ClassView: Hierarchical Video Shot Classification, Indexing, and Accessing

Jianping Fan, Ahmed K. Elmagarmid, *Senior Member, IEEE*, Xingquan Zhu, Walid G. Aref, *Member, IEEE*, and Lide Wu

***Abstract*—Recent advances in digital video compression and networks have made video more accessible than ever. However, the existing content-based video retrieval systems still suffer from the following problems. 1) *Semantics—sensitive video classification problem* because of the *semantic gap* between low-level visual features and high-level semantic visual concepts; 2) *Integrated video access problem* because of the lack of efficient video database indexing, automatic video annotation, and concept-oriented summary organization techniques. In this paper, we have proposed a novel framework, called *ClassView*, to make some advances toward more efficient video database indexing and access. 1) *A hierarchical semantics-sensitive video classifier* is proposed to shorten the semantic gap. The hierarchical tree structure of the semantics-sensitive video classifier is derived from the domain-dependent concept hierarchy of video contents in a database. Relevance analysis is used for selecting the discriminating visual features with suitable importances. The Expectation-Maximization (EM) algorithm is also used to determine the classification rule for each visual concept node in the classifier. 2) *A hierarchical video database indexing and summary presentation technique* is proposed to support more effective video access over a large-scale database. The hierarchical tree structure of our video database indexing scheme is determined by the domain-dependent concept hierarchy which is also used for video classification. The presentation of visual summary is also integrated with the inherent hierarchical video database indexing tree structure. Integrating video access with efficient database indexing tree structure has provided great opportunity for supporting more powerful video search engines.**

***Index Terms*—Video classification, video database indexing, video retrieval, visual summarization.**

I. Introduction

As a result of decreasing cost of storage devices, increasing network bandwidth capacities, and improved compression techniques, digital video is more accessible than ever. To help users find and retrieve relevant information effectively and facilitate new and better ways of entertainment, advanced technologies need to be developed for indexing, browsing, filtering, searching, and updating the vast amount of information available in video databases [1], [2]. The recent development of content-based video retrieval systems has advanced our capabilities for searching videos via color, layout, texture, motion, and shape features [3]–[11]. However, these content-based video retrieval systems still suffer from the following challenging problems.

- **Semantics-sensitive video classification problem:** When very large video data set comes into view, efficient video database indexing can no longer be ignored [12]. However, the traditional database indexing trees [13]–[18], such as R-tree, SR-tree, and SS-tree, are unsuitable for video database indexing and management because of the *curse of dimensionality* [49]. Video retrieval can be performed in an efficient way by classifying the similar videos into the same cluster [10], [11]. Unfortunately, there is a *semantic gap* between low-level visual features and high-level semantic visual concepts [22]–[32]. The traditional pure feature-based data clustering techniques are unsuitable for video classification because of the semantic gap [19]–[21]. Decision tree classifier is also widely used for supervised data classification [33]–[35], but it may consist of too many internal nodes which are consequently very difficult to comprehend and interpret. Even after pruning, the decision tree structures induced by the existing machine learning algorithms can be extremely complex and the constructed tree structures do not make sense for video database indexing. Therefore, the semantics-sensitive video classifier is expected not only to shorten the semantic gap but also to provide an effective scheme for video database indexing, automatic video annotation, and concept-oriented summary presentation. In this paper, we propose an efficient hierarchical semantics-sensitive video classifier and its hierarchical tree structure is derived from the domain-dependent concept hierarchy of video contents.
- **Integrated video access problem:** There are three widely-accepted but independent approaches to access the video in a database, as follows.
  - *Query-by-example* is widely used in the existing video retrieval systems. Query-by-example is necessary in a situation where naive users cannot clearly describe what they want via keywords or they do not want to search the large-scale video database via hierarchical summary browsing. However, the query-by-example approach suffers from at least two problems. The first one is that not all database users have example video

Manuscript received March 1, 2001; revised July 29, 2002. This work was supported by National Science Foundation under Grants 0208539-IIS, 9972883-EIA, 9974255-IIS, and 9983249-EIA, and by grants from HP, IBM, Intel, NCR, Telcordia and CERIAS. L. Wu was supported by the National Science Foundation of China under Contract 69935010. Parts of this work were performed while J. Fan was with the Department of Computer Science, Purdue University. The associate editor coordinating the review of this paper and approving it for publication was Prof. Wayne Wolf.

J. Fan is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: jfan@uncc.edu).

A. K. Elmagarmid, X. Zhu, and W. G. Aref are with the Department of Computer Science, Purdue University, West Lafayette, IN 47907 USA.

L. Wu is with the Department of Computer Science, Fudan University, Shanghai 200433, China.

Digital Object Identifier 10.1109/TMM.2003.819583

clips at hand. Even the video database system interface can provide some templates of video clips, there is still a gap between the various requirements of different users and the limited templates that can be provided by the database interface. The second one is that the naive users may prefer to query the video database via the high-level semantic visual concepts or hierarchical summary browsing through the concept hierarchy of video contents. The major difficulty for the existing video retrieval systems is that they are unable to let users query video via the high-level semantic visual concepts and enable concept-oriented hierarchical video database browsing [12].

– *Query-by-keywords* is also used in some content-based video retrieval systems based on manual text annotation [3], [6]. The keywords, which are used for describing and indexing the videos in the database, are subjectively added by database constructionist without a well-defined structure. Since the keywords used for video indexing are subjective, the naive users cannot find exactly what they want because they may not be so lucky to use the same keyword as the database constructionist did. Moreover, manual text annotation is too expensive for large-scale video collections.

  In order to avoid the problem of subjective and expensive manual video annotation in our system, the videos are annotated automatically via the well-defined and widely-accepted concept hierarchy, where only the keywords used for constructing and interpreting the domain-dependent concept hierarchy are selected for video annotation. An efficient high-level video retrieval technique is provided by using the keywords which are used for interpreting the high-level semantic visual concepts. We call this, integrating the query-by-example with the query-by-keywords, as the ***integrated video access problem 1***.

– *Hierarchical browsing* is also widely accepted by the naive internet users to access text document via *Yahoo*, *Google* search engines. The naive users should be interested in hierarchical browsing the summaries that are presented on different visual concept levels, rather than having to use visual features or keywords to describe their requests. However, most existing video retrieval systems do not support concept-oriented hierarchical video database browsing because of the lack of efficient video summary presentation structure [12]. In order to support video browsing, some pioneer works have been proposed in the past [36]–[39]. However, these existing techniques just focus on browsing a video sequence and they did not address how to support the concept-oriented hierarchical video database browsing [40].

  A key issue to the concept-oriented hierarchical video database browsing is whether the visual summaries found make sense to the naive users and how to interpret the contextual and logical relationships of the visual summaries on different visual concept levels. We call this, integrating the concept-oriented hierarchical video database browsing with the inherent database indexing structure, as the ***integrated video access problem 2***.

Based on above observations, we propose a novel framework, called ***ClassView***, to make some advances in overcoming these problems. This paper is organized as follows. Section II proposes a semantics-sensitive video database model which can support more effective video database indexing and access. A novel semantics-sensitive video classification technique is proposed in Section III. Section IV presents a hierarchical indexing structure and an efficient video access procedure. We have also provided the performance analysis of our techniques in the related Sections. We conclude in Section V.

II. Semantics-Sensitive Video Database Model

Several high-dimensional database indexing trees have been proposed in the past and they are expected to be used for video database indexing [13]–[18], but they suffer from the problem of *curse of dimensionality* because the visual features used for video representation and indexing are normally in high-dimensions [49]. One reasonable solution is first to classify videos into a set of clusters and then to perform the dimension reduction on these clusters independently [41], [42], the traditional database indexing trees can supposedly be used for indexing these video clusters independently with relatively low-dimensional features. However, the pure feature-based clustering techniques are unsuitable for video classification because of the semantic gap [22]–[32]. Decision tree classifier is very attractive for video classification via learning from the labeled training examples [33]–[35], but its internal nodes do not make sense for video database indexing. The semantics-sensitive video classifier is expected not only to be efficient for bridging the semantic gap but also to provide an effective video database indexing scheme, thus the tree structure of the semantic video classifier should be related to the concept hierarchy of video contents. Unfortunately, the ***video database model*** problem has not been addressed efficiently for supporting video access over the large-scale database [1], [2].

There are two widely-accepted approaches to characterize video in the database: *shot-based* and *object-based*. In this paper, we focus on the shot-based approach because video shots are good choice as the basic unit for video content indexing [36]–[38]. In order to support more efficient video database management, we classify video shots into a set of hierarchical database management units as shown in Fig. 1. In order to achieve hierarchical video database management, we need to address the following key problems. 1) How many ***levels*** should be included in this video database model and how many ***nodes*** should be included on each database level? 2) Do these nodes in the classifier make sense to human beings? 3) What kind of ***discriminating features*** should be selected and what kind of ***classification rule*** should be used for each visual concept node?

We solve the first and second problems by deriving the database model from the concept hierarchy of video contents. Obviously, the concept hierarchy is domain-dependent and a *video*

news example is given in Fig. 2. This domain-dependent concept hierarchy defines the contextual and logical relationships between the higher level visual concepts and the relevant lower level visual concepts. ***ClassView*** has provided the following techniques to achieve this novel framework.

- ***A semantics-sensitive video classifier*** to shorten the semantic gap between the low-level visual features and the high-level semantic visual concepts. The hierarchical tree structure of our semantics-sensitive video classifier is derived from the domain-dependent concept hierarchy of video contents and is provided by domain experts or obtained by using WordNet [43], [44]. Each visual concept node in this classifier defines a specific semantic visual concept which makes sense to human beings, the contextual and logical relationships between the higher level visual concept nodes and their relevant sublevel visual concept nodes are defined by the domain-dependent concept hierarchy. Relevance analysis is integrated with the Expectation-Maximization (EM) algorithm to determine the discriminating features and classification rule for each visual concept node by learning from the labeled training examples.

  Since different visual features capture different aspects of perception of visual concepts, the video database management units (i.e., visual concept nodes) as shown in Fig. 1 should be characterized and indexed by their discriminating features with different significances. The ***basic assumption*** of our work is that the semantically similar video shots should be close to each other in their warped subspace defined by their discriminating features, even though they may be far from each other in their original feature space. Note that the goal of semantics-sensitive video classification is not to understand videos the way human beings do, but to classify the unlabeled video clips to the known semantic visual concepts defined by the domain-dependent concept hierarchy so that more efficient video database indexing and access can be supported [26]. After the classification, the unlabeled video shots inherit the semantic labels assigned for the visual concept nodes they belong to, thus automatic video annotation is supported by using the widely-accepted keywords (i.e., keywords used for constructing and interpreting the domain-dependent concept hierarchy) with a well-defined structure (i.e., concept hierarchy).

