30, pp. 607–626, 1997.
7. Y. Deng and B.S. Manjunath, "NeTra-V: Toward an object-based video representation," IEEE Trans. on CSVT, Vol. 8, pp. 616–627, 1998.
8. C. Faloutsos, W. Equitz, M. Flickner, W. Niblack, D. Petkovic, and R. Barber, "Efficient and effective querying by image content," Journal of Intelligent Information Systems, Vol. 3, pp. 231–262, 1994.
9. J. Fan, M.S. Hacid, X. Zhang, and A.K. Elmagarmid, "Semantic video object extraction towards content-based indexing," in IASTED Int. Conf. on Internet and Multimedia Systems and Application, Las Vegas, Nov. 19–23, 2000, pp. 430–435.
10. J. Fan, D.K.Y. Yau, W.G. Aref, and A. Rezgui, "Adaptive motion-compensated video coding scheme towards content-based bit rate allocation," Journal of Electronic Imaging, Vol. 9, No. 4, pp. 521–533, 2000.
11. J. Fan et al., "Spatiotemporal segmentation for compact video representation," Signal Processing: Image Communication, Vol. 16, pp. 553–566, 2001.
12. B. Furht, S.W. Smoliar, and H.J. Zhang, Video and Image Processing in Multimedia Systems, Kluwer Academic Publisher, Norwell, MA, 1995.
13. B. Gunsel, A.M. Ferman, and A.M. Tekalp, "Temporal video segmentation using unsupervised clustering and semantic object tracking," J. Electronic Imaging, Vol. 7, pp. 592–604, 1998.
14. A. Guttman, "R-trees: A dynamic index structure for spatial searching," in ACM SIGMOD'84, 1984, pp. 47–57.
15. A. Humrapur, A. Gupta, B. Horowitz, C.F. Shu, C. Fuller, J. Bach, M. Gorkani, and R. Jain, "Virage video engine," in SPIE Proc. Storage and Retrieval for Image and Video Databases V, San Jose, CA, Feb. 1997, pp. 188–197.
16. Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: Querying databases through multiple examples," in Proc. of VLDB'98, 1998.
17. A.K. Jain, A. Vailaya, and X. Wei, "Query by video clip," ACM Multimedia Systems, Vol. 7, pp. 369–384, 1999.
18. K.V.R. Kanth, D. Agrawal, and A. Singh, "Dimensionality reduction for similarity searching in dynamic databases," in ACM SIGMOD, 1998, pp. 166–176.
19. N. Katayama and S. Satoh, "The SR-tree: An index structure for high dimensional nearest neighbor queries," in ACM SIGMOD, 1997.
20. A. Pentland, R.W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," Int. J. Computer Vision, Vol. 18, pp. 233–254, 1996.
21. Y. Rui and T.S. Huang, "A novel relevance feedback technique in image retrieval," in Proc. ACM Multimedia'99, 1999, pp. 67–70.
22. Y. Rui, T.S. Huang, and S. Mehrotra, "Constructing table-of-content for videos," Multimedia Systems, Vol. 7, pp. 359–368, 1999.
23. Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," IEEE Trans. on CSVT, Vol. 8, pp. 644–655, 1998.
24. S. Satoh and T. Kanade, "Name-It: Association of face and name in video," in Proc. of Computer Vision and Pattern Recognition, 1997.
25. G. Sheikholeslami, W. Chang, and A. Zhang, "Semantic clustering and querying on heterogeneous features for visual data," in ACM Multimedia'98, 1998, pp. 3–11.
26. A. Thomasian, V. Castelli, and C.-S. Li, "Clustering and singular value decomposition for approximate indexing in high dimensional space," in CIKM'98, Bethesda, MD, USA, 1998, pp. 201–207.
27. H.J. Zhang, J. Wu, D. Zhong, and S. Smoliar, "An integrated system for content-based video retrieval and browsing," Pattern Recognition, Vol. 30, pp. 643–658, 1997.

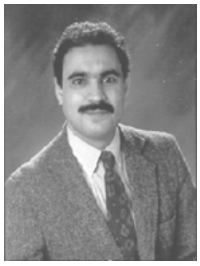
28. D. Zhong, H.J. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," in Proc. SPIE, 1996, pp. 239–246.

Jianping Fan received his MS degree in Theoretical Physics from Northwestern University in 1994, and his PhD in Optical Storage and Computer Science from Shanghai Institute of Optics and Fine mechanics, Chinese Academy of Sciences in 1997. He spent half year in Department of Computer Science, Fudan University at Shanghai as a researcher, and one and half years in Department of Information System Engineering, Osaka University as a JSPS researcher. From 1999 to 2001, he was a researcher at Department of Computer Science, Purdue University, West Lafayette. He is now an assistant professor at Department of Computer Science, University of North Carolina at Charlotte. His research interests include image processing, computer vision, video content computing, indexing and security.

Xingquan Zhu received his Ph.D degree in Computer Science from Fudan University, Shanghai, China in 2001, and B.S, M.S degrees from the Xidian University, Shannxi, China, in 1995 and 1998, respectively. Currently, he is a post-doctoral research assistant in Department of Computer Science, Purdue University, USA. His research interests include image processing, video processing, content based image/video retrieval and video database. He received the SIEMENS and INTEL scholarships in 1999 and 2000 respectively for his Ph.D thesis research-key techniques on content-based video retrieval.

Mohand-Said Hacid graduated as an engineer in Computer Science from the University of Tizi-Ouzou, Algeria, in 1987, and received his PhD degree in computer science from the National Institute of Applied Sciences,

Lyon, France, in 1991. He is currently a professor at the University Claude Bernard Lyon 1, Lyon, France. He has been a visiting researcher at the Theoretical Computer Science Laboratory, Aachen University of Technology, Germany, and at the Indiana Center for Database Systems, Purdue University, Indiana, USA. His research interests include knowledge representation and reasoning, data models and query languages for multimedia databases and semi-structured databases.



Ahmed K. Elmagarmid received a Presidential Young Investigator award from the National Science Foundation, and distinguished alumni awards from Ohio State University and the University of Dayton in 1988, 1993 and 1995 respectively. Professor Elmagarmid is the editor-in-chief of Distributed and Parallel Databases: An International Journal and of the book series on Advances in Database Systems, and serves on the editorial boards of: IEEE Transactions on Knowledge and Data Engineering, Information Sciences, and Journal of Communications Systems. He has served on the editorial boards of IEEE Transactions on Computers and the IEEE Data Engineering Bulletin. He is on the steering committees for the IEEE International Conference on Data Engineering, and the IEEE Symposium on Research Issues in Data Engineering and served on the organization committees of several international conferences. Professor Elmagarmid is the Director of the Indiana Center for Database Systems (ICDS) and the newly formed Indiana Telemedicine Incubator. His research interests are in the areas of video databases, multidatabases, data quality and their applications in Telemedicine and digital government. He is the author of several books in databases and multimedia. He has served widely as an industry consultant and/or adviser to Telcordia, Harris, IBM, MCC, UniSql, MDL, BNR etc. Professor Elmagarmid received his B.S. degree in Computer Science from the Univesrity of Dayton and his M.S. and Ph.D. degrees from The Ohio State University in 1977, 1981 and 1985 respectively. He has served as a faculty member at the Pennsylvania State University from 1985–1988 and has been with the Department of Computer Science at Purdue University since 1988. He is a senior member of the IEEE.