Concept-Oriented Indexing of Video Databases: Toward Semantic Sensitive Retrieval and Browsing

Jianping Fan, Hangzai Luo, and Ahmed K. Elmagarmid, *Senior Member, IEEE*

***Abstract*—Digital video now plays an important role in medical education, health care, telemedicine and other medical applications. Several content-based video retrieval (CBVR) systems have been proposed in the past, but they still suffer from the following challenging problems: *semantic gap*, *semantic video concept modeling*, *semantic video classification*, and *concept-oriented video database indexing and access*. In this paper, we propose a novel framework to make some advances toward the final goal to solve these problems. Specifically, the framework includes: 1) *a semantic-sensitive video content representation framework* by using principal video shots to enhance the quality of features; 2) *semantic video concept interpretation* by using flexible mixture model to bridge the semantic gap; 3) *a novel semantic video-classifier training framework* by integrating feature selection, parameter estimation, and model selection seamlessly in a single algorithm; and 4) *a concept-oriented video database organization technique* through a certain domain-dependent concept hierarchy to enable semantic-sensitive video retrieval and browsing.**

***Index Terms*—Database, video analysis, video browsing, video indexing, video retrieval.**

I. Introduction

As STORAGE and bandwidth capacities increase, digital video now plays an important role in a wide range of multimedia applications. As large-scale video collections come into view, there is an urgent need for characterization efforts on semantic video classification, so that the users can select the relevant video clips at the semantic level. Unfortunately, our current ability on semantic video classification is so far primitive because of the following challenging issues.

- **Semantic-Sensitive Video Analysis:** The performance of semantic video classifiers largely depends on the quality of features (i.e., the *ability* of the selected low-level perceptual features to discriminate among various semantic video concepts). On the other hand, the quality of features also depends on the effectiveness of the underlying video patterns that are selected for video content representation and feature extraction. Most existing content-based video retrieval (CBVR) systems select video shots [1]–[5], homogeneous video regions, or semantic video objects [5]–[9] as the underlying video patterns for video content representation and feature extraction. The difficulty of using the video shots and homogeneous video regions for video content representation and feature extraction is the lack of means to relate the low-level perceptual features to the semantic video concepts [11]–[13]. The major problem for using the semantic video objects for video content representation and feature extraction is that automatic semantic video object extraction in general is very hard, if not impossible [14]–[28]. Moreover, most existing CBVR systems only use the shot-based or region-based low-level visual features. However, original video is a synergy of multimodal inputs such as audio, vision, and image-text [29]–[32]. Thus, new video content representation frameworks, that can not only provide more discriminating multimodal perceptual features, but also avoid performing uncertain semantic video object extraction, are strongly expected to enhance the quality of features.
- **Semantic Video Concept Modeling:** The major difficulty of the existing CBVR systems is that they are unable to support video access at the semantic level because of the semantic gap. Thus, bridging the *semantic gap* may be the biggest challenge that we face in supporting content-based video retrieval and it has recently received much attention [33]–[53]. To bridge the semantic gap, the rule-based (i.e., model-based) approaches use domain knowledge to define the perceptional rules for extracting semantic video concepts [33]–[41]. Some researchers also used the relevance feedback to bridge the semantic gap in the retrieval loop [59]–[65]. Statistical machine learning has also been used to bridge the semantic gap by discovering nonobvious correlations (i.e., hidden rules) among multimodal inputs [42]–[53]. However, no existing work has addressed the underlying multimodal context integration model that can be used to explore the joint effects among the multimodal inputs for semantic video concept interpretation.
- **Semantic Video Classification:** Many semantic video-classification techniques have been proposed in the past, but the limited number of pages does not allow us to survey all these related work. Instead we try to emphasize some of the work that is most related to our proposed work. The related semantic video-classification techniques can be classified into two categories [32].
  1) *Rule-based* (i.e., *model-based*) *approach* by using domain knowledge to define the perceptional rules and achieve semantic video classification

Manuscript received September 19, 2000; revised November 5, 2003. This work was supported by National Science Foundation under 0208539-IIS and 9974255-IIS and by grants from the AO Foundation, HP, IBM, Intel, NCR, Telcordia, and CERIAS. The associate editor coordination the review of this manuscript and approving it for publication was Dr. Jean-Luc Dugelay.

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

A. K. Elmagarmid is with the Department of Computer Science, Purdue University, West Lafayette, IN 47907 USA.

Digital Object Identifier 10.1109/TIP.2004.827232

1057-7149/04$20.00 © 2004 IEEE

[33]–[41]. One advantage of the rule-based approach is the ease to insert, delete, and modify the existing rules when the nature of the video classes changes. However, effective semantic video-classification techniques should discover not only the perceptional rules that can be perceived by human inspection, but also the hidden significant correlations (i.e., hidden rules) among multimodal inputs. Therefore, the rule-based approach is only attractive for the video domains such as news and films that have well-defined story structure for the semantic units (i.e., film and news making rules) [36]–[41].

2) *Statistical approach* by using statistical machine learning to bridge the semantic gap [42]–[53]. The statistical approach can support more effective semantic video classification by discovering nonobvious correlations (i.e., hidden rules) among different video patterns. However, its performance largely depends on the success of the underlying classifier training framework and the ability of the selected low-level multimodal perceptual features on discriminating among various semantic video concepts.

- **Feature Selection and Dimension Reduction:** Theoretically, having more features should give us more discriminating power to enable more effective semantic video classification [72]–[75]. However, the time requirements for classifier training often grow dramatically with the feature dimensions, thus including more features makes it very difficult to obtain good estimates of many parameters for the classifier and renders the classifier training algorithm impractical. An important question for supporting more effective semantic video classification is how to select a good subset of features. A good choice of feature subset may not only improve the classifier's performance (i.e., accuracy), but also aid in finding smaller classifier models and result in better understanding and interpretation of the classifier.
- **Concept-Oriented Video Database Organization and Access:** Research developments in Computer Vision and Database related disciplines have traditionally been independent and unrelated [10]. Even today, there is a lack of research synergy between the two fields. When truly large video data sets come into view, database indexing can no longer be ignored to support more effective CBVR systems. However, the traditional database indexing structures are unsuitable for video database organization because they suffer from the problems of the *curse of dimensions* [54]–[58].

The essential goal of concept-oriented video database organization is to enable video access at the semantic level and to support naive users to specify and evaluate their query concepts more effectively and efficiently [57], [58]. There are three widely accepted approaches to achieving semantic video retrieval: 1) *query-by-example via online relevance feedback* [59]–[65]; 2) *query-by-keyword* [57], [58]; and 3) *video database browsing* [76]–[82]. Each approach represents a useful way of accessing a video database. Approach 1) allows a user to specify his/her query concept and retrieve the database via an example video clip. Approach 2) is convenient for users who want to search for video clips based on semantic concepts as described in keywords. Approach 3) is attractive for naive users who have no prior knowledge of the video collections in a video database and no precise query concepts in mind. However, each of these approaches has its limitations. For Approach (1), most existing techniques have not yet achieved the level that allow a naive user to specify his/her initial query concept effectively when he/she does not have good examples at hand. For Approach 2), the main obstacle is the lack of means for automatic text annotation of large-scale video collections. For Approach 3), browsing based on semantic concepts is yet to be realized due to the lack of suitable concept-oriented video database organization structure.

Based on these observations, this paper proposes a novel framework to address these challenging problems in a certain **medical education video domain**, which has strong application impact but has never been addressed by other researchers. In summary, the contributions of this paper include:

- a novel semantic-sensitive video content characterization framework by using principal video shots to enhance the ability of the low-level multimodal perceptual features on discriminating among various semantic video concepts;
- a probabilistic semantic video concept modeling framework by using flexible mixture model to bridge the semantic gap;
- a novel classifier training framework by integrating feature subset selection, parameter estimation and classifier model selection seamlessly in a single algorithm;
- a novel concept-oriented video summarization and database organization technique to enable semantic-sensitive video retrieval and browsing over large-scale video collections.

This paper is organized as follows. Section II introduces a novel framework to support semantic-sensitive video analysis. Section III proposes a probabilistic semantic video concept modeling framework to bridge the semantic gap. A novel semantic video-classification algorithm is proposed in Section IV. Section V presents a concept-oriented video summarization and database organization technique to enable semantic-sensitive video retrieval and browsing. Section VI gives the theoretical analysis of the performance of our framework. We conclude in Section VII.

II. Semantic-Sensitive Video Content Analysis

While a CBVR system for medical education is not necessarily capable of understanding semantics of medical video clips as medical experts do, it is necessary to understand: what are the suitable *concept-sensitive video patterns* for interpreting the ***semantic medical concepts*** in a certain domain for medical education videos? A good semantic-sensitive video content representation framework should be able to enhance the quality of features (i.e., enhance their ability to discriminate among various semantic medical concepts) and avoid performing uncertain semantic video object extraction.

Based on this understanding, we have developed a novel framework by using ***principal video shots*** (i.e., concept-sensitive video patterns) for video content representation and feature extraction. In a certain medical education video domain, the semantic medical concepts that should be indexed may be limited and thus can be pre-defined by medical experts. On the other hand, these pre-defined semantic video concepts are implicitly or explicitly related to some domain-dependent ***multimodal salient objects*** (visual, auditory, and image-textual salient objects) because video creation in a certain medical education domain is not really random but with the concept-driven multimodal salient objects. Thus the concept-sensitive *principal video shots* are defined as the integration units of the concept-driven multimodal salient objects associated with the relevant video shots.

The visual salient objects for semantic-sensitive video content characterization are not necessary the semantic video objects but some domain-dependent and concept-driven *regions of interest* that are effective to characterize the pre-defined semantic medical concepts. The auditory and image-textual salient objects for concept-sensitive video content characterization are not necessary the recognized speech and image-text but some domain-dependent auditory and image-textual patterns that are explicitly related to the pre-defined semantic medical concepts. For example, the presences of semantic medical concepts, such as *lecture presentation, gastrointestinal surgery, diagnosis, dialog*, and *traumatic surgery*, are implicitly related to the *visual salient objects* such as "human faces," "blood-red regions," , "gastrointestinal regions," and "skin regions," the *auditory salient objects* such as "single-human speech," "multiple-human speech (dialog talking)," "medical equipment noise," "silence," and the *image-textual salient objects* such as "text titles," "slides," and "sketch." While the concept-driven and domain-dependent multimodal salient objects are not exactly the multimodal semantic objects, they can have *certain perceptual properties in common* as the relevant multimodal semantic objects have and thus they are able to relate their low-level multimodal perceptual features to the relevant semantic medical concepts under certain vision purposes.