  The hierarchical tree structure of our semantics-sensitive video classifier, which is determined by the domain-dependent concept hierarchy, is also used as the database indexing tree structure for supporting hierarchical video database management. For each visual concept node of the proposed hierarchical video database indexing tree, we use Gaussian functions to approximate the distribution of its video shots in their warped feature subspace with a certain degree of accuracy. We use the following parameters to represent and index the visual concept node R, shown at the bottom of the page, where L_R is its semantic label for the visual concept node R (inherited from the keyword-based interpretation of the concept hierarchy), Ξ_R is the subset of the discriminating features for the corresponding visual concept node R, D_R is the number of its discriminating features, $\theta_{\mathbf{R}}$ indicates the weights associated with these discriminating features, μ_R and σ_R will be used to approximate the Gaussian distribution $\rho(X, \mu_R, \sigma_R)$ of the video shots with the feature values X assigned to the corresponding visual concept node R, with μ_R and σ_R being the mean and the variance respectively. The node seeds, $\{ ST_1, \ldots, ST_m \}$, which are the principal video shots for the corresponding visual concept node R, are used for representing the high-level semantic visual concept because of the lack of featural support for the higher level visual concepts.
- ***A hierarchical database indexing and summary presentation technique*** to support more effective video access. The video database indexing and management structure is inherently provided by the domain-dependent concept hierarchy which is also used for video classification. The organization of visual summaries is also integrated with the inherent hierarchical database indexing tree structure, thus the concept-oriented hierarchical video database browsing can be supported.

III. Semantics-Sensitive Video Classifier

Video analysis and feature extraction are necessary steps for supporting hierarchical semantics-sensitive video classification [45]. In our approach, a MPEG video sequence is first partitioned into a set of video shots by using our automatic video shot detection technique. In general, threshold setting plays a critical role in automatic video shot detection [46]. The thresholds for shot detection should be adapted to the activities of video contents. It is impossible to use a universal threshold that can satisfy various conditions because the thresholds for different video sequences or even different video shots within the same sequence should be different. Our previous technique can adapt the thresholds for video shot detection according to the activities of various video sequences [46], but it cannot adapt the thresholds for different video shots within the same sequence.

In order to adapt the thresholds to the ***local activities*** of different video shots within the same sequence, we use a small

$$\begin{aligned}
\text{semantic label} &: L_R, \text{subset for discriminating features} : \Xi_R \\
\text{dimensions} &: D_R, \text{feature weights} : \theta_{\mathbf{R}} = (\theta_{\mathbf{1}}, \ldots, \theta_{\mathbf{D_R}}) \\
\text{Gaussian parameters} &: \text{mean } \mu_{\mathbf{R}} = (\mu_{\mathbf{1}}, \ldots, \mu_{\mathbf{D_R}}), \text{variance } \sigma_{\mathbf{R}} = (\sigma_{\mathbf{1}}, \ldots, \sigma_{\mathbf{D_R}}) \\
\text{node seeds} &: \{ST_1, \ldots, ST_m\}
\end{aligned}$$

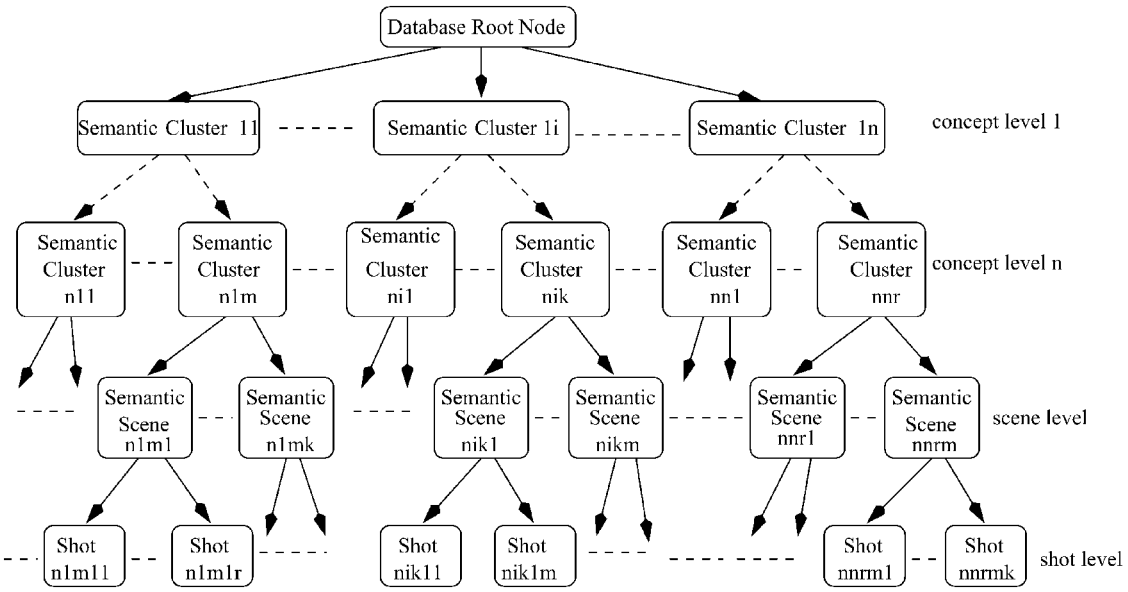

Fig. 1. Proposed hierarchical video database model, where the cluster may include several levels according to the concept hierarchy and a video scene basically consists of a sequence of connected or unconnected shots.

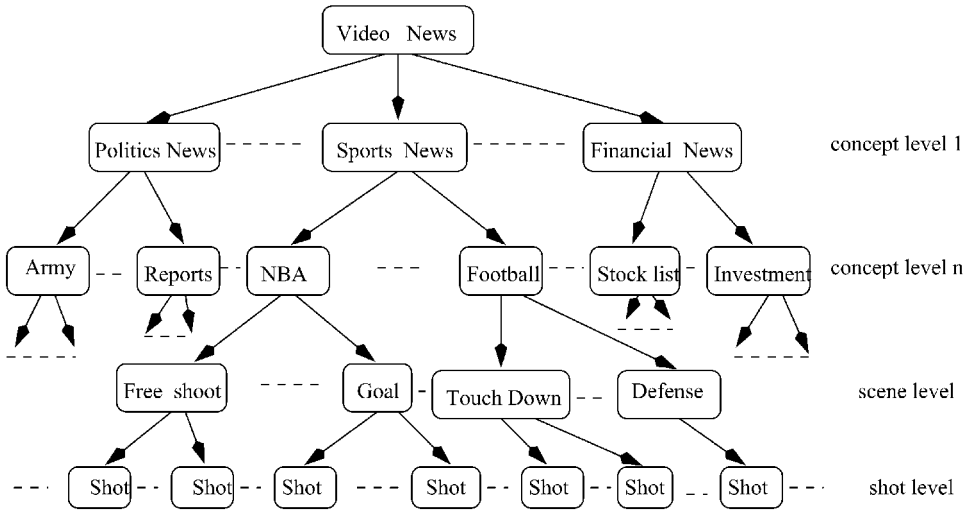

Fig. 2. Concept hierarchy for *video news* used in our system.

window (i.e., 20 frames in our current work) and the threshold for each window is adapted to its local visual activity. The video shot detection results shown in Figs. 3–6 are obtained from several video data sources used in our system, such as *movies*, *video news* and *medical videos*. Each video shot is then processed to extract a set of visual features such as 256-dimensional HSV color histogram (i.e., 16 components for H, four components for S and four components for V), 32-dimensional texture feature via directional edge histogram (i.e., 32 directions), Tamura texture features, nine-dimensional directional motion histogram, and camera motions.

In this paper, we focus on generating semantic video scenes and upper-level visual concepts such as clusters from these obtained video shots, so that more efficient database management structure can be supported. The semantic video classifier is built in a bottom-up fashion as shown in Fig. 7. As mentioned in Section II, the hierarchical tree structure of the classifier, i.e., ***levels*** and ***nodes***, is first determined according to the domain-dependent concept hierarchy of video contents and is given by the domain experts or obtained via WordNet [43], [44]. Once such hierarchical video classification structure is given, we use a set of labeled training examples to determine the discriminating features (i.e., feature subspace) and classification rule for each visual concept node via relevance analysis. For each visual concept node, a labeled training example is in terms of a set of shot-based low-level visual features $\Xi = \{F_l|_{l=1}^{n}\}$ and the semantic label L provided by the domain experts or the naive users. There are two measurements for defining the similarity among the labeled training video shots under the given visual concept node, as follows.

- ***Visual similarity*** via comparing their shot-based low-level visual features.
- ***Semantic similarity*** via comparing their high-level semantic labels (i.e., visual concepts).