As illustrated in Fig. 1, the "bridgeless" semantic gap between the concept-insensitive low-level multimodal signals and the ***elementary semantic medical concepts*** is bridged by two steps: 1) bridging the ***semantic gap 1*** by detecting the concept-driven and domain-dependent multimodal salient objects automatically and 2) bridging the ***semantic gap 2*** by using a statistical classification technique to implicitly link the concept-sensitive principal video shots into the relevant elementary semantic medical concepts under certain vision purposes.

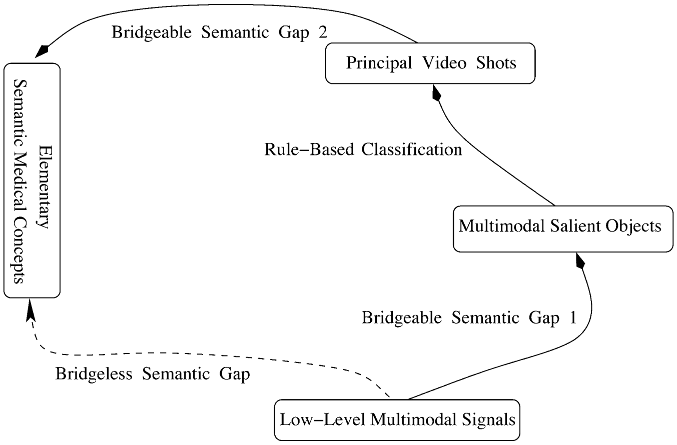

Bridgeable Semantic Gap 2
Principal Video Shots
Elementary Semantic Medical Concepts
Rule−Based Classification
Multimodal Salient Objects
Bridgeable Semantic Gap 1
Bridgeless Semantic Gap
Low−Level Multimodal Signals

Fig. 1. Proposed semantic-sensitive video content representation framework by using concept-sensitive principal video shots, where the "bridgeless" semantic gap between the concept-insensitive low-level multimodal signals and the elementary semantic medical concepts is now divided into two "small" bridgeable gaps.

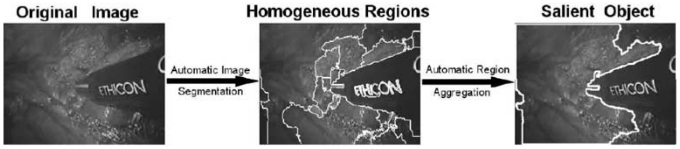


Fig. 2. Flowchart for our automatic salient object detection function, where the neighboring images regions with the same semantic label are automatically aggregated to form a certain type of the concept-sensitive salient objects.

To support this novel video content representation framework, the concept-insensitive video shots are first determined automatically by using adaptive shot detection techniques [11]–[13]. The auditory features have also been integrated with the visual features to detect the perceptual content changes among frames [29]–[32]. Based on the medical knowledge given by our medical consultants, a set of multimodal salient object detection functions have been designed and each function is able to detect one certain type of these pre-defined concept-driven and domain-dependent multimodal salient objects under certain vision purposes.

We use our visual salient object detection function for "gastroinstinal regions" as an example to show how we can design our multimodal salient object detection functions. Our visual salient object detection function for "gastrointestinal regions" consists of the following three components as shown in Fig. 2.

1) Image regions with homogeneous color or texture are obtained by using our automatic image segmentation techniques [22], [23]. This automatic image segmentation procedure are performed on a set of video frames that consist of the visual salient object of "gastrointestinal regions." These video frames are selected from different medical video clips with various illuminations.
2) The homogeneous image regions, that are implicitly related to the visual salient object of "gastrointestinal regions", are annotated and certified by our medical consultants and medical students. Region-based low-level visual features, such as dominant colors and variances, Tamura textures, object density (i.e., coverage ratio between object region and relevant rectangular box for object representation), height-width ratio for the object rectangular box, are extracted for characterizing the visual properties of these labeled image regions. To generate the detection function for the visual salient object of "gastrointestinal regions," an automatic image region classification technique is performed to determine

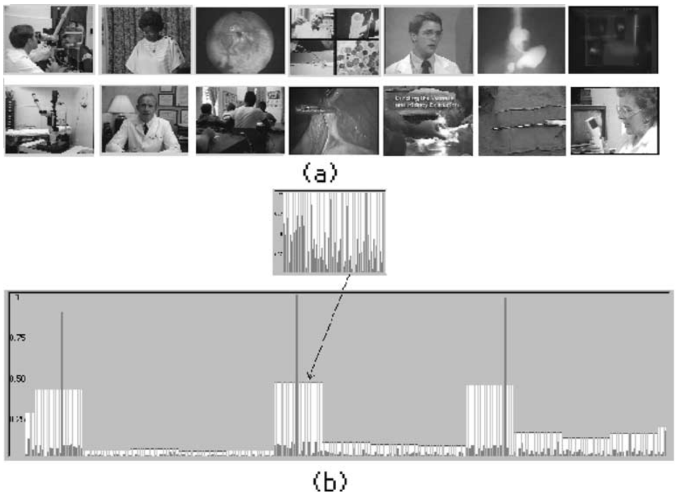


Fig. 3. Video shot detection results from a medical education video. (a) Part of the detected shot boundaries. (b) The corresponding color histogram difference and the determined thresholds for different video shots, where the small window shows the local properties of the color histogram difference.

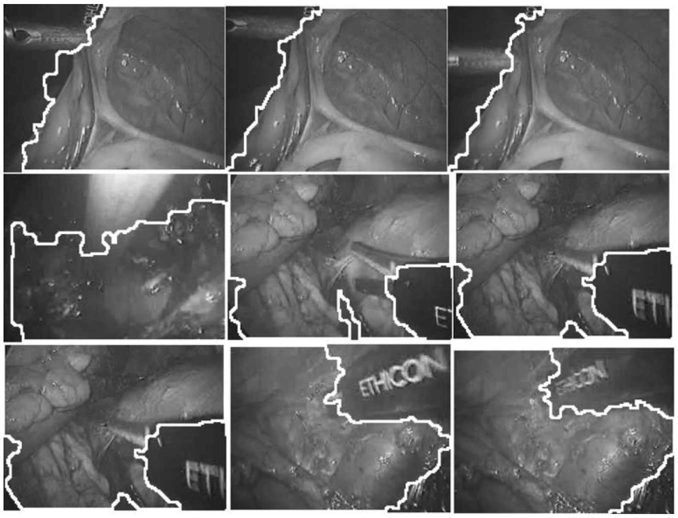


Fig. 4. Results on visual salient object detection for "gastrointestinal regions," where the white lines indicate the boundaries for the gastrointestinal regions.

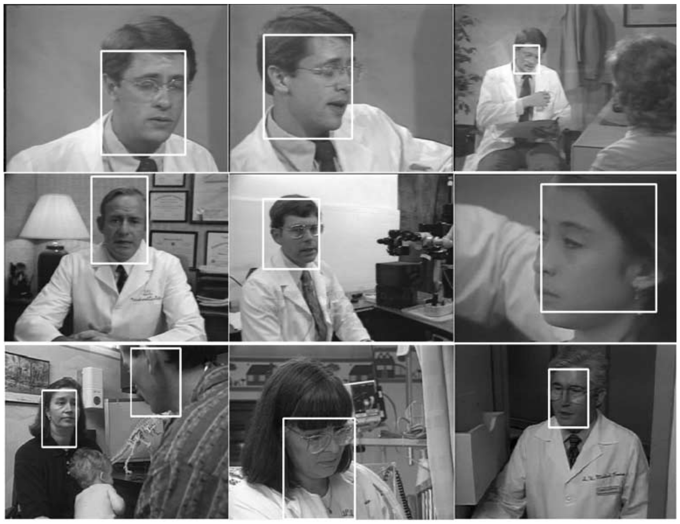


Fig. 5. Object detection results for "human face" from medical education videos.

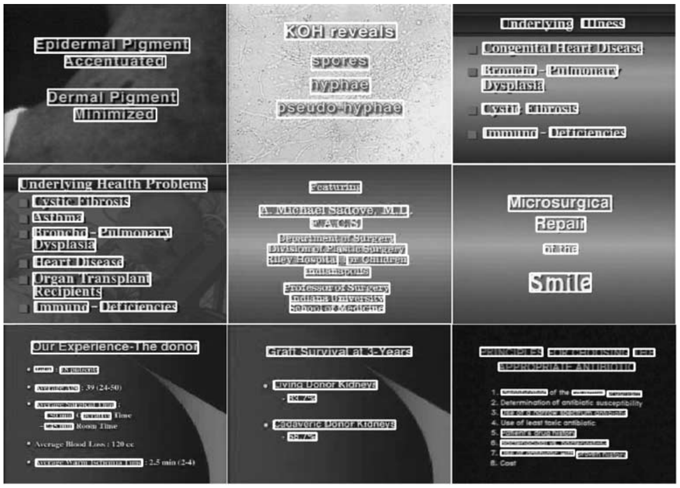


Fig. 6. Object detection results for "lecture slide" from medical education videos.

the implicit relationship between the semantic labels and the region-based low-level visual features by using the support vector machine (SVM). The connected homogeneous image regions with the same semantic label are aggregated as the visual salient object of "gastrointestinal regions".

3) The temporal tracking technique is used to integrate the visual salient object detection results of "gastrointestinal regions" within the same video shot as a single output.

Our video shot detection results from a medical video clip are shown in Fig. 3. Our multimodal salient object detection results for "gastrointestinal regions," "human face," and "lecture slide" are shown in Figs. 4–6, respectively. We have also proposed a semi-automatic salient object generation technique via a human-computer interaction procedure [25], [26]. As shown in Fig. 7, the human user can first define the boundary of a salient object, and this human-defined object boundaries are then refined by a intra-frame *snaking* procedure [28]. An automatic image-segmentation technique is then performed on the determined semantic objects to obtain their region relationship graphs. The region relationship graphs tell us which regions should be aggregated to form the salient objects and this can be taken as an interactive object model-definition procedure. The salient objects are then tracked among frames within a video shot.

After these pre-defined concept-driven and domain-dependent multimodal salient objects are obtained, a rule-based classification technique is used to generate the concept-sensitive *principal video shots*. The concept-driven multimodal salient objects and the associated video shots are integrated as the concept-sensitive principal video shots for semantic-sensitive video content representation and feature extraction.

III. Semantic Video Concept and Database Modeling

It seems that no existing CBVR system has fully answered the following questions [10].

- Which *video database model* can be used to support concept-oriented video database organization and access?

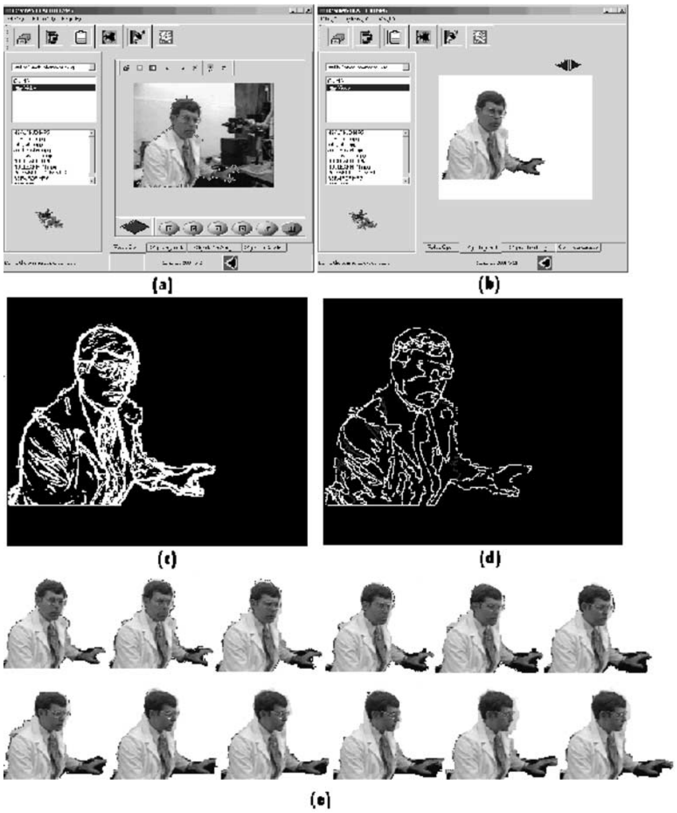


Fig. 7. Object extraction results via a semi-automatic approach. (a) Human-computer interaction interface. (b) Human-defined semantic object after intra-frame snaking. (c) Color edges of semantic object. (d) Region boundary of semantic object. (e) Temporal tracking results, where some background pixels are included.

- Which *semantic video concept interpretation model* can be used to bridge the semantic gap?

Unlike traditional relational databases, video documents are generally unstructured. In order to support more efficient video database management in our system, the principal video shots in database are classified into a set of multi-level manageable units (i.e., semantic medical concept nodes) as shown in Fig. 8. In order to build this multilevel video database management structure, we have to face two critical problems.

- How many *levels* should be included and how many *nodes* should be used at each level?
- How should the *model* for each database node be determined for decision-making (i.e, video classification and video retrieval)?

In this paper, we have proposed a novel framework to organize the large-scale video collections according to a certain domain-dependent concept hierarchy, thus the database management structure (number of levels and number of nodes at each level) is derived from the concept hierarchy for a certain medical education video domain. The concept hierarchy defines the contextual and logical relationships between a upper semantic concept cluster (i.e., high-level database manageable unit) and its relevant deeper semantic medical concepts (i.e., sub-level database management units) [58]. The deeper the level of the concept hierarchy, the narrower the coverage of the subjects, thus the database manageable units at the deeper level can represent more specific subjects of a video. On the other hand, the database manageable units at the upper level can cover more distinct subjects of videos. In our current works, the deepest level of the concept hierarchy (i.e., leaf nodes of the database) is defined as the domain-dependent ***elementary semantic medical concepts***.

To classify the principal video shots into the most relevant semantic medical concept nodes, we have also proposed a novel multimodal video context integration model for semantic medical concept interpretation via flexible mixture model as shown in Fig. 9. The class distribution of the principal video shots that are implicitly related to the elementary semantic medical concept C_j is approximated by using a flexible mixture model with κ Gaussian functions

$$P\left(X, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right) = \sum_{i=1}^{\kappa} P\left(X \mid S_i, \omega_{s_i}, \theta_{s_i}\right) P(S_i) \quad (1)$$

where κ indicates the optimal number of Gaussian functions, $\Theta_{c_j} = \{\theta_{s_i}, i = 1, \ldots, \kappa\}$ is the set of the parameters (i.e., mean and co-variance) for these Gaussian functions, $\omega_{c_j} = \{\omega_{s_i}, i = 1, \ldots, \kappa\}$ is the set of the relative weights among these Gaussian functions, $\omega_{s_i} = P(S_i)$ is the relative weight for the ith Gaussian function, and $X = (x_1, \ldots, x_n)$ is the n-dimensional multimodal perceptual features which are used for representing the relevant principal video shots. For example, five different types of concept-sensitive principal video shots (i.e., principal video shots consist of the multimodal salient objects such as human faces, slides, text titles, sketch, and human speech) are explicitly related to the elementary semantic medical concept "lecture presentation." The data distribution for each type of these relevant concept-sensitive principal video shots is approximated by using multiple mixture Gaussian functions.

The fundamental assumptions of our flexible mixture model are: 1) there is a many-to-one correspondence between mixture Gaussian functions and different types (classes) of various principal video shots and 2) different types (classes) of various principal video shots are independent in their multimodal perceptual feature space. For a certain semantic medical concept, the optimal number of mixture Gaussian functions and their relative weights are acquired automatically through a machine learning process. Using the flexible mixture model for probabilistic semantic medical concept interpretation enables to remain the variability (heterogeneity) among various semantic medical concepts, thus it will offer a number of additional theoretical advantages.

IV. Semantic Video Classification

As described in Figs. 8 and 9, our hierarchical video-classification framework includes two major steps.

1) **First Classification:** classifying the principal video shots into the most relevant elementary semantic medical concepts.
2) **Second Classification:** assigning the principal video shots to the relevant high-level semantic concept clusters according to a certain domain-dependent concept hierarchy.

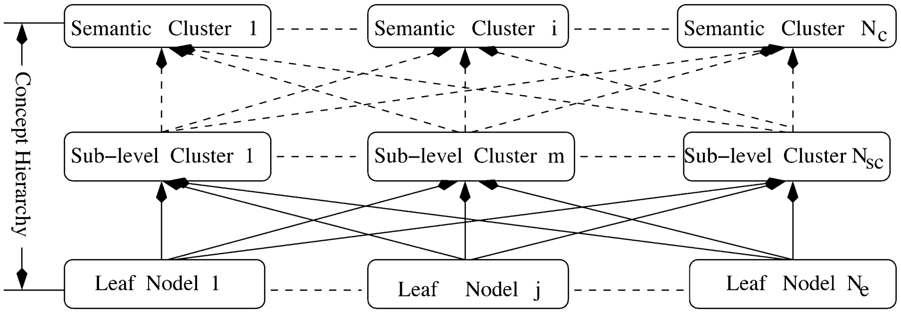


Fig. 8. Proposed hierarchical video database model, where the subcluster may consist of several levels according to the domain-dependent concept hierarchy.

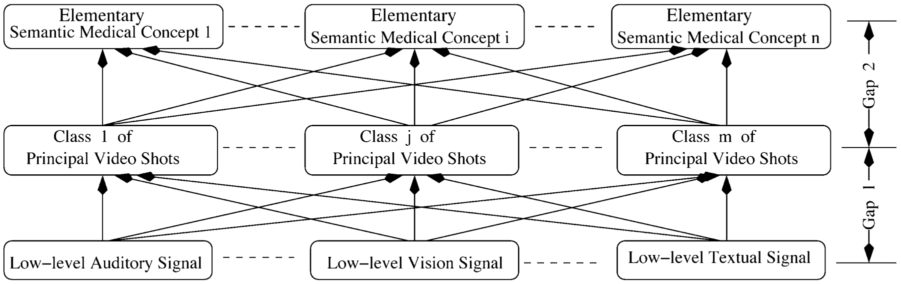

Elementary Semantic Medical Concept 1
Elementary Semantic Medical Concept i
Elementary Semantic Medical Concept n
Class 1 of Principal Video Shots
Class j of Principal Video Shots
Class m of Principal Video Shots
Low−level Auditory Signal
Low−level Vision Signal
Low−level Textual Signal
Gap 2
Gap 1

Fig. 9. Composition relationships between the elementary semantic video concepts and the relevant concept-sensitive principal video shots.

To enable more effective semantic medical video classification, the central goal of this paper is to automatically determine the optimal *multimodal context integration model* (i.e., flexible mixture model). We use *one-against-all* rule to label the training samples $\Omega_{c_j} = \{X_l, C_j(X_l) \,|\, l = 1, \ldots, N\}$: *positive samples* for a certain elementary semantic medical concept C_j and others are *negative samples*. Each labeled sample is a pair $(X_l, C_j(X_l))$ that consists of a set of n-dimensional multimodal perceptual features X_l and the semantic label $C_j(X_l)$ for the corresponding sample.

The posterior probability $P(C_j \,|\, X, \kappa, \omega_{c_j}, \Theta_{c_j})$, that a principal video shot with the multimodal perceptual features X can be assigned to the elementary semantic medical concept C_j is determined by a Bayesian framework. However, the traditional classifier induction techniques only estimate the Gaussian parameters Θ_{c_j} and the relative weights ω_{c_j} by using *maximum likelihood* (ML) criterion but ignore the estimation of the optimal model structure κ by using a fixed number of mixture Gaussian functions. On the other hand, the classification accuracy (posterior probability) $P(C_j \,|\, X, \kappa, \omega_{c_j}, \Theta_{c_j})$ is implicitly related to both the likelihood and the optimal model structure κ. If the given Gaussian mixture model does not match the real class distribution, a better estimate of the likelihood may not correspond to a higher classification accuracy $P(C_j \,|\, X, \kappa, \omega_{c_j}, \Theta_{c_j})$. Instead of using ML criterion, we use *maximum a posterior probability* (MAP) as the criterion for classifier induction, as follows:

$$\left(\hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right) = \arg\min \left\{ -\sum_{l=1}^{N} \log P\left(C_j \;\middle|\; X_l, \kappa, \omega_{c_j}, \Theta_{c_j}\right) \right\}. \tag{2}$$

The MAP estimation can be achieved automatically by using the expectation-maximization (EM) algorithm [83]–[87]. Unfortunately, the EM estimation of κ is not well defined. *Minimum description length* (MDL) criterion has been widely used to determine the optimal model structure (i.e., the optimal number κ of mixture Gaussian functions) by penalizing the complex model candidates with a large κ [48]. However, determining the optimal model structure by using MDL may not be appropriate and our main concern for semantic video classifcation is to achieve higher classification accuracy not just to minimize the description length.