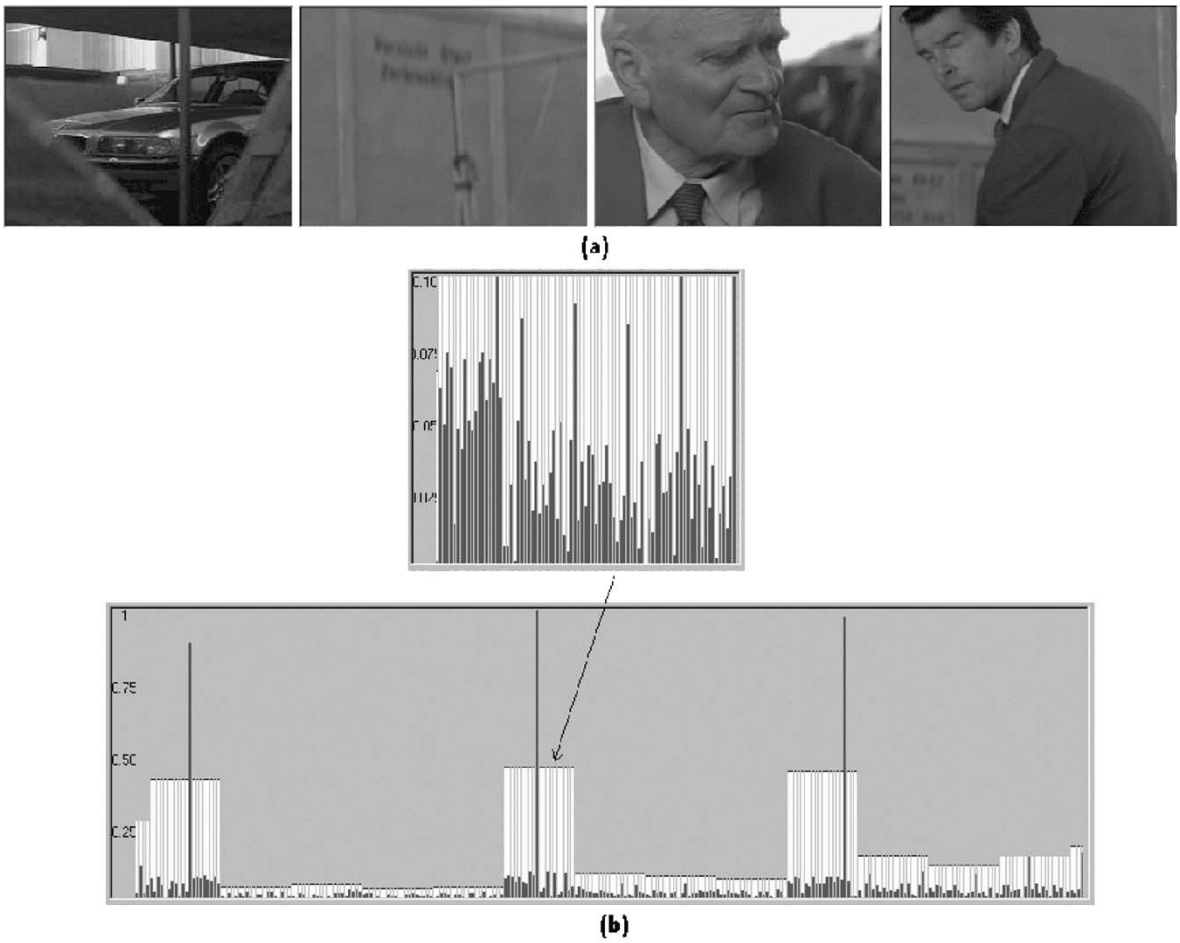

Fig. 3. Video shot detection results from a movie: (a) part of the detected scene cut frames; (b) the corresponding color histogram difference and the determined thresholds for different video shots, where the values of color histogram difference in a small window is also given.

The feature-based similarity distance $D_F(T_\delta, T_\gamma)$ between two video shots T_δ and T_γ is defined as

$$D_F(T_\delta, T_\gamma) = \sum_{F_l \in \Xi} \frac{1}{\alpha_l} \cdot D_{F_l}(T_\delta, T_\gamma), \quad \sum_{l=1}^{n} \frac{1}{\alpha_l} = 1 \tag{1}$$

where $D_{F_l}(T_\delta, T_\gamma)$ denotes the similarity distance between T_δ and T_γ according to their lth visual feature F_l, α_l is the weight for the lth visual feature, Ξ is the set of original visual features, and n is the total number of visual features as described above which are initially extracted for video shot representation.

The concept-based semantic similarity distance $D_S(T_\delta, T_\gamma)$ between two video shots T_δ and T_γ can be defined as

$$D_S(T_\delta, T_\gamma) = \begin{cases} 0, & L_\delta = L_\gamma \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

where L_δ and L_γ is the semantic labels for the video shots T_δ and T_γ. There are only two possibilities for the concept-based semantic similarity between two labeled video shots under the given semantic label for the corresponding visual concept node: *similar* versus *dissimilar*.

Our hierarchical semantics-sensitive video classifier focuses on bridging the gap between these two similarity measurements, so that the feature-based visual similarity can correspond to the concept-based semantic similarity by selecting the discriminating features with suitable importance. We have integrated relevance analysis with the EM algorithm to shorten the semantic gap and determine the feature subspace and classification rule for each visual concept node.

A. Semantic Visual Concept Generation

We now describe how to build this hierarchical semantics-sensitive video classifier. For the first classification (from video shots to semantic video scenes), we first use a set of labeled training examples to select the discriminating features and their importances and determine the classification rule for each visual concept node at the scene level. These labeled training examples can be partitioned into four groups according to their feature-based visual similarity and concept-based semantic similarity.

- *Type one positive examples* which are both semantically and visually similar.
- *Type two positive examples* which are both semantically and visually dissimilar.
- *Type one negative examples* which are visually similar but semantically dissimilar.
- *Type two negative examples* which are semantically similar but visually dissimilar.

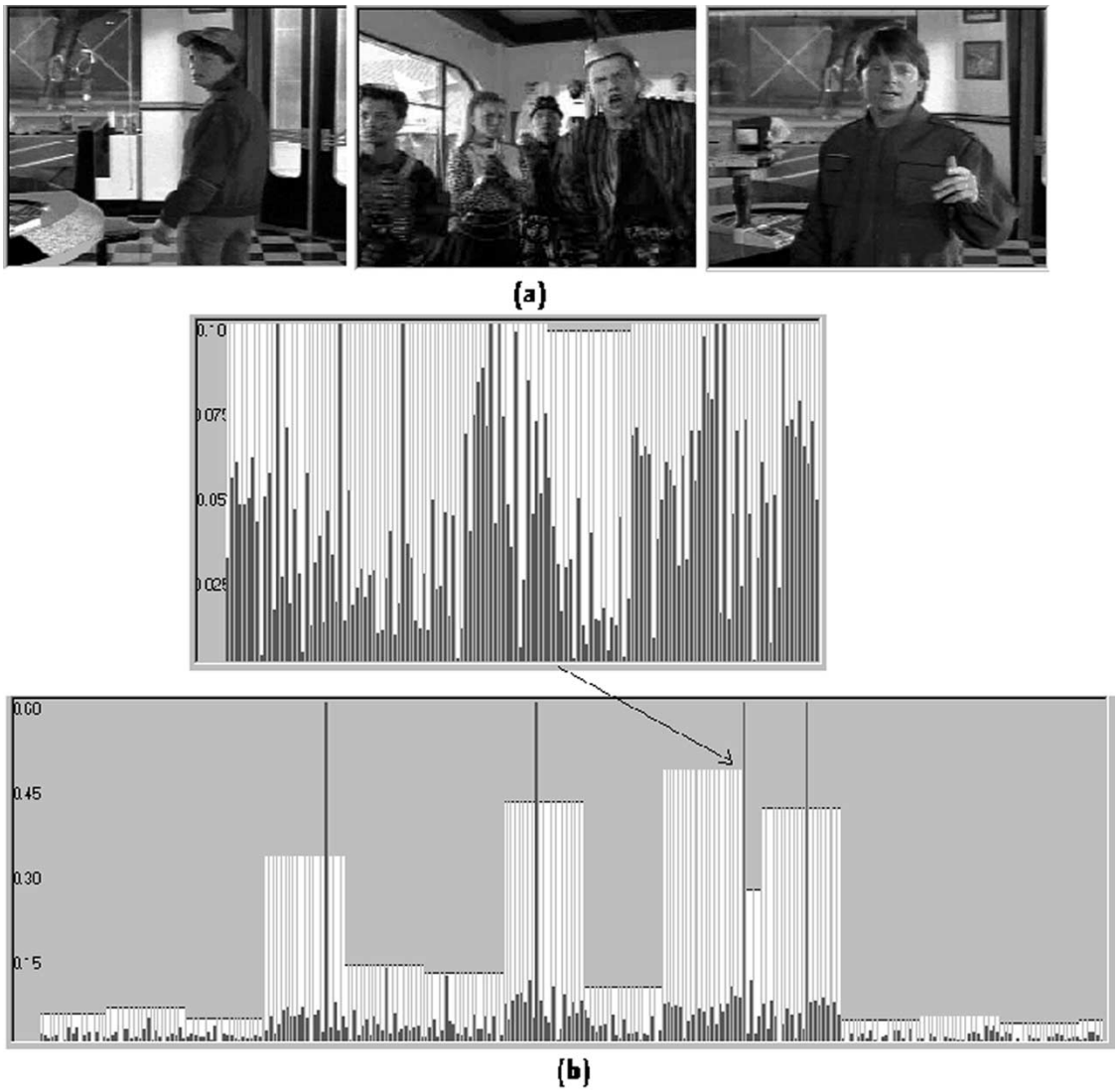

Fig. 4. Video shot detection results from a movie: (a) part of the detected scene cut frames; (b) the corresponding color histogram difference and the determined thresholds for different video shots, where the values of color histogram difference in a small window is also given.

Type One Negative Examples:

$$\max\left\{D_F(T_\delta,T_\gamma)=\sum_{F_j}\frac{1}{\alpha_j}\cdot D_{F_j}(T_\delta,T_\gamma)\right\},\quad F_j\in\Xi_1 \tag{3}$$

where Ξ_1 indicates the subset for the discriminating features which make the type one negative examples far from each other, m_1 is the total number of the discriminating features, $m_1 \ll n$, and α_j is the weight for the jth dimensional feature F_j.

Type Two Negative Examples:

$$\min\left\{D_F(T_\delta,T_\gamma)=\sum_{F_i}\frac{1}{\alpha_i}\cdot D_{F_i}(T_\delta,T_\gamma)\right\},\quad F_i\in\Xi_2 \tag{4}$$

where Ξ_2 indicates the subset for the discriminating features which make the type two negative examples close to each other, m_2 is the total number of the discriminating features, $m_2 \ll n$, and α_i is the weight for the ith feature F_i.

Type One Positive Examples:

$$\min\left\{D_F(T_\delta,T_\gamma)=\sum_{F_h}\frac{1}{\alpha_h}\cdot D_{F_h}(T_\delta,T_\gamma)\right\},\quad F_h\in\Xi_1\cup\Xi_2 \tag{5}$$

where $\Xi_1 \cup \Xi_2$ indicates the subset for the discriminating features for the type one positive examples and $\Xi_1 \cap \Xi_2 = \emptyset$, and $m_1 + m_2$ is the total number of these discriminating features, $m_1 + m_2 < n$.

Type Two Positive Examples:

$$\max\left\{D_F(T_\delta,T_\gamma)=\sum_{F_k}\frac{1}{\alpha_k}\cdot D_{F_k}(T_\delta,T_\gamma)\right\},\quad F_k\in\Xi-(\Xi_1\cup\Xi_2) \tag{6}$$

where $\Xi-(\Xi_1\cup\Xi_2)$ indicates the subset for the discriminating features for the type two positive examples, $n-(m_1+m_2)$ is the total number of these discriminating features, and α_k is the weight of the kth feature F_k.