To estimate the optimal flexible mixture model, we propose an ***adaptive EM algorithm*** by integrating feature selection, parameter estimation and model selection (i.e., selecting the optimal number κ of Gaussian functions) seamlessly in a single algorithm and it takes the following steps.

Step 1) The class distribution of various principal video shots, that are explicitly related to the elementary semantic medical concept C_j, is approximated by using a flexible mixture model. The data distribution for a certain type (class) of principal video shots is approximated by using multiple Gaussian functions. Thus the number of mixture Gaussian functions is initially set as $\kappa = m + 1$, where m is the total number of different types (classes) of various principal video shots that are explicitly related to the semantic medical concept C_j (i.e., m is obtained from the domain knowledge given by our medical consultants). One more Gaussian function is added for the hidden video patterns.

Step 2) To hold the many-to-one correspondence assumption, the optimal number κ of mixture Gaussian functions is adapted to the underlying class distributions of various principal video shots that are explicitly related to the elementary semantic medical concept C_j.

To determine the most discriminating features for representing the elementary semantic medical concept C_j, a feature subset with large discrimination power is selected by making the intra-concept distance small but the inter-concept distance large. Based on a number of labeled positive and negative samples, this discriminative feature subset $\hat{X}_{c_j}$ is determined automatically from the intersection of the intra-concept and inter-concept distance distributions

$$\hat{X}_{c_j} = \left\{ \arg\min \sum_{l=1}^{N-1} \sum_{k=l+1}^{N} \delta_{lk} \frac{D_{lk}}{N_s} \right\} \cap \left\{ \arg\max \sum_{l=1}^{N-1} \sum_{k=l+1}^{N} (1-\delta_{lk}) \frac{D_{lk}}{N_v} \right\} \quad (3)$$

where $\delta_{lk} = 1$ iff $C_j(X_l) \equiv C_j(X_k)$, else $\delta_{lk} = 0, D_{lk}$ is the similarity distance between a pair of labeled positive and negative samples X_l and X_k. $N_s = \sum_{l=1}^{N-1} \sum_{k=l+1}^{N} \delta_{lk}$ and $N_v = \sum_{l=1}^{N-1} \sum_{k=l+1}^{N} (1-\delta_{lk})$ are the numbers of labeled sample pairs for the positive and negative cases.

To hold the independence assumption, linear discriminant analysis is performed to obtain a transformed feature space such that the independence among different classes of various principal video shots can be maximized [87]

$$W = \arg\max \left\{ \frac{|W^T S_b W|}{|W^T S_w W|} \right\}, \quad \hat{Y}_{c_j} = W^T \hat{X}_{c_j} \quad (4)$$

where S_w is the intra-concept scatter matrix and S_b is the inter-concept scatter matrix, W is the feature transformation matrix, $\hat{X}_{c_j}$ is the set of the original multimodal perceptual features, and $\hat{Y}_{c_j}$ is the set of the representative features in the transformed feature space.

Linear discriminant analysis has reduced the obscuring noise (i.e., irrelevant multimodal perceptual features with less important influences to the relevant elementary semantic medical concept) and has discovered a more expressive feature subset by using a linear combination of the original multimodal perceptual features. This linear feature transformation also represents the video contents more compactly in a transformed features space where the data are clustered and easier to select more accurate model structure. Our experimental results have confirmed that using linear discriminant analysis for feature transformation not only increases the classification accuracy (i.e., decrease the misclassification ratio), but also dramatically reduces the optimal number of principal Gaussian functions and the amount of labeled samples that are needed for accurate classifier training (shown in Figs. 10 and 11).

Step 3) The traditional EM algorithm is used to estimate the parameters for the given κ Gaussian functions iteratively [83]–[87]. The *E-step* calculates the probabilistic labels (concept membership) for the training samples by using the current estimate of $\omega_{c_j}, \Theta_{c_j}$. The *M-step* calculates a new estimate for $\hat{\omega}_{c_j}, \hat{\Theta}_{c_j}$ by using all the labeled samples. After a point of (local) maximum is reached, a weak Bayesian classifier is built by using the estimated parameters. The performance of this weak Bayesian classifier is obtained by testing a small number of labeled samples that are not used for classifier training. If the average performance of this weak classifier is good enough, $P(C_j \,|\, \hat{Y}_{c_j}, \kappa, \omega_{c_j}, \Theta_{c_j}) \geq \delta_1$, go to step 6). Otherwise, go to step 4).

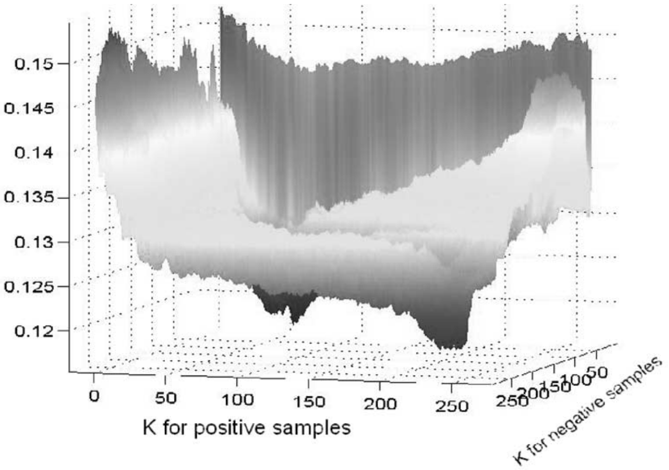


Fig. 10. Surface of misclassification ratio (i.e., missing-recall) for skin classification with different number of mixture Gaussian components, where the original perceptional features (i.e., without KLT) is used and thus multiple local minimum points appear and a bigger optimal number of mixture Gaussian components $\kappa = 216$ is obtained.

Step 4) A new Gaussian component, $P(\hat{Y}_{c_j} \,|\, S_{\kappa+1}, \omega_{s_{\kappa+1}}, \theta_{s_{\kappa+1}})$, is added to the flexible mixture model with the relative weight $\omega_{\kappa+1}$. The class distribution of the principal video shots that are implicitly related to the elementary semantic medical concept C_j is refined as

$$\begin{aligned} P\left(\hat{Y}_{c_j}, C_j, \kappa+1, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right) &= \omega_{\kappa+1} P\left(\hat{Y}_{c_j} \,\middle|\, S_{\kappa+1}, \omega_{s_{\kappa+1}}, \theta_{s_{\kappa+1}}\right) \\ &\quad + (1-\omega_{\kappa+1}) P(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}). \end{aligned} \quad (5)$$

The traditional EM algorithm is then used to estimate the Gaussian parameters $\hat{\Theta}_{c_j}$ and the relative weights $\hat{\omega}_{c_j} = \{\omega_1, \ldots, \omega_{\kappa+1}\}$ for $\kappa + 1$ Gaussian functions. The Kullback–Leibler distance Δ is used to quantify the "closeness" between two probability distributions $P(\hat{Y}_{c_j}, C_j, \kappa+1, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$ and $P(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j})$. The Kullback–Leibler distance is calculated as [88]

$$\begin{aligned} \Delta = \int & P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right) \\ & \times \log \frac{P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right)}{P\left(\hat{Y}_{c_j}, C_j, \kappa+1, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right)} dY. \end{aligned} \quad (6)$$

Step 5) If $\Delta \leq \delta_2$ or the iteration times $(\kappa - m) \geq \delta_3$, go to step 6). Otherwise, one more relevant feature F is added to $\hat{X}_{c_j}$ and linear discriminant analysis is performed on $\hat{X}_{c_j} \cup F$ to obtain a new representative feature set $\hat{Y'}_{c_j}$. This additional feature F is selected by maximizing the posterior probability $P(C_j \,|\, \hat{Y'}_{c_j}, \kappa, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$. If the classifier accuracy with one more feature F is decreased, $P(C_j \,|\, \hat{Y}_{c_j}, \kappa, \omega_{c_j}, \Theta_{c_j}) \geq P(C_j \,|\, \hat{Y'}_{c_j}, \kappa, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$, go to step 6). Otherwise, the "closeness" Δ between two distributions $P(\hat{Y'}_{c_j}, C_j, \kappa, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$ and $P(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j})$ is calculated by using (6). If $\Delta \leq \delta_2$, set $\hat{X}_{c_j} = \hat{X}_{c_j} \cup F$, go back step 3.

Step 6) Output mixture Gaussian parameters κ, Θ_{c_j}, and ω_{c_j}.

We have also achieved a theoretical justification for the convergence of the proposed *adaptive EM algorithm*. In our proposed *adaptive EM algorithm*, the parameter spaces for the two approximated models that are estimated incrementally have the following relationship:

$$\begin{cases} \hat{\kappa} = \kappa + 1 \\ \hat{\omega}_{c_j} = \left\{ \left(1 - \omega_{s_{\kappa+1}}\right) \omega_{c_j}, \omega_{s_{\kappa+1}} \right\} \\ \hat{\Theta}_{c_j} = \left\{ \Theta_{c_j}, \theta_{s_{\kappa+1}} \right\} \end{cases}$$
$$\begin{cases} \omega_{c_j} = \{\omega_1, \ldots, \omega_{s_\kappa}\} \\ \Theta_{c_j} = \{\theta_1, \ldots, \theta_{s_\kappa}\} \end{cases} \tag{7}$$
$$\begin{aligned} P&\left(\hat{Y}_{c_j}, C_j, \kappa + 1, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right) \\ &= \omega_{\kappa+1} P\left(\hat{Y}_{c_j} \,\middle|\, S_{\kappa+1}, \theta_{s_{\kappa+1}}\right) \\ &\quad + (1 - \omega_{\kappa+1}) P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right). \end{aligned}$$

The real class distribution $P(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j})$ is defined as the underlying optimial model that our proposed *adaptive EM algorithm* should converge to. Thus, we put the real class distrbution $P(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j})$ as the first augument in the following discussion. Given the approximated class distributions $P(\hat{Y}_{c_j}, C_j, \hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$ and $P(Y_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j})$ that are estimated sequentially, the Kullback–Leibler distances, between the real class distribution $P(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j})$ and the approximated class distrbutions, is calculated as

$$\begin{aligned} \Delta_1 &= \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log \frac{P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right)}{P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right)} dY \\ \Delta_2 &= \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log \frac{P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right)}{P\left(\hat{Y}_{c_j}, C_j, \hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right)} dY \end{aligned} \tag{8}$$

where the Kullback–Leibler distances, Δ_1 and Δ_2, are always nonnegative [88].

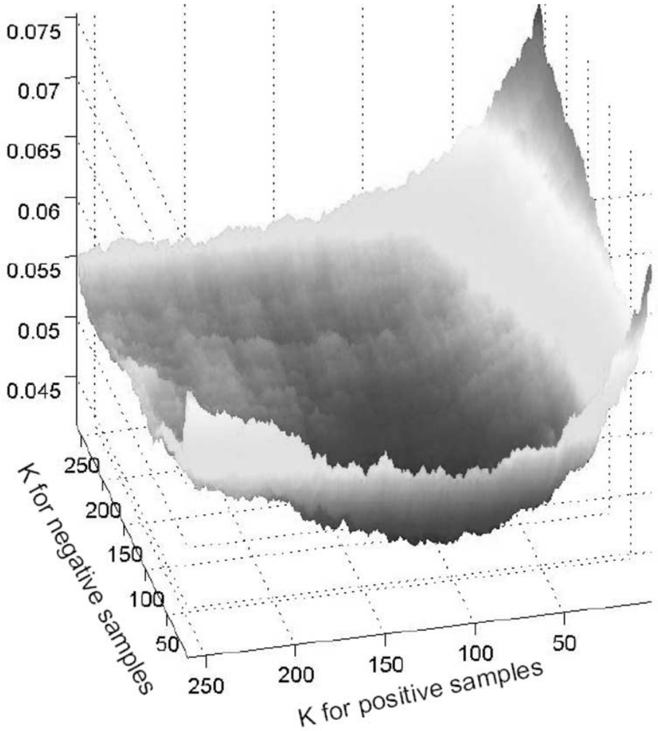


Fig. 11. Surface of misclassification ratio (i.e., missing-recall) for skin classification with different number of mixture Gaussian components, where KLT is used to derive more expressive feature subset and thus only few local minimum point appears and a smaller optimal number of principal Gaussian components $\kappa = 76$ is obtained.

Thus, the difference D between Δ_1 and Δ_2 is able to reflect the convergence of our *adaptive EM algorithm*. The difference D is calculated as

$$\begin{aligned} D &= \Delta_1 - \Delta_2 \\ &= \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log \frac{P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right)}{P\left(\hat{Y}_{c_j}, C_j, \hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right)} dY \\ &\quad - \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log \frac{P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right)}{P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right)} dY \\ &= - \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log P\left(\hat{Y}_{c_j}, C_j, \hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right) dY \\ &\quad + \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right) dY \\ &= \int P\left(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j}\right) \\ &\qquad \times \log \frac{P\left(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right)}{P\left(\hat{Y}_{c_j}, C_j, \hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}\right)} dY. \end{aligned} \tag{9}$$

By considering the implicit relationships among $\kappa, \hat{\kappa}, \kappa^*, \omega_{c_j}, \hat{\omega}_{c_j}, \omega^*_{c_j}, \Theta_{c_j}, \hat{\Theta}_{c_j}, \Theta^*_{c_j}$, and $P(\hat{Y}_{c_j}, C_j, \kappa^*, \omega^*_{c_j}, \Theta^*_{c_j})$, $P(\hat{Y}_{c_j}, C_j, \hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j})$, $P(\hat{Y}_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j})$, we can prove

$$\begin{cases} D \leq 0, & \text{if } \hat{\kappa}, \kappa \leq \kappa^* \\ D > 0, & \text{if } \hat{\kappa}, \kappa > \kappa^* \end{cases}. \tag{10}$$

Hence, our *adaptive EM algorithm* can reduce the divergence sequentially and thus it can be converged to the underlying optimal model incrementally. By selecting a suitable threshold δ_2, we can also control its convergence rate. Our experimental results also match our theoretical proof convincingly as shown in Fig. 12. Before our adaptive EM algorithm converges to the optimal model, adding more Gaussian functions will increase the classifier's performance, while after our adaptive EM algorithm converges to the optimal model, adding more Gaussian functions will decrease the classifier's performance.

After the semantic video classifiers for the N_e elementary semantic medical concepts are in place, the classifier training for the high-level semantic concept clusters is achieved by two steps.

1) The flexible mixture model for a certain high-level semantic concept cluster is determined by using a general combination of N_κ mixture Gaussian functions for the relevant elementary semantic medical concepts, that are under the corresponding semantic concept cluster node in a certain domain-dependent concept hierarchy. To determine the optimal flexible mixture model for a certain semantic concept cluster, the mixture Gaussian functions for the relevant elementary semantic medical concepts with less prediction power are removed iteratively.
2) The weights among the residual mixture Gaussian functions are then refined automatically by learning from the labeled training samples.

Once the hierarchical video classifier is in place, the task of semantic medical video classification can be summarized as follows. The principal video shots and their multimodal perceptual features are first extracted automatically from the test medical video clips. Linear discriminant analysis is then used to obtain more representative feature subset for video content representation and indexing. Given an unlabeled principal video shot S_i and its transformed feature values Y_i, it is finally assigned to the best matching elementary semantic medical concept C_j that corresponds to the maximum posterior probability

$$P(C_j \mid Y_i, \Theta) = \frac{P(C_j) P\left(Y_i, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right)}{\sum_{j=1}^{N_e} P(C_j) P\left(Y_i, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}\right)} \tag{11}$$

where $\Theta = \{\omega_{c_j}, \Theta_{c_j}, j = 1, \ldots, N_e\}$ is the set of the mixture Gaussian parameters and relative weights for the classifier, $\omega_{c_j} = P(C_j)$ is the prior probability (i.e., relative weight) of the elementary semantic medical concept C_j in the database for the labeled samples. The principal video shot S_i is then assigned into the relevant high-level semantic concept clusters. Our semantic medical video-classification results at the elementary semantic medical concept level are given in Figs. 13 and 14.

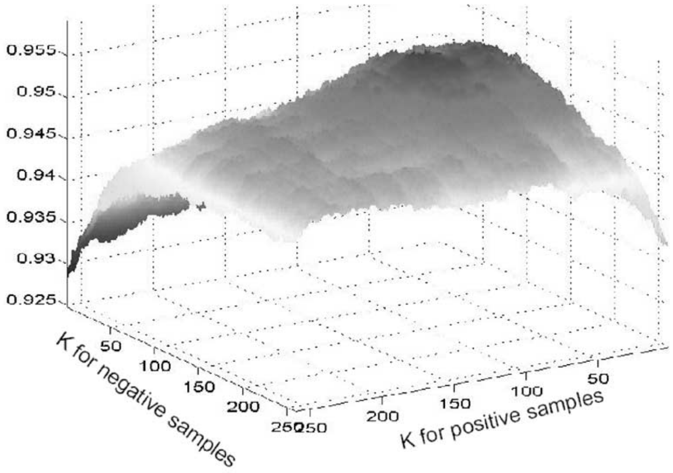


Fig. 12. Classification accuracy increases when more mixture Gaussian components are added before it reaches the optimal model $\kappa = 76$. The classification accuracy decreases when more mixture Gaussian components are added after it is bigger the optimal model $\kappa = 76$.

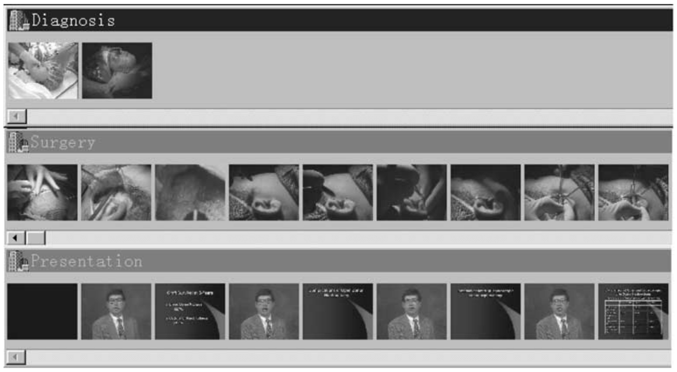


Fig. 13. Principal video shot classification results for a test video which consists of three semantic medical concepts: "Presentation," "Traumatic Surgery," and "Diagnosis."

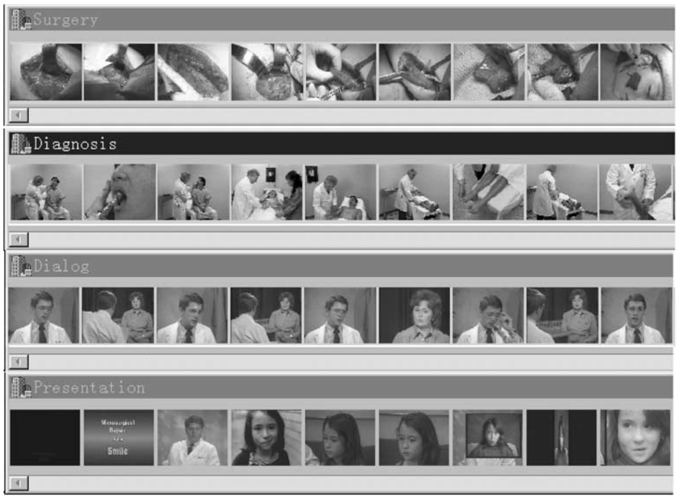


Fig. 14. Principal video shot classification results for a test video which consists of four semantic medical concepts: "Traumatic Surgery," "Dialog," "demo presentation," and "Diagnosis."