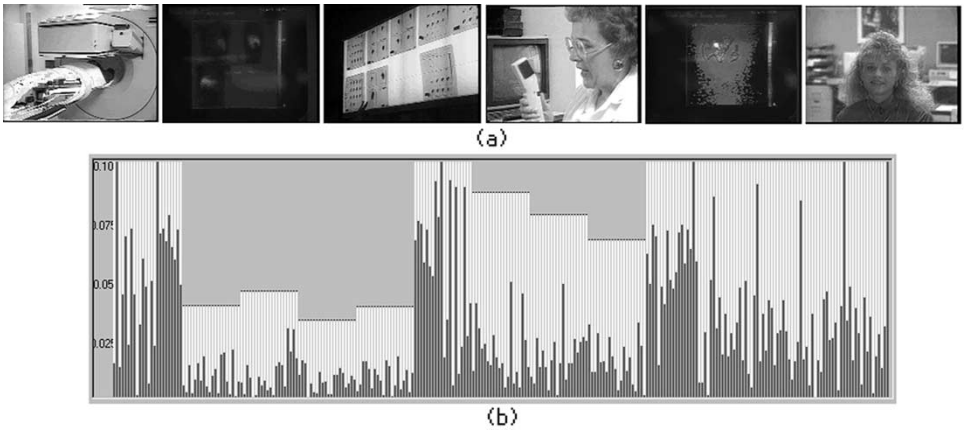

Fig. 5. Video shot detection results from a medical video: (a) part of the detected scene cut frames; (b) the corresponding color histogram difference and the determined thresholds for different video shots.

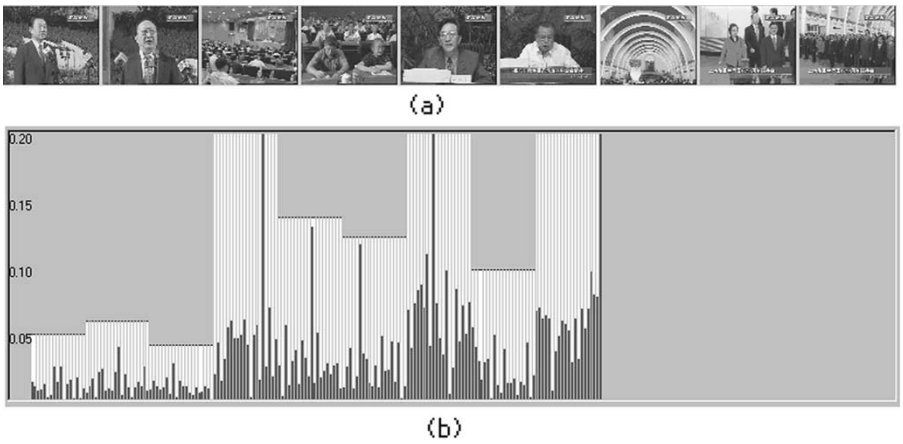

Fig. 6. Video shot detection results from a video news: (a) part of the detected scene cut frames; (b) the corresponding color histogram difference and the determined thresholds for different video shots.

The type one negative examples have small average similarity distances $D_F(T_\delta, T_\gamma)$ according to their original visual features, but they should have big dimensional similarity distances according to some discriminating features (i.e., $D_{F_l}(T_\delta, T_\gamma)$) because they are semantically dissimilar. The visual features, which indicate these type one negative examples to be far from each other, should be selected as the discriminating features with high importances for the corresponding visual concept node.

The type two negative examples have big average visual similarity distances $D_F(T_\delta, T_\gamma)$ according to their original visual features, but they should have small dimensional similarity distances according to some discriminating features (i.e., $D_{F_l}(T_\delta, T_\gamma)$) because they are semantically similar. The visual features, which indicate these type two negative examples to be close to each other, should be selected as the discriminating features with high importances for the corresponding visual concept node.

Based on above observations, the visual features, which indicate the type one negative examples to be far from each other and the type two negative examples to be close to each other, should be selected as the discriminating features with high importance for video shot representation and classification for the corresponding visual concept node.

Instead of searching the weights α from the high-dimensional original feature space (i.e., Ξ) [28], [29], we first use decision tree to obtain the feature subsets, Ξ_1 and Ξ_2, for the corresponding visual concept node [33]–[35]. Determining the feature subsets first via decision tree has reduced the search burden of relevance analysis dramatically. The Lagrangian optimization technique is then used for obtaining the dimensional weights for these discriminating visual features [28], [29]:

$$\min\{D_F(T_\delta, T_\gamma) = \sum_{Fi} \frac{1}{\alpha_l} \cdot D_{Fi}(T_\delta, T_\gamma)$$
$$\text{subject to } \sum_{l=1}^{n} \frac{1}{\alpha_l} = 1, \quad F_l \in \Xi_1 \cup \Xi_2. \tag{7}$$

We now describe how to learn the classification rules for the visual concept nodes at the scene level in the classifier. The posterior probability $P(S|X, \alpha)$ of a video shot T with the feature values X being in a video scene S can be computed via Bayes law [47]:

$$P(S|X, \alpha) = \frac{P(X|S, \alpha)P(S)}{P(X|S, \alpha)P(S) + P(X|\bar{S}, \alpha)P(\bar{S})} \tag{8}$$

where $P(X|S, \alpha)$ and $P(X|\bar{S}, \alpha)$ indicate the conditional probabilities for the presence and absence of the video scene S, $P(S)$

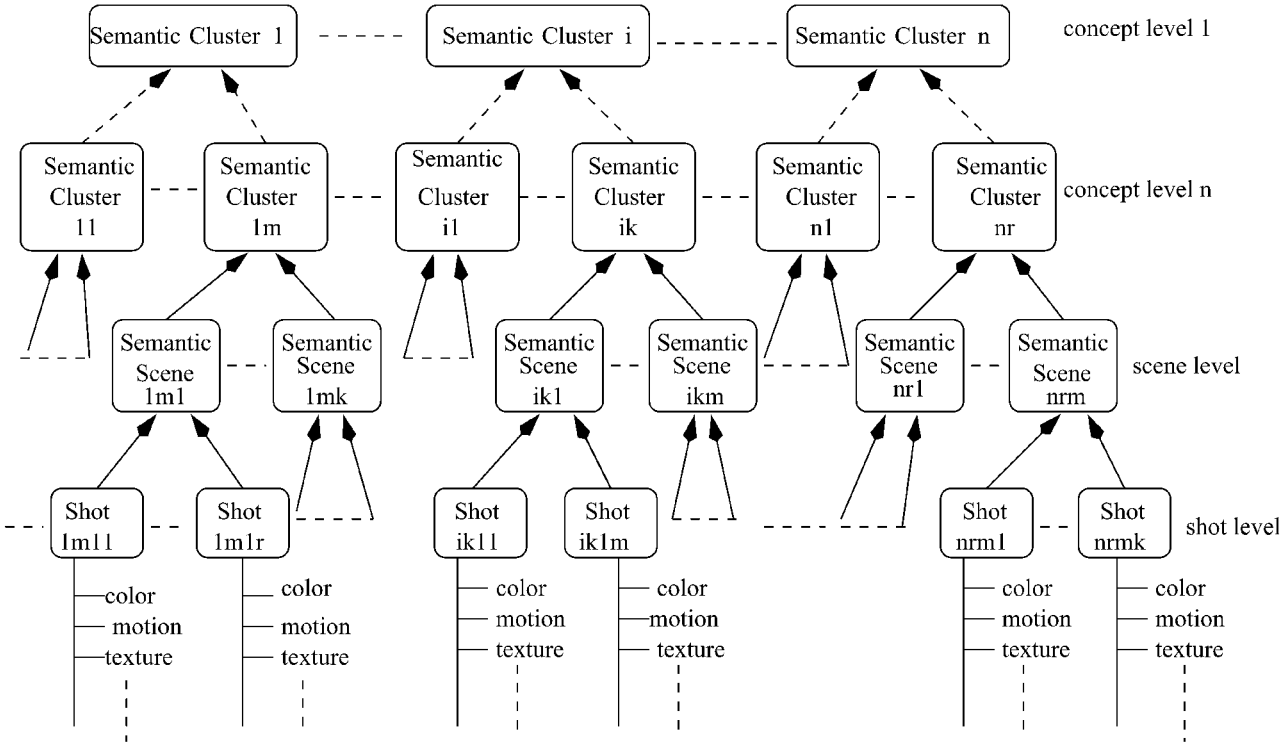

Fig. 7. Bottom-up procedure for building the hierarchical video classifier, where the semantic cluster may include multiple levels according to the concept hierarchy.

and $P(\bar{S})$ represent the probabilities for the presence and absence of the video scene S in the training data set, andα is the feature weights for the corresponding visual concept node.

The prior probability $P(S)$ of the video scene S can be easily estimated from the labeled training data set by dividing N_S, the number of instances in the video scene S, by the total number of instances N in the training data set, $P(S) = N_s/N$. The conditional density $P(X|S, \alpha)$ for each video scene is modeled as a Gaussian distribution or be obtained from a large training data set by using histogram approximation [47]. The classification rule for the video scene node S is determined by maximizing the function

$$E_S = \sum_i \sum_s P(X_i|S_s, \alpha) \log P(S_s|X_i, \alpha). \tag{9}$$

The second level classification (from semantic scene to semantic clusters) is also modeled by a set of probabilities:

$$P(C|X, \beta) = \frac{P(X|C, \beta)P(C)}{P(X|C, \beta)P(C) + P(X|\overline{C}, \beta)P(\overline{C})} \tag{10}$$

where $P(X|C, \beta)$ and $P(X|\overline{C}, \beta)$ indicate the conditional probabilities for the presence and absence of the cluster C, $P(C)$ and $P(\overline{C})$ represent the probabilities for the presence and absence of the cluster C in the training data set, and β is the feature weights for the corresponding visual concept node at the cluster level. The classification rule for the cluster node C is determined by maximizing the function

$$E_C = \sum_i \sum_p P(X_i|C_p, \beta) \log P(C_p|X_i, \beta). \tag{11}$$

The semantic cluster may consist of several levels according to the concept hierarchy of video contents, we just discuss one level here as an example because the classification rules for the other levels can also be obtained via a similar approach.

After the hierarchical semantics-sensitive video classifier is generated, an ***interactive pruning*** step is performed to simplify the classifier so that all these visual concept nodes are meaningful to the human users. For each visual concept node of the proposed hierarchical video classifier, Gaussian functions are used to approximate the distribution of its video shots in its warped feature subspace (see Section II). The mean and variances of the distribution of video shots, feature subspace (i.e., discriminating features) and their weights, semantic label, and node seeds are selected for visual concept node representation and indexing.