It is important to note that once an unlabeled principal video shot is classified, the semantic labels for the relevant elementary semantic medical concept and the high-level semantic concept clusters that it is assigned to become the semantic labels for the corresponding principal video shot. Moreover, the membership between the principal video shots and the elementary semantic medical concepts could be highly nonlinear with different probabilities. One certain principal video shot may consist of multiple types (classes) of various multimodal salient objects, thus it can be classified into multiple elementary semantic medical concepts when these multimodal salient objects are implicitly related to different elementary semantic medical concepts. Thus, multiple semantic labels for the relevant elementary semantic medical concepts and their relevant high-level semantic concept clusters become the semantic labels for the corresponding principal video shot with different probabilities. Our probabilistic semantic video-classification and annotation algorithm could remain the variability (heterogeneity) within the same semantic medical concept and thus offer a number of additional theoretical advantages compared with other classification techniques with a binary "hard" decision. This probabilistic video annotation technique is very attractive to enable semantic video retrieval such that the naive users will have more flexibility to specify their query concepts via different keywords. One certain medical video clip may consist of multiple types (classes) of various principal video shots, the semantic labels for the relevant semantic medical concepts are finally taken as the semantic labels for the corresponding medical video clip. Such ***automatic probabilistic video annotation*** via semantic classification will make it possible for semantic video retrieval via keywords.

V. Concept-Oriented Video Database Organization and Access

After the elementary semantic medical concepts and the relevant semantic concept clusters are obtained, we turn our attention to use them to provide concept-oriented video database indexing, retrieval and browsing.

A. Concept-Oriented Video Database Indexing

After all the unlabeled principal video shots are classified into the relevant elementary semantic medical concept nodes and the high-level semantic concept clusters, these elementary semantic medical concept nodes become the leaf nodes of the video database, upon which the nonleaf nodes of the video database can be constructed as the high-level semantic concept clusters. The parent–child relationships in the database indexing structure correspond to the underlying inter-level relationships in a certain domain-dependent concept hierarchy.

To support more effective video database access, it is necessary to find a good way to characterize the database nodes (i.e., semantic medical concept nodes) jointly by using their class distributions in the high-dimensional feature space, visual summaries and semantic labels. Thus, the following novel techniques are used to support statistical video database indexing.

- We use the underlying flexible mixture model to characterize and index the statistical property of each database node (i.e., semantic medical concept node) in its discriminant feature subspace. The underlying flexible mixture model, that is used for semantic medical concept modeling and classification, is able to approximate the class distribution for the relevant concept-sensitive principal video shots with a certain degree of accuracy.
- Each database node (i.e., semantic medical concept node) is jointly described by the semantic label (i.e., keyword), visual summary, and statistical properties of the class distribution for the relevant concept-sensitive principal video shots in their discriminant feature subspace.

Thus, the following parameters will be used to represent a database node (i.e., semantic medical concept node) Q:

$$\begin{aligned}&\text{semantic label } L_Q, \text{ feature subset}: \ X_Q\\&\text{flexible model parameters}: \ \Theta_Q, \omega_Q, \kappa_Q\\&\text{visual summary}: \ V_Q\end{aligned}\qquad(12)$$

where L_Q is the semantic label for the database node (i.e., semantic medical concept node) Q, Θ_Q, ω_Q, and κ_Q are the model parameters that are used for semantic medical concept interpretation and indexing, X_Q is the feature subset that is used for medical content representation, and V_Q is the visual summary for the database node Q. Based on this proposed joint database node representation and indexing approach, more effective query concept specification and video database access framework can be supported.

B. Hierarchical Semantic Video Summarization

Most existing CBVR systems do not support hierarchical browsing [10]. Users, however, are not only interested in searching for specific video clips (e.g., query-by-example). They would also like to browse and navigate through the video databases. A key issue to hierarchical video browsing is whether the visual summaries at different database nodes and the hierarchical relationships among different database levels make sense to the user. Such requirements have created great demands for effective and efficient approaches to organize the visual summaries through a certain domain-dependent concept hierarchy [54]–[58].

Our hierarchical video-classification framework has resulted in a hierarchical concept-oriented video organization in a database and thus more effective concept-oriented video browsing can be supported. To enable concept-oriented video browsing, we have developed a novel semantic-sensitive video summarization technique and it includes two parts.

1) ***Semantic summary at video clip level:*** Our semantic video-classification technique is able to support efficient context understanding for a certain medical video clip; thus, two heuristic rules are used to generate the concept-sensitive visual summary automatically: a) the principal video shots, that consist of the most frequent semantic medical concept in a certain medical video clip, are selected as the concept-sensitive visual summary for the corresponding medical video clip and b) as mentioned above, one certain principal video shot could be implicitly related to multiple elementary semantic medical concepts. The principal video shots, that consist of multiple elementary semantic medical concepts and thus provide a compact but sufficient representation of the original medical contents, are also selected as the

concept-sensitive visual summary for the corresponding medical video clip.

2) ***Semantic summary at semantic concept level:*** The *icon principal video shots* (i.e., most informative principal video shots) for a certain database node (i.e., semantic medical concept node) are obtained by using *independent component analysis* [72]–[74]. The icon principal video shots are treated as the concept-sensitive visual summary for the corresponding semantic medical concept node.

Our multiple-level semantic video summarization results are given in Figs. 15 and 16.

C. Hierarchical Video Retrieval

To support more effective video database access, it is very important to address two key problems. How can the video database system provide an intuitive approach for the naive users to specify their query concepts effectively? How can the underlying query processor evaluate the users' query concepts effectively? Thus, it is very important to integrate three video access approaches (i.e., query by exmaple via online relevance feedback, query by keywords, and concept-oriented video browsing) in a unified framework.

1) Intuitive Query Concept Specification: To provide an intuitive approach for the naive users to specify their query concepts, we have proposed the following.

a) **Query Concept Specification via Browsing:** Our proposed concept-oriented database organization technique can support the users to get a good idea of the video context quickly through browsing the visual summaries for the semantic medical concept nodes. After the naive users browse the visual summaries, they can pick up one or multiple video clips as their query examples.

b) **Query Concept Specification via Keywords:** Keywords are most useful for the naive users to specify their query concepts and communicate with the CBVR systems at the semantic level. However, the keywords, which are used for achieving automatic video annotation, may be too abstract to describe the details of video contexts. The query results, that are initially obtained by keywords, may include a large number of semantically similar video clips sharing the same semantic medical concept node. However, the naive users can specify their query concepts by selecting the most suitable video clips as their query examples in the browsing procedure.

c) **Query Concept Specification via Pattern Combinations:** Our proposed semantic video analysis and semantic medical concept interpretation techniques have also provided a query concept interpretation language for the naive users to specify their query concepts by using the concept-sensitive principal video shots (i.e., building blocks of semantic medical concepts) and the underlying semantic medical concept interpretation models. Based on the underlying semantic medical concept interpretation models (i.e., query concept interpretation language), the naive users can interpret their query concepts easily and effectively by using the general combinations of the preattentive concept-sensitive principal video shots that are explicitly relevant to their query concepts (one example is shown in Fig. 17).

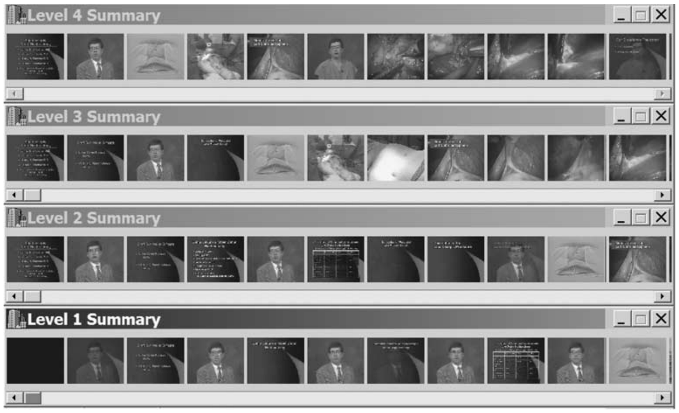


Fig. 15. Multiple-level semantic video summarization results.

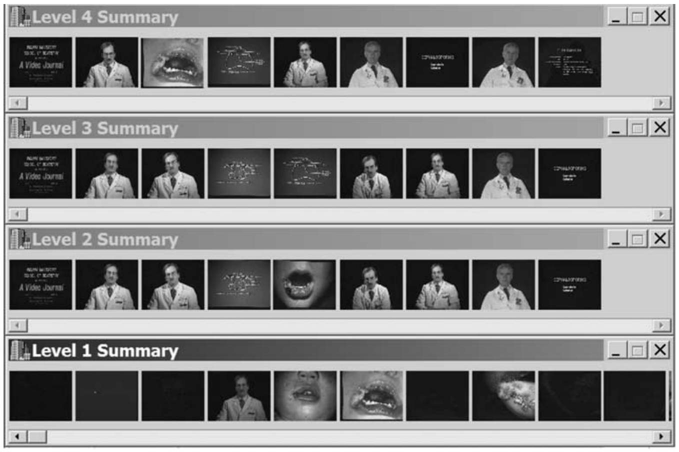


Fig. 16. Multiple-level semantic video summarization results.

2) Query Concept Evaluation for Query-by-Example: After the query concepts are interpreted by the selected video clips, similarity search is performed through the underlying video database indexing structure so that the most similar video clips can be obtained. The naive users can then label these retrieved video clips as *relevant* or *irrelevant* according to their subjectivity [59]–[65]. Rocchio's formula could possibly be used to determine the new query vector for the next iteration. However, Rocchio's formula cannot predict the most suitable search direction for the next iteration, thus there is no guarantee that the search results can be improved progressively and be converged to the "optimal" target quickly [64].

To solve this convergence problem, we have developed an effective scheme by combining an informative sampling technique with an optimal search direction prediction method to achieve more effective online relevance feedback. The scheme takes the following major steps.