B. Video Shot Classification

Once the hierarchical semantics-sensitive video classifier is in place, it is used to semantically classify the unlabeled videos. The video shots are first extracted from the unlabeled videos by using our automatic video shot detection technique. The task of video shot classification can then be summarized as follows. Given an unlabeled video shot T (obtained from the unlabeled video) and its n-dimensional feature values X, it is first assigned to the best matching semantic video scene that corresponds to the maximum probability, and subsequently to the best matching semantic cluster, in a bottom-up fashion according to the domain-dependent concept hierarchy. Along with each step of classifying a new video shot into a specific visual concept node, the means and variances for the corresponding visual concept node is also updated by involving this new video shot.

Fig. 8. Semantic scenes generated from a *video news*: (a) reports; (b) army; (c) reports; (d) sports news.

Fig. 9. Semantic scene of *dialog* generated from medical video.

It is important to note that once an unlabeled video shot is classified, the semantic labels for the corresponding visual concept nodes that it is assigned to becomes its semantic labes; therefore the corresponding unlabeled video shot is automatically associated with a hierarchy of semantic labels. Such semantic labels (keywords used for constructing and interpreting the concept hierarchy) make it possible for video query via semantic visual concepts (see Section IV-A).

High-level semantic visual concept nodes (i.e., cluster levels) may have lower featural support because the variances of the shot-based low-level visual features may be very large for these semantically similar video shots residing in. In order to avoid this problem, we have identified a number of **principal video shots** (i.e., concept abstract) for each high-level visual concept node and these principal video shots can also be taken as the node seeds for video classification and access. The basic requirement of seed selection is that the selected principal video shots (i.e., seeds) for the higher level visual concept node should convey all the potential visual concepts in its relevant sublevel visual concept nodes. We currently select these principal video shots for each visual concept node via a semi-automatic approach, full-automatic techniques are expected in the future.

To answer query-by-example, the naive users can select to achieve similarity video search by using the average properties of clusters (i.e., Gaussian distribution) or the node seeds (i.e., principal video shots assigned to the high-level visual concept nodes).

C. Performance Analysis

In order to evaluate the real performance of our hierarchical semantics-sensitive video classification technique, we selected three testing video sources: *video news*, *movies*, and *medical videos*. The video shots are first detected automatically by using our adaptive shot detection technique. A set of low-level visual features as described above are then extracted for video shot representation and indexing. In order to support video access via the high-level visual concepts in our systems, we first select a training example set and the domain experts are involved to annotate these training examples manually according to the domain-dependent concept hierarchy. Currently, part of these training examples are also annotated by the naive users. After the manual annotation (i.e., high-level visual concepts from human point of view) and the low-level visual features are available for these training examples, our hierarchical semantics-sensitive video classification technique is then performed to obtain the classification rule, feature subspace and dimensional weights for each high-level visual concept node in the classifier.

After the hierarchical semantics-sensitive video classifier is obtained, we then obatin its performance based on three testing video sources. The average performance of our hierarchical semantics-sensitive video classifier shown in Table I is obtained from three video sources: *movies*, *video news*, and *medical videos*. The video scene generation results from two sources are shown in Figs. 8 and 9.

For one-level and two-state image classification techniques [26], [30], their classification accuracy can be achieved higher than 90%. As compared with these traditional one-level and

TABLE I
THE AVERAGE PERFORMANCE OF OUR SEMANTICS-SENSITIVE VIDEO CLASSIFIER

Test Data Types	Test Data Numbers (shots)	Concept Levels	Correction Ratio
Medical Videos	1200	3	60.2%
News Videos	1100	3	56.3%
Movie Videos	3500	4	47.6%

two-state semantic image classification techniques, one can find that the accuracy ratio for our hierarchical semantics-sensitive video classifier is not perfect as we expected. The reasons are as follows.

a) The relevance analysis has been integrated with the EM algorithm to bridge the semantic gap by exploiting the statistical properties of the semantically similar video shots, but there is a semantic gap between the low-level visual features and the high-level visual concepts. Thus the statistical properties of these shot-based low-level visual features are too general to characterize the relevant semantic visual concepts. On the other hand, representing the high-level visual concepts by using the limited node seeds cannot effectively convey the statistical properties of the semantically similar video shots residing in the same visual concept node.

b) It is not the best solution to use the feature weighting technique to bridge the semantic gap, especially when the shot-based low-level visual features are unsuitable for characterizing the relevant visual concepts. The concept-based semantic similarity among these training examples is too label-intensive because the manual labels may be defined by the domain experts or the naive users just according to the semantic categorizations instead of the visual perceptions.

c) The accuracy of our semantic video classifier also depends on the distribution of the training example set and some selected training examples may be irrelevant to the corresponding visual concept, thus result in poor performance.

d) As shown in Fig. 10, the performance of our video shot classifier depends the size of the training data set, a large training data set often increases the accuracy of the classification as shown in Fig. 10. However, it is very expensive for obtaining the large-scale training example set. The limited size of the training data set for each specific visual concept node depends on the dimensions of its discriminating features and the number of its relevant sublevel visual concept nodes.

e) Our hierarchical semantics-sensitive video classifier focuses on addressing the video classification problem with multiple levels (i.e., concept hierarchy) and multiple states (i.e., each visual concept cluster consists of multiple sublevel clusters), thus the variances of the shot-based low-level visual features for the semantically similar video shots should be very large and thus result in poor performance. One simple but reasonable solution is to treat the hierarchical video classifier as a set of independent one-level and two-state classifiers, thus each semantic visual concept in our hierarchical semantics-sensitive video classifier will be generated by a specific one-level and two-state classifier, but the relationships among these visual concepts on the same level will be lost and thus the generated visual concepts may have heavy overlaps in their low-level feature spaces.

f) The single video shot may consist of multiple semantic visual concepts and induces very different subjective interpretations, thus the concept-based semantic similarity between the labeled video shots suffers from the subjectivity problem.

g) As defined in (8) and (10), the conditional probabilities for the absences of the corresponding visual concepts are defined as the joint probabilities $P(X|\bar{S},\alpha)$ and $P(X|\bar{C},\beta)$. However, the irrelevant video shots (i.e., indicating the absence of the corresponding visual concept) may consist of multiple visual concepts, thus they will not follow the same joint probability and mixture probabilities should be used.

IV. HIERARCHICAL DATABASE INDEXING AND ACCESSING

Once video shots are organized via our hierarchical semantics-sensitive video classifier, the next issue is how to provide more efficient and effective video database indexing through this structure, so that fast video retrieval and browsing can be supported in our system. Recall that each visual concept node in our hierarchical video classifier consists of a set of relevant sublevel visual concept nodes, and all these visual concept nodes are further associated with a set of video shots and their distributions in the low-level visual feature space.

The distributions of video shots from different semantic clusters may overlap in their original feature space. The distribution of video shots from the same semantic cluster may have large variance in the original feature space. If the traditional database indexing structures are used for representing these high-level visual concept nodes via hyper-rectangular or even hyper-sphere boxes, they suffer from the problem of *curse of dimensionality* because of heavy overlap of the distributions of video shots from different visual concept nodes at the same level. In order to make the semantically similar video shots be close to each other and make the semantically dissimilar video shots be far from each other, we use the discriminating features with different importances for each visual concept node. The distributions of video shots for different visual concept nodes at the same database level are isolated as shown in

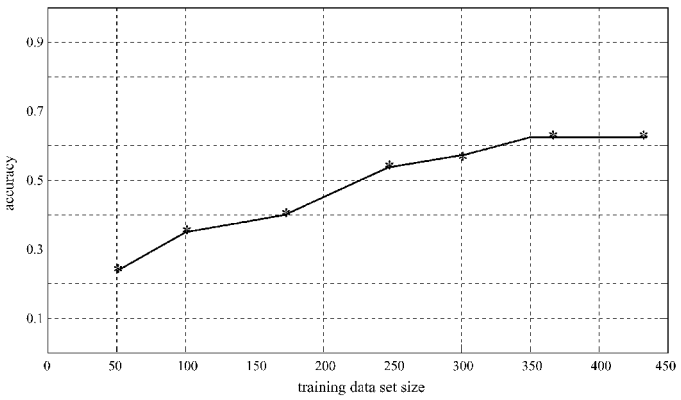

Fig. 10. Learned classification accuracy based on different training data size.

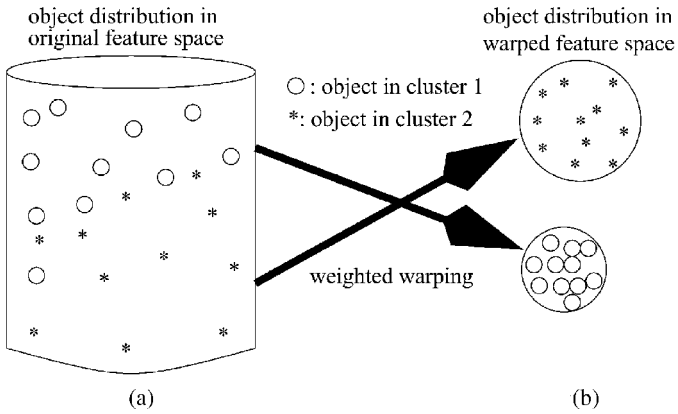

Fig. 11. Feature space transformation to support better node representation: (a) data distribution for two clusters in the original feature space; (b) data distribution for two clusters in their warped spaec with different discriminating features and weights.

Fig. 11 even they have heavy overlap in their original feature space. As mentioned in Section II, we use Gaussian function to approximate the distribution of video shots for each visual concept node in its warped feature subspace with a certain degree of accuracy.

We use geometric hashing to build the database indices. The database indexing structure includes a set of hash tables for different visual concept levels: a root hash table for keeping track of the information about all semantic clusters in the database, a hash table for each cluster to preserve the information about all its sublevel clusters, a set of hash tables for each internal nodes, a hash table for each leaf cluster for indexing all its scenes, and a hash table for each scene providing indices of all its video shots, and every index point to the disk pages that the corresponding video shot resides. The leaf cluster node as shown in Fig. 1, which indicates the end of the concept hierarchy, may include large number of video scenes or shots and it is inefficient to index the leaf cluster node by using only one hash table. The leaf cluster node can further be partitioned into a set of groups according to its distribution of video shots. This hierarchical partitioning of a leaf cluster node will end when the number of video shots in each group is less than a predefined threshold $\log N_R \ll D_R$, where N_R is the total number of video shots in the group, and D_R is the dimensions of the discriminating features for the corresponding leaf cluster node [48].

A. Integrated Video Query

Our hierarchical video database indexing structure can also support more powerful query-by-example. As mentioned above, the naive users can select two approaches to achieve query-by-example: similarity search by using the average properties and similarity search by using the node seeds.

To answer a query-by-example by using the average properties, our video database system first extracts the features of the query-example $X = (x_1, \ldots, x_r, \ldots, x_n)$, which are then compared to those of the semantic clusters as shown in Fig. 12. The similarity distance d_{qi} between the query-example and the cluster C_i in its warped feature subspace is calculated as

$$d_{qi} = \sum_{r=1}^{D_{c_i}} \frac{1}{\beta_{c_{i_r}}} \cdot d_r(x_r, \mu_{c_{i_r}}) \tag{12}$$

where $d_r(x_r, \mu_{c_{i_r}})$ is the similarity distance between the query-example and the semantic cluster C_i according to their rth representative feature. If

$$d_{qi} \leq \sum_{r=1}^{D_{c_i}} \frac{1}{\beta_{c_{i_r}}} \cdot \sigma_{c_{i_r}} \tag{13}$$

then the query processor will subsequently obtain the weighted sum of the Gaussian probabilities ρ that the subject of the query-example belongs to C_i:

$$\rho(X, \mu_{c_i}, \sigma_{c_i}) = \sum_{r=1}^{D_{c_i}} \frac{1}{\beta_{c_{i_r}}} \cdot \rho_r(x_r, \mu_{c_{i_r}}, \sigma_{c_{i_r}}). \tag{14}$$

Similarly, we can also get other potential semantic clusters where the subject of query-example also belongs to and the corresponding Gaussian probabilities. The semantic cluster with the maximum sum of weighted Gaussian probabilities ρ is first selected (Fig. 12). In the same fashion, the query processor can subsequently find the relevant sublevel clusters, groups, scenes, and then video shots that reside in the database. The video query results from *movie* and *news* clusters are shown in Figs. 13 and 14. The average performance of hierarchical video database indexing structure is given in Fig. 15. One can find that including more hierarchical visual concept levels in the database indexing structure can reduce the query time, but it also induces lower query accuracy. The average CPU time for video query is shown in Fig. 16. The average query time depends on three parameters: the number of the levels of the visual concepts in the concept hierarchy, the size of the feature subspaces for the relevant visual concept nodes, and the size of the leaf node.

In order to avoid the low featural support for the high-level semantic visual concepts (i.e., clusters), we have used the node seeds (i.e., principal video shots) for visual concept representation and indexing. Given a query video shot $T = \{x_1, x_2, \ldots, x_n\}$ characterized by full n-dimensional representative features, a similarity search is performed as follows.

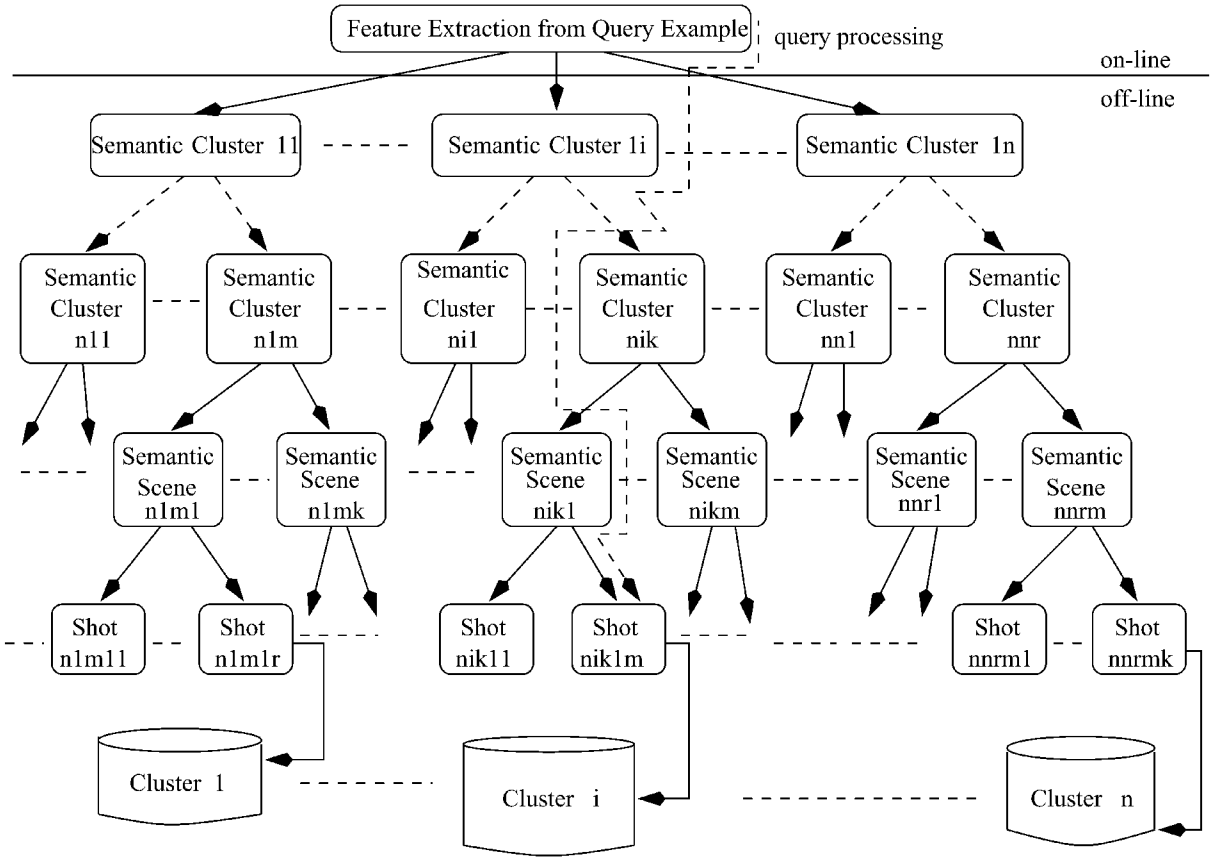

Fig. 12. Hierarchical query procedure: The query processor will first select the most relevant cluster, and then its sublevel cluster, and then the scene, and finally the video shot before linking it to its point in the storage disk.

Fig. 13. Shot-based video query results from a *movie* cluster.

Fig. 14. Shot-based video query results from a *video news* cluster.

a) The query processing subsystem first tries to find the best matched seed from each cluster. Since there are several node seeds for each cluster, the weighted feature-based similarity distance $D_F(T, ST_{C_i}^j)$ between the query shot T and the jth seed $ST_{C_i}^j$ of the cluster C_i is calculated:

$$D_F^i(T, ST_{C_i}^j) = \sum_{h=1}^{D_i} \frac{1}{\beta_{C_{i_h}}} \cdot D_{F_h}(T, ST_{C_i}^j) \tag{15}$$

where $D_{F_h}(T, ST_{C_i}^j)$ is the similarity distance between the query shot T and the jth seed $ST_{C_i}^j$ of cluster C_i on the basis of their hth dimensional features. The best matched seed in the corresponding cluster C_i can be determined as

$$D_F(T, ST_{C_i}^l) = \min\{D_F^i(T, ST_{C_i}^j) | ST_{C_i}^j \in \{ST_1, \ldots, ST_m\}\}. \tag{16}$$

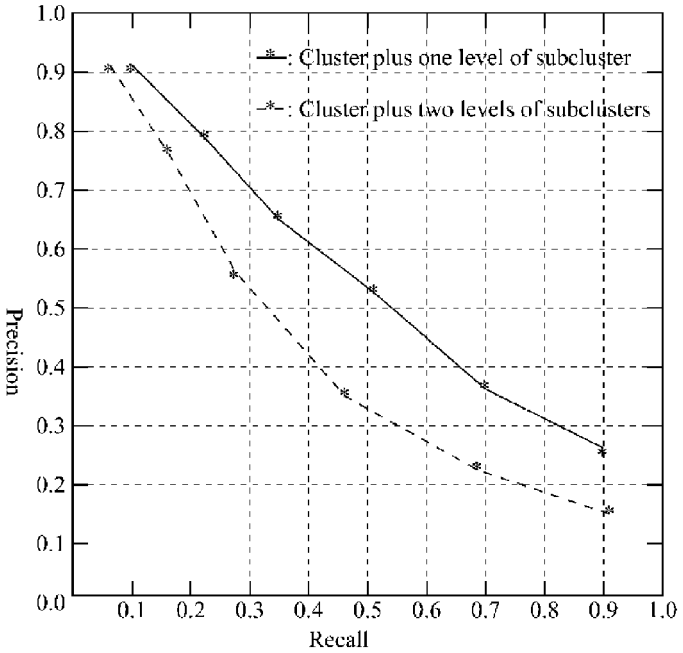

Fig. 15. Average performance of hierarchical video database indexing technique, and each point is obtained via 810 queries from three different video sources used in our system.

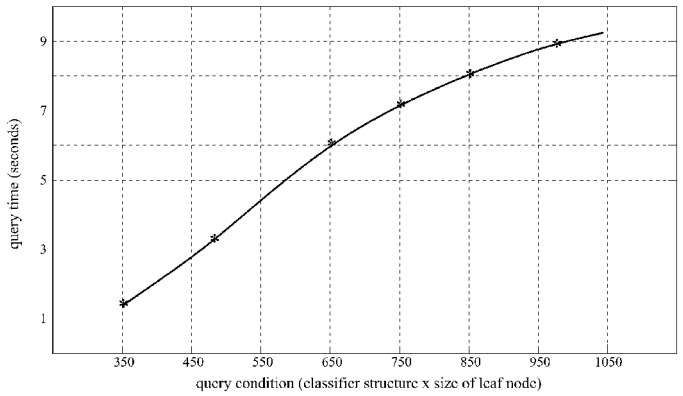

Fig. 16. Average query cost of our hierarchical video database indexing technique which is obtained from three video sources.

b) The query processing subsystem then find the best matched cluster:

$$\begin{cases} D_F(T, c_r) = \min\{D_F(T, ST^l_{c_i}) | c_i \in \{C_{11}, \ldots, C_{1n}\} \\ \rho(X, \mu_{c_r}, \sigma_{c_r}) = \max\{\rho(X, \mu_{c_i}, \sigma_{c_i}) | c_i \in \{C_{11}, \ldots, C_{1n}\}. \end{cases} \tag{17}$$

c) The query processing subsystem can then find the best matched sublevel clusters and finally obtain the best matched group by using the similar approach. The users can then decide which video they want by browsing the content abstractions of the ranked query results.