- **Informative Sample Selection:** The irrelevant video data samples, which are obtained in a previous query and located in the nearest neighbor sphere of the current query seed, are used for shrinking the sampling area for the current query iteration [64]. Specifically, the nearest neighborhoods of these irrelevant samples (shown as dash

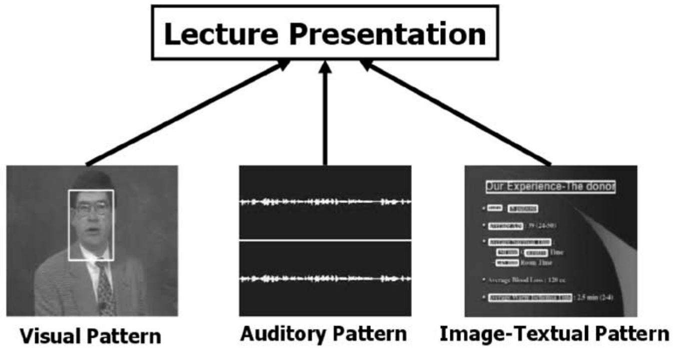


Fig. 17. Query concept specification via a general combination of the preattentive concept-sensitive principal video shots.

circles in Fig. 18) are taken out from the sampling area of the current query iteration. The most informative video clips residing in the shrunk sampling area are subsequently displayed to the naive users as the seed for next iteration of query [64], [65] (see Fig. 18).

- **Best Search Direction Prediction:** Relevance feedback with the user in the loop can improve the the query results subsequently, and thus the nearest neighbor spheres for subsequent query iterations are be reduced in size repeatedly, as shown in Fig. 18. The best search direction for the next query iteration predicted by combining such iterative nearest neighbor sphere reduction with the above introduced technique for informative sampling. Similarity search can converge quickly with the prediction of the best search direction.
- **Query Refinement:** Only the previous query vector and the positive samples are used to determine the new query vector for the next iteration based on the revised Rocchio's formula

$$Q = \alpha Q' + \beta \left(\frac{1}{N_p} \sum_{i \in D_p} V_i \right) \tag{13}$$

where Q and Q' are the *new query vector* for the next iteration and the *current query vector* respectively, α and β are some suitable constants, V_i denotes the feature vectors for the positive samples, D_p is the set of the positive samples, and N_p is the cardinality of D_p. For each query concept, only the discriminating perceptual features are used for generating the new query vector. After the query concept and the relevant discriminating feature subspace are refined, we have developed a Bayesian framework for selecting the matching candidates.

3) Query Concept Evaluation for Query-by-Patterns: After the query concepts are initially specified by the naive users with a general combination of the preattentive principal video shots, our query processor can first interpret the users' query concepts with multiple mixture Gaussian functions that are used to approximate the class distrbutions of the selected principal video shots. The weights among multiple mixture Gaussian functions for these selected principal video shots can be pre-defined by the users or be learned by the system incrementally.

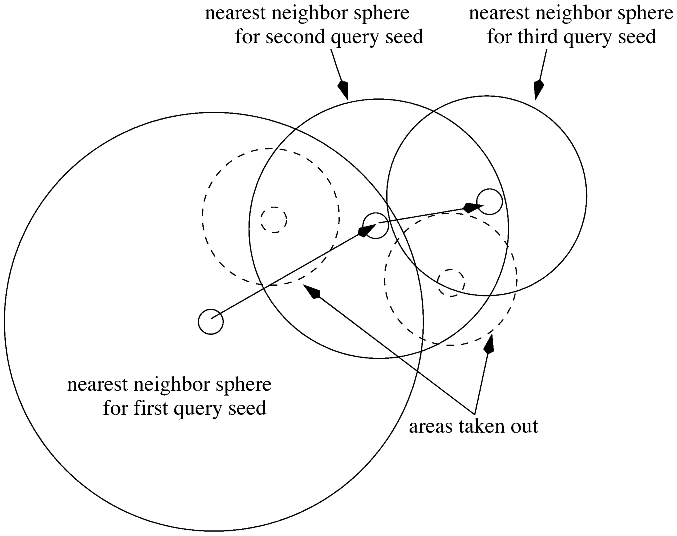

nearest neighbor sphere for second query seed
nearest neighbor sphere for third query seed
nearest neighbor sphere for first query seed
areas taken out

Fig. 18. Proposed adaptive nearest neighbor search and informative sampling scheme.

In order to capture the users' subjectivity more effectively, it is very important to adapt the query processor to the potential concept drift [62], [63]. For semantic video retrieval, we focus on addressing the *gradual concept drift* and it can be induced by two factors: 1) the users' interpretation for a certain semantic medical concept changes gradually because of the appearance of *hidden video context* and 2) the users' interpretation for a certain semantic medical concept changes gradually because of the disappearance of *existing video context*. Based on this understanding, we have proposed an ***incremental EM algorithm*** to adapt the query processor to the gradual concept drift automatically.

To characterize the difference of the semantic medical concept interpretation along the time, a new time factor is represented explicitly in the flexible mixture model for semantic medical concept interpretation $P(X, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}, t)$ as follows:

$$P\left(X_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}, t\right) = \sum_{i=1}^{\kappa} P\left(X_{c_j}, C_j \mid S_i, \theta_{s_i}, t\right) \omega_{s_i}. \tag{14}$$

To detect the query concept drift over time, the Kullback–Leibler distance Δ is used to quantify the divergence between $P(Y_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}, t)$ and $P(Y_{c_j}, C_j, \kappa', \omega'_{c_j}, \Theta_{c_j}, t)$ by adding more training samples which are labeled recently by the users. The Kullback–Leibler distance is calculated as [88]

$$\Delta = \int P\left(Y_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}, t\right) \times \log \frac{P\left(Y_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}, t\right)}{P\left(Y_{c_j}, C_j, \kappa', \omega'_{c_j}, \Theta'_{c_j}, t\right)} \, dY \tag{15}$$

where the query concept model structure κ is fixed but the model parameters ω_{c_j} and Θ_{c_j} may be changed after adding latest new samples.

If $\Delta \geq \delta_2$, the gradual query concept drift is detected. To address the gradual query concept drift induced by the ***appearance of hidden video context***, our *adaptive EM algorithm* is used to generate a new query concept model and feature subset by adding more Gaussian functions iteratively, $\hat{\kappa}, \hat{\omega}, \hat{\Theta}_{c_j}, \hat{X}_{c_j}$.

If the gradual query concept drift is induced by the ***disappearance of existing video context***, one or more existing Gaussian functions with the least prediction power are removed from flexible mixture model. Our *adaptive EM algorithm* is performed to obtain a new query concept model and feature subset $\hat{\kappa}, \hat{\omega}_{c_j}, \hat{\Theta}_{c_j}, \hat{X}_{c_j}$ iteratively. If a mixture Gaussian function $P(Y_{c_j} \mid S_l, \omega_{s_l}, \theta_{s_l})$ is removed from the underlying flexible mixture model, the weights among the residual mixture Gaussian functions are then refined automatically by

$$P\left(Y_{c_j}, C_j, \kappa, \omega_{c_j}, \Theta_{c_j}, t\right) = \frac{1}{1-\omega_{s_l}} \sum_{i=1}^{\kappa-1} P\left(X \mid S_i, \omega_{s_i}, \theta_{s_i}\right) \omega_{s_i}, \quad i \neq l. \quad (16)$$

VI. Performance Analysis

Our experiments are conducted on two image/video databases: skin database (i.e., marked face database) from Purdue University and medical video database. The skin database consists of 1265 face images and 150 face images are selected as the labeled samples for classifier training. The medical video database includes more than 35 000 principal video shots that are obtained from 45 h of MPEG medical education videos, where 1500 principal video shots are selected as the training samples and labeled by our medical consultant.

A. Benchmark Matrics

The success of semantic video classifier depends on five major factors: 1) the ***effectiveness*** of the underlying video content representation framework; 2) the ***correction*** of the basic assumption that the real data distributions can be approximated by using mixture Gaussian functions; 3) the ***ability*** of the selected multimodal perceptual features to discriminate among various semantic medical concepts; 4) the ***significance*** of the classifier induction algorithm; and 5) the ***size*** of labeled samples and the ***relative size ratio*** between positive samples and negative samples.

Our algorithm and system evaluation works focus on:

- evaluating the performances of two major video content representation frameworks by using concept-insensitive "pure" video shots or concept-sensitive principal video shots;
- comparing the performance differences between our proposed probabilistic classification algorithms and other existing techniques, especially SVM because SVM was reported to be successful for high-dimensional "hard" binary classification;
- Comparing the performance differences for our proposed classification and feature subset selection algorithms by using different sizes of labeled samples and different relative size ratios between the positive samples and the negative samples.

The*first benchmark metric* is the *classification accuracy* (i.e., misclassification ratio versus classification accuracy ratio). The classification accuracy ρ and misclassification ratio $\bar{\rho}$ are defined as

$$\begin{cases} \rho = \frac{1+\nu+\gamma}{1+\nu+\gamma+\theta+\eta} \\ \bar{\rho} = \frac{1+\theta+\eta}{1+\nu+\gamma+\theta+\eta} \end{cases} \quad (17)$$

where ν is the set of true positive samples that are related to the corresponding semantic medical concept and classified correctly, γ is the set of true negative samples that are irrelevant to the corresponding semantic medical concept and classified correctly, θ is the set of false positive sample that are related to the corresponding semantic medical concept but misclassified, and η is the set of false negative samples that are irrelevant to the corresponding semantic medical concept but classified incorrectly.

The ***second benchmark metric*** is the *retrieval accuracy* (i.e., precision versus recall weighted by different retrieval purposes). The weighted precision ϱ and recall $\bar{\varrho}$ are defined as

$$\begin{cases} \varrho = \frac{1+2\lambda^2\tau}{1+2\lambda^2\tau+(1+\lambda^2)\xi} \\ \bar{\varrho} = \frac{1+2\beta^2\tau}{1+2\beta^2\tau+(1+\beta^2)\varepsilon} \end{cases} \quad (18)$$

where τ is the set of true positive samples that are relevant to the query concept and returned by a certain query correctly, ξ is the set of false negative samples that are irrelevant to the query concept but returned by a certain query incorrectly, ε is the set of false positive samples that are relevant to the query concept but not returned by a certain query correctly, and $\lambda \in [1, \infty)$ and $\beta \in [1, \infty)$ are the weighting parameters to specify the retrieval purposes by controlling the influences of false positive and false negative samples on ϱ and $\bar{\varrho}$. A large value of λ indicates that the users' retrieval purposes will focus on the total number of ture positive samples returned by the system. A large value of β indicates that the users' retrieval purposes will focus on obtaining more true positive samples but neglecting how many relevant false positive samples residing in the database. When $\lambda = 1$ and $\beta = 1$, ϱ and $\bar{\varrho}$ become the traditional precision and recall.