As introduced in Section III, each video shot inherits a hierarchy of the semantic labels (i.e., keywords used for constructing and interpreting the domain-dependent concept hierarchy) for the corresponding visual concept nodes that it belongs to. Semantic label is also used as an attribute for video database indexing as described in Section II, thus query by high-level visual concept can also be supported by using the semantic labels assigned to the video shots. To answer the query by visual concept, the query processor performs the similar procedure that is used for achieving query-by-exmaple, but only the attribute of the semantic label is used for computing the concept-based semantic similarity. Since the low-level visual features and the high-level semantic visual concepts are integrated for video database indexing in our system, more powerful video database search engine has been provided.

B. Hierarchical Video Database Browsing

Browsing has the advantage to keep the user in the loop in the search process (i.e., *I know it when I see it*), however, most the existing video retrieval systems do not support browsing because of the lack of efficient visual summary presentation structure. As mentioned before, a key issue to video database browsing is whether the visual summaries found make sense to users and whether the contextual and logical relationship of the visual summaries at different database level (which will be accessed by the users hierarchically) is well-defined.

In our current work, the video shots are classified into a set of hierarchical database management units according the domain-dependent concept hierarchy, and this can help us understand the context of video contents in the database. Since the contextual and logical relationships of the related visual concepts is defined via the domain-dependent concept hierarchy, our semantics-sensitive video classification and management structure can provide a good environment for organizing and presenting visual summaries hierarchically.

For each high-level visual concept node in our database model (i.e., a semantic cluster, sublevel cluster), we could create its visual summary either in terms of its most significant or principal components (i.e., lower-level visual concepts) or the representative video shots which are closest to the centers of each component, depending on the desired level of details. Our system can support three types of browsing: 1) browsing the whole video database via the summaries of all the semantic clusters, 2) browsing a semantic cluster via the summaries of its relevant sublevel clusters, and 3) browsing a leaf cluster via the summaries of its video scenes. For the high-level visual concepts, their visual summaries consist of the principal components of their relevant sublevel visual concepts as shown in Fig. 17. For the leaf visual concept node (i.e., semantic video scene) in the concept hierarchy, it is unsuitable to use the discontinuous key frames to provide the visual summary for the semantic video scene because the adjoining video shots may be classified into the same semantic video scene. Thus it is very important to detect the boundary of the corresponding semantic video scene and use a "significant" and continuous video pieces to produce a short but meaningful summary for the corresponding semantic video scene. As shown in Fig. 18, we integrated the adjoining and relevant video shots as the visual summary for the corresponding semantic video scene.

C. Related Problem Discussion

Because classifying video shots into a set of high-level semantic visual concepts according to the domain-dependent concept hierarchy can help us understand the logical and

Fig. 17. Visual summaries for video news cluster at the cluster level.

Fig. 18. Visual summaries for *video news* cluster at the scene level.

contextual relationships among the related visual concepts, our semantics-sensitive hierarchical video classification technique can support more effective video database indexing and access. Annotating videos by using the predefined keywords listed on the domain-dependent concept hierarchy could reduce the diversity of the annotation results from different people. Unfortunately, it still suffers from the following challenging problems.

- **Subjectivity problem:** The concept hierarchy should be domain-dependent because bridging the semantic gap, in general, is still impossible for current computer vision and machine learning techniques. For each semantic visual concept (i.e., each node on the classifier and database indexing structure), its feature subspace, feature weights (i.e., importances) and classification rule are predetermined without considering the user's subjectivity [50]–[52]. The same domain-dependent concept hierarchy is used as the inherent database indexing tree structure for video management and holden for all the users. While it is very important to enable real-time updating of these predetermined feature subspaces, feature weights, classification rules and even the inherent video database indexing structure according to the user's subjectivity, it is impossible to perform this real-time updating for the large-scale video database. In our current work, the concept hierarchy and the semantic labels for the training examples are provided by the domain experts or obtained according to some common knowledge. It is very important to generate this concept hierarchy atomatically according to the user's subjectivity, however, it is very hard if not impossible for current machine learning techniques to achieve this [54]. Using the common knowledge or the domain knowledge from the experts is a good tradeoff for us to address this hard problem now, obviously, automatic techniques are expected to be provided in the future.

  Our hierarchical semantics-sensitive video classifier tries to bridge the semantic gap between the low-level visual features and the high-level semantic visual concepts, so that the feature-based visual similarity can correspond to the concept-based semantic similarity by learning from the limited labeled training examples. Obviously, all these are schieved under some specific domain. On the other hand, the semantic labels for these training examples are also provided by some domain experts or naive users. Therefore, the performances of the proposed feature selection and video classification techniques also suffer from the subjectivity problem.
- **High-level summary determination problem:** Proper identification of principal components is very important in supporting visual summarization for the high-level visual concepts. However, it is not a trivial work. In our current work, we define the visual summary for the high-level visual concepts by exploiting the domain knowledge or selecting the principal video shots which are close to the centers of their relevant sublevel visual concepts. It is important to develop more effective techniques in the near future for determining the principal components automatically for the high-level visual concepts.
- **Integrated similarity problem:** It is very important to integrate query-by-example with query-by-keywords because the users may prefer to access video contents via high-level visual concepts. However, using the shot-based low-level visual features along cannot characterize the high-level semantic visual concepts effectively even the discriminating feature subspace is extracted. On the other hand, manual annotations and keywords are too subjective. It is important to integrate the feature-based visual similarity with the concept-based semantic similarity for defining more effective similarity measurement among the high level visual concepts [50]–[52]. Unfortunately, it is not an easy work to determine the importance between these two similarity measurements because the importance also depends on the user's subjectivity.

On the other hand, the online relevance feedback approach is more attractive for supporting semantic video retrieval because it keeps the naive users in the loop of retrieval. Since the naive users can exchange their subjective judgments with the database system interactively, the online relevance feedback approach is more suitable for serving a large population of naive users [5], [28], [29]. However, the conventional online rele-

vance feedback techniques suffer from the following problems when they are applied for video retrieval over the large-scale database: 1) Few works have been done to integrate the online relevance feedback with the inherent database indexing structure, thus the conventional online relevance feedback techniques cannot scale to the database size [53]. The conventional nearest neighbor search is also unsuitable for supporting online relevance feedback because it treats all the visual features with the same importance. If the naive users do not have a good example to start a query, query refinement around some bad examples is misleading and also very time-consuming. 2) The expected numbers of query iterations and samples for each query iteration (i.e., query results deployed to the naive users interactively) should be small enough because the naive users may be impatient to browse and label large-scale samples. Support Vector Machine techniques have been used to address the problem of limited samples by regarding video retrieval as a strict two-class classification problem. Since the feature space for video shot representation and indexing is normally in high-dimensions, it is still an open problem to learn the stable classification rules from the limited samples in real-time [50]. Our hierarchical video database indexing technique can support more effective video access over the large-scale database. Therefore, the next attractive research issue is how we can support more effective relevance feedback based on this proposed hierarchical video database indexing structure.

V. Conclusions and Future Works

We have proposed a novel framework, called ***ClassView***, to make some advances in overcoming the problems suffered by the existing content-based video retrieval systems. A hierarchical semantics-sensitive video classifier is proposed to shorten the semantic gap between the low-level visual features and the high-level semantic concepts. The hierarchical structure of the semantics-sensitive video classifier is derived from the domain-dependent concept hierarchy of video contents in the database. Relevance analysis is used to shorten the semantic gap by selecting the discriminating visual features and suitable importances. The EM algorithm is used to determine the classification rule for each visual concept node. A hierarchical video database indexing and summary presentation technique is also proposed to support more effective video access over the large-scale database. Integrating video querying with video browsing has provided great opportunity for supporting more powerful video search engines.

While we are not claiming to be able to solve all the problems related to content-based video retrieval, we have made some advances toward the final goal, close to human-level video retrieval by using the domain-dependent concept hierarchy. The following research issues should be addressed in the future to avoid the limitations of our hierarchical semantics-sensitive video classification and indexing techniques.

- Research in semantics-sensitive video classification is currently limited by the relative lack of large-scale labeled training data set. It should be very attractive to generate the classification rules by integrating the unlabeled video clips with the limited labeled video clips. Since the unlabeled training examples may consist of different visual concepts, they will not follow the joint probability and they may also degrade the classification performance. Mixture probability model is expected to be used for avoiding this joint probability problem when using the unlabeled training examples learns more accurate classification rules.
- Video characterization and classification via integration of multiple media, such as video, audio, and text information such as closed caption, will provide more meaningful results. At the same time, it is urgent to address problems of the normalization of multiple cues and the automatic determination of their importances for semantic visual similarity judgment.
- Shot-based low-level visual features may be too general to characterize the semantic visual concepts associated with the video shots. A single video shot may consist of multiple semantic visual concepts, thus the single video shot should be permitted to be classified into the related multiple visual concept nodes. Semantic video classification can be improved by detecting the salient objects such as human faces, skin color regions, from the video shots because the presence or absence of the salient objects can indicate the presence or absence of the relevant visual concepts more effectively.
- The *basic assumption* of our work is too strong for some video types, thus our current works include high classification errors for the movie video source which includes more complex visual concepts. High-dimensional data visualization should be developed for evaluating the effectiveness of the semantic video classifier because the semantically similar video shots should be close to each other in their warped discriminating feature subspace.
- It is very important to enable the real-time updating of these predetermined feature subspaces, dimensional weights, classification rules or even the inherent concept hierarchy according to the user's subjectivity for the large-scale video database. Our hierarchical video database indexing structure can support more effective video retrieval and concept-oriented hierarchical video database browsing, thus it is very attractive to support the online relevance feedback over this hierarchical video database indexing structure and achieve more effective query optimization for the large-scale video database. The final users will ultimately evaluate the performances of the inherent video database representation and indexing model, semantics-sensitive video classification under the given database model, query optimization and concept-oriented hierarchical video database browsing in the task of content-based video retrieval. It is very important to study the human factors in supporting content-based video retrieval through this proposed prototype system.

Our recent research works focus on providing some practical solutions for these challenging problems.

Acknowledgment

The authors thank the reviewers for their useful comments and suggestions to make this paper more readable. They also

thank R. Gandhi at UNC-Charlotte for his great help in implementing part of this prototype system.

REFERENCES

[1] P. Salembier, R. Qian, N. O'Connor, P. Correia, I. Sezan, and P. van Beek, "Description schemes for video programs, users and devices," *Signal Process.: Image Commun.*, vol. 16, pp. 211–234, 2000.

[2] A. Beritez, S. Paek, S.-F. Chang, A. Puri, Q. Huang, J. R. Smith, C.-S Li, L. D. Bergman, and C. N. Judice, "Object-based multimedia content description schemes and applications for MPEG-7," *Signal Process.: Image Commun.*, vol. 16, pp. 235–269, 2000.

[3] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, vol. 38, pp. 23–31, 1995.

[4] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *Int. J. Comput. Vis.*, vol. 18, pp. 233–254, 1996.

[5] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 644–655, 1998.

[6] A. Humrapur, A. Gupta, B. Horowitz, C. F. Shu, C. Fuller, J. Bach, M. Gorkani, and R. Jain, "Virage video engine," *Proc. SPIE, Storage and Retrieval for Image and Video Databases V*, pp. 188–197, Feb. 1997.

[7] 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. Circuits Syst. Video Technol.*, vol. 8, pp. 602–615, Sept. 1998.

[8] S. Satoh and T. Kanade, "Name-It: Association of face and name in video," presented at the Proc. Computer Vision and Pattern Recognition, 1997.

[9] Y. Deng and B. S. Manjunath, "NeTra-V: Toward an object-based video representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 616–627, Sept. 1998.

[10] H. J. Zhang, J. Wu, D. Zhong, and S. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognit.*, vol. 30, pp. 643–658, 1997.

[11] A. K. Jain, A. Vailaya, and X. Wei, "Query by video clip," *ACM Multimedia Syst.*, vol. 7, pp. 369–384, 1999.

[12] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1349–1380, Dec. 2000.

[13] A. Guttman, "R-trees: A dynamic index structure for spatial searching," in *ACM SIGMOD'84*, 1984, pp. 47–57.

[14] D. B. Lomet and B. Salzberg, "The hB-tree: A multiattribute indexing method with good guaranteed performance," in *Proc. ACM Symp. Trans. Database Syst.*, vol. 15, 1990, pp. 625–658.

[15] K. Lin, H. V. Jagadish, and C. Faloutsos, "The TV-tree: An index structure for high dimensional data," *VLDB Journal*, 1994.

[16] D. A. White and R. Jain, "Similarity indexing with the SS-tree," in *Proc. 12th Int. Conf. Data Engineering*, New Orleans, LA, 1996, pp. 516–523.

[17] N. Katayama and S. Satoh, "The SR-tree: An index structure for high dimensional nearest neighbor queries," presented at the ACM SIGMOD, 1997.

[18] S. Berchtold, D. A. Keim, and H. P. Kriegel, "The X-tree: An index structure for high-dimensional data," presented at the VLDB, 1996.

[19] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining," *VLDB*, 1994.

[20] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large database," *ACM SIGMOD*, 1996.

[21] S. Gupta, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large database," *ACM SIGMOD*, 1998.

[22] W. Zhou, A. Vellaikal, and C. Kuo, "Rule-based video classification system for basketball video indexing," *ACM Multimedia*, 2000.

[23] J. Huang, S. R. Kumar, and R. Zabih, "An automatic hierarchical image classification scheme," in ACM Multimedia, Bristol, U.K., 1998.

[24] G. Sheikholeslami, W. Chang, and A. Zhang, "Semantic clustering and querying on heterogeneous features for visual data," in ACM Multimedia , Bristol, U.K., 1998.

[25] Q. Huang, Z. Liu, and A. Rosenberg, "Automated semantic structure reconstruction and representation generation for broadcast News," in *Proc. SPIE*, 1999.

[26] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 947–963, Sept. 2001.

[27] T. P. Minka and R. W. Picard, "Interactive learning using a society of models," *Pattern Recognit.*, vol. 30, p. 565, 1997.

[28] Y. Rui and T. S. Huang, "A novel relevance feedback technique in image retrieval," in *Proc. ACM Multimedia'99*, 1999, pp. 67–70.

[29] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "MindReader: Querying databases through multiple examples," in *Proc. VLDB'98*, 1998, pp. 218–227.

[30] A. Vailaya, A. Jain, and H. J. Zhang, "On image classification: City versus landscape," *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries*, pp. 3–8, 1998.

[31] H. Yu and W. Wolf, "Scenic classification methods for image and video databases," *Proc. SPIE*, vol. 2606, pp. 363–371, 1995.

[32] R. S. Michalski and R. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-5, pp. 396–410, 1983.

[33] J. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[34] W. Buntine, "Learning classification tree," *Statist. Comput.*, vol. 2, pp. 63–73, 1992.

[35] M. I. Jordan, "A statistical approach to decision tree modeling," *Mach. Learn.*, 1996.

[36] Y. Rui, T. S. Huang, and S. Mehrotra, "Constructing table-of-content for videos," *ACM Multimedia Syst.*, vol. 7, pp. 359–368, 1999.

[37] B.-L. Yeo and M. M. Yeung, "Classification, simplification and dynamic visualization of scene transition graphs for video browsing," *Proc. SPIE*, vol. 3312, pp. 60–70, 1997.

[38] 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," *Proc. SPIE: Multimedia Storage and Archiving Systems IV*, vol. 3846, pp. 148–164, Sept. 1999.

[39] J. R. Smith, "VideoZoom spatial-temporal video browsing," *IEEE Trans. Multimedia*, vol. 1, pp. 157–171, June 1999.

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

[41] 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, pp. 201–207.

[42] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimentionality reduction using genetic algorithm," *IEEE Trans. Evol. Comput.*, vol. 4, pp. 164–171, July 2000.

[43] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to WordNet: An on-line lexical database," *Int. J. Lexicography*, vol. 3, pp. 235–244, 1990.

[44] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation," in *Proc. SPIE*, 2001.

[45] M. M. Yeung and B.-L. Yeo, "Video visualization for compact presentation and fast browsing of picturial content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 771–785, Oct. 1997.

[46] J. Fan, D. K. Y. Yau, W. G. Aref, and A. Rezgui, "Adaptive motion-compensated video coding scheme toward content-based bitrate allocation," *Journal of Electronic Imaging*, vol. 9, no. 4, 2000.

[47] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2000.

[48] Y. Manolopoulos, Y. Theodoridis, and V. J. Tsotras, *Adavanced Database Indexing*. Norwell, MA: Kluwer, 2000.

[49] S. J. Raudys and A. K. Jain, "Small samples size effects in statistical pattern recognition," *IEEE Trans. Pattern Anal. Mchine Intell.*, vol. 13, pp. 252–264, Mar. 1991.

[50] X. Zhu and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Syst.*, 2002.

[51] C. Zhang and T. Chen, "Active Learning for Information Retrieval: Using 3D Models as an Example," Carnegie Mellon Univ., Pittsburgh, PA, AMP01–04, 2002.

[52] X. Zhu and T. S. Huang, "Unifying keywords and visual contents in image retrieval," *IEEE Multimedia*, no. 2, pp. 23–33, Apr.-June 2002.

[53] P. Wu and B. S. Manjunath, "Adaptive nearest neighbor search for relevance feedback in large image database," *ACM Multimedia*, 2001.

[54] A. B. Benitez and S.-F. Chang, "Multimedia knowledge integration, summarization and evaluation," presented at the 2002 International Workshop on Multimedia Data Mining, Edmonton, AB, Canada, July 23–26, 2002.

Jianping Fan received the M.S. degree in theory physics from Northwestern University, Xian, China, in 1994, and the Ph.D. degree in optical storage and computer sceince from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, in 1997.

He was a Researcher at Fudan University, Shanghai, during 1998. From 1998 to 1999, he was a Researcher with the Japan Society for Promotion of Sciences (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From Septermber 1999 to 2001, he was researcher in Department of Computer Science, Purdue University, West Lafayette, IN. He is now an Assistant Professor in the Department of Computer Science, University of North Carolina at Charlotte. His research interests include nonlinear systems, error correction codes, image processing, video coding, semantic video computing, and content-based video indexing and retrieval.

Ahmed K. Elmagarmid (M'88-SM'93) received the M.S. and Ph.D. degrees in computer and information sciences from The Ohio State University (OSU), Columbus, in 1980 and 1985, respectively.

He is now a Professor of computer science at Purdue University, West Lafayette, IN, as well as an industry consultant. His areas of research interests are data quality, video databases, heterogeneous databases, and distance learning. He is the Founding Editor-in-Chief of *International Journal on Distributed and Parallel Databases*. He serves as a Editor of *Information Science Journal*, *International Journal of Communication Systems*, and the book series *Advanced Database Systems* (Norwell, MA, Kluwer, 1995).

Dr. Elmagarmid received a National Science Foundational PYI Award in 1988 and was named a "Distinguished Alumnus" of OSU in 1993, and the University of Dayton, Dayton, OH, in 1995. He has served on the editorial board of IEEE TRANSACTIONS ON COMPUTERS and he is now the associate editor for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. He is a Chair of the steering committee of the Symposium on Research Issues on Data Engineering and was one of its founders; he serves on the steering committee of IEEE ICDE and has served on its program chair and general chair. He has served on various program and organization committees. He is a Member of the ACM.

Xingquan Zhu received the M.S. degree in communication from Xidian University, Xian, China, and the Ph.D degree in computer science from Fudan University, Shanghai, China.

He is currently a Researcher with the Department of Computer Science, Purdue University, West Lafayette, IN. He previously spent four months with Microsoft Research Center, Beijing, China, where he was working on relevance feedback for image indexing and retrieval. His research interests include image processing, content-based video retrieval and indexing.

Walid G. Aref (S'90-M'93) received the Ph.D. degree in computer science from the University of Maryland, College Park, in 1993.

He has been with Matsushita Information Technology Laboratory and the University of Alexandria, Egypt. Currently, he is an Associate Professor, Department of Computer Science, Purdue University, West Lafayette, IN. His research interests include efficient query processing and optimization algorithms and data mining in spatial and multimedia databases.

Dr. Aref is a member of the ACM.

Lide Wu graduated from Fudan University, Shanghai, China, in 1958.

He was a Visiting Scholar with Princeton University, Princeton, NJ, in 1980, and with Brown University, Providence, RI, in 1981. He was the Dean of the Department of Computer Science, Fudan University, from 1982 to 1983. He is now a Professor with Fudan University. His main research interests are computer vision, natural language processing, and video database systems.