B. Implementation Issues

We have extracted a set of *multimodal perceptual features* to represent the principal video shots and enable more effective semantic video classification. The multimodal perceptual features include shot-based global visual features, object-based local visual features, shot-based auditory features, and shot-based image-textual features. The shot-based global visual features include 32-bin histograms of principal (dominant) colors and color variances within the same principal video shot, 9-bin edge histogram as the texture and structure feature. We did not include shot-based motion features because the motion features do not have strong impact for medical content representation and semantic medical video classification, this property for medical education videos is very different from that for other video domains such as news and films. The object-based local visual features include object density, dominant colors and variances, height-width ratio, Tamura texture features. We focus on the *shot-based image-textual features* rather than recognizing written image-text, the image-text segmentation outputs within the same principal video shot are integrated as a single bitmap for extracting the suitable shot-based image-textual features such as average length ratio between the length of the image-textual regions and the size of video frames, average

TABLE I
Average Performance (i.e., Classification Accuracy Ratio Versus Misclassification Ratio) of Our Semantic Video Classifier Based on Principal Video Shots

concepts	lecture presentation	traumatic surgery	dialog	diagnosis	gastroinstinal surgery
adaptive	79.6%	81.7%	78.5%	74.7%	80.6%
EM	20.8%	18.9%	21.8%	25.8%	9.5%
traditional	71.6%	72.4%	69.8%	67.2%	69.8%
EM	28.5%	28.1%	30.6%	33.4%	30.3%

TABLE II
Average Performance (i.e., Classification Accuracy Ratio Versus Misclassification Ratio) of Our Semantic Video Classifier Based on "Pure" Video Shots

concepts	lecture presentation	traumatic surgery	dialog	diagnosis	gastroinstinal surgery
adaptive	65.4%	66.9%	71.2%	59.8%	67.1%
EM	35.1%	33.5%	29.3%	41.1%	33.0%
traditional	58.4%	59.8%	62.4%	51.7%	58.6%
EM	42.1%	40.6%	37.2%	49.2%	41.5%

width ratio, and coverage ratio within a shot. We also focus on the *shot-based auditory features*, such as loudness, frequencies, pitch, fundamental frequency, and frequency transition ratio, rather than recognizing speech.

The thresholds for system implementation include: δ_1 for classification accuracy, δ_2 for the closeness between two data distributions, and δ_3 for the maximum iteration times. In our current implementation, we set $\delta_1 = 90.0\%$ for skin database and $\delta_1 = 80.0\%$ for medical video database. We set $\delta_2 = 0.075$ for defining the closeness of the data distributions that are estimated sequentially with different number of mixture Gaussian components. To control the iteration times for estimating the optimum number κ of mixture Gaussian components, we set $\delta_3 = 25$ for medical video classification (i.e., with KLT). For skin classification, $\delta_3 = 300$ if the original perceptual features are directly used for parameter estimation and model selection, $\delta_3 = 100$ if KLT is used for deriving a more expressive feature subset.

C. Performance Evaluation

Human faces in our database include various backgrounds and illuminations, thus we extract 32-bin HSV color histogram for each 3×3 image block. We have obtained very high classification accuracy 95.5% for the skin database. As shown in Figs. 10 and 11, the optimal numbers of mixture Gaussian components for positive and negative examples are selected with the highest classification accuracy. From Figs. 10 and 11, we have also found that our adaptive EM algorithm can be converged to the underlying optimal model as described by (10). After our adaptive EM algorithm converges to the underlying optimal model, adding more mixture Gaussian functions to the flexible mixture model will descrease the classifier performance. This experimental conclusion matches our theoretical proof in (10) for the convergence of our adaptive EM algorithm very well. One can also find that the optimal number κ of mixture Gaussian components for skin classification is very large because the face images for different illumination conditions are included in our skin database. In our experiments, we find that $\kappa = 76$ if Karhunen–Loeve transformation (KLT) is used for deriving more expressive feature subset and $\kappa = 216$ if the original perceptual features are directly used.

The average performance of our semantic medical video-classification technique is given in Tables I and II, they are obtained by averaging *classification accuracy* and *misclassification ratio* for the same semantic medical concept over 33 500 testing medical video clips. We have compared the performance differences of our semantic video classifier by using different video content charaterization and representation frameworks via principal video shots or "pure" video shots. We find that our semantic video classifier based on principal video shots has better performance than the same classifier that is based on "pure" video shots, because the multimodal perceptual features obtained from the principal video shots are more effective to discriminate among various semantic medical concepts.

We have also compared the performance differences of our classifier with and without KLT. The experimental results are given in Tables III. One can find that our semantic video classifier has better performance by performing KLT on the original perceptual multimodal features, because the KLT has reduced the obscuring noise (i.e., irrelevant multimodal perceptual features with less important influences to the relevant semantic medical concept) and discovered a more expressive feature subset by using a linear combination of the original high-dimensional perceptual features. This linear feature transformation represents video contents in a new features space where the data are clustered and easier to select the effective model structure of mixture Gaussian components. From

TABLE III
AVERAGE PERFORMANCE (I.E., CLASSIFICATION ACCURACY RATIO VERSUS MISCLASSIFICATION RATIO) OF OUR SEMANTIC VIDEO CLASSIFIER WITH AND WITHOUT KLT

concepts	lecture presentation	traumatic surgery	dialog	diagnosis	gastroinstinal surgery
without	79.6%	81.7%	78.5%	74.7%	82.1%
KTL	20.8%	18.9%	21.8%	25.8%	18.0%
with	86.8%	88.6%	85.8%	83.9%	88.9%
KLT	13.4%	11.5%	14.3%	16.4%	11.2%

TABLE IV
OPTIMAL NUMBERS κ OF GAUSSIAN COMPONENTS FOR FOUR SEMANTIC MEDICAL CONCEPTS WITH AND WITHOUT KLT

concepts	lecture presentation	traumatic surgery	dialog	diagnosis	gastroinstinal surgery
with	4	11	6	8	15
KLT					
without	26	46	33	39	52
KLT					

TABLE V
AVERAGE PERFORMANCE (I.E., CLASSIFICATION ACCURACY RATIO VERSUS MISCLASSIFICATION RATIO) FOR SEVERAL CLASSIFIERS WITH KLT

concept	lecture presentation	traumatic surgery	dialog	diagnosis	gastroinstinal surgery
adaptive	86.8%	88.6%	85.8%	83.9%	88.9%
EM	13.4%	11.5%	14.3%	16.4%	11.2%
SVM	83.5%	89.4%	80.5%	84.6%	89.3%
	16.6%	10.7%	18.6%	15.5%	10.8%
C4.5	68.5%	67.7%	68.3%	66.6%	68.1%
	27.6%	28.4%	27.8%	29.5%	32.0%

Tables III and IV, one can find that using KLT for feature transformation not only increases the classification accuracy (i.e., decreases the misclassification ratio) but also dramatically reduces the optimal number of principal Gaussian components. The optimal numbers κ of mixture Gaussian components for five semantic medical concepts with and without KLT in our test are given in Table IV.

We have also compared the performance differences between our classifier and other well-known classifiers such as SVM and C4.5. The test is performed on the same medical video data set by using the same video content characterization framework (i.e., via principal video shots). The test results are given in Table V. One can find that our classifier has better average performance as compared with other classifiers. The testing results have also shown that SVM is also successful for binary video classification; however, C4.5 is not a good choice for semantic video classification because hundreds of its inter-nodes (decision nodes) do not make sense to human beings.

The performance difference for our adaptive EM algorithm with different feature dimensions is given in Fig. 19. Theoretically, having more features should give us more discriminating power to support more accurate classifier training. However, more features will also make it very difficult to obtain the good estimates of many parameters for the classifier and thus adding more irrelevant features will also decrease the classifier accuracy, as shown in Fig. 19.

The search time T_e for our CBVR system is the sum of two times: the time T_s for comparing the relevant video clips in the database and the time T_r for ranking the relevant results. If no database indexing structure is used for organizing this search procedure, the total retrieval time is

$$T_e = T_s + T_r = N_T \cdot T_m + O(N_T \log N_T) \qquad (19)$$

where N_T is the number of videos in the databases, T_m is the basic time to calculate the feature-based similarity distance between two video clips, and $O(N_T \log N_T)$ is the time to rank N_T elements.

Our concept-oriented video database indexing structure can provide fast retrieval because only the relevant database management units are compared with the query example.

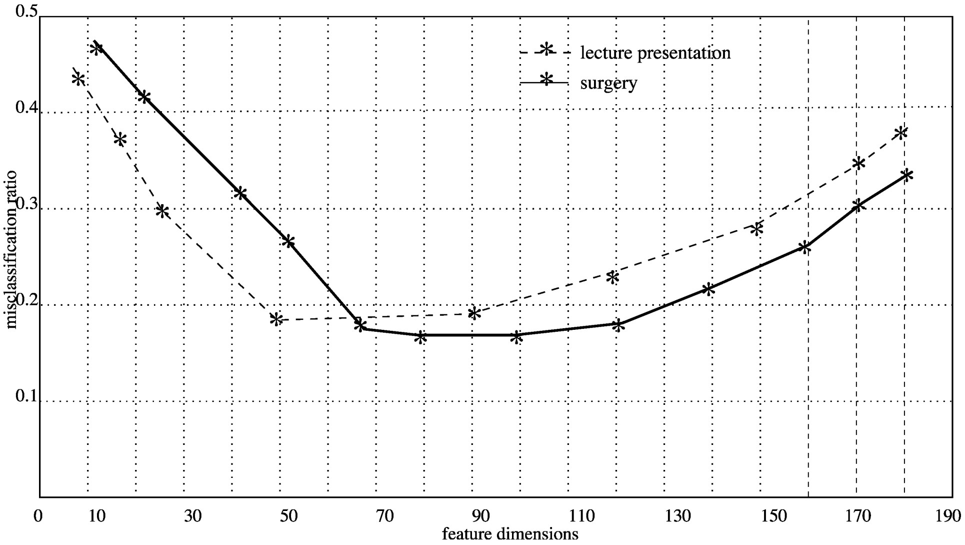


Fig. 19. Relationship between the misclassification ratio and the sizes of feature dimensions.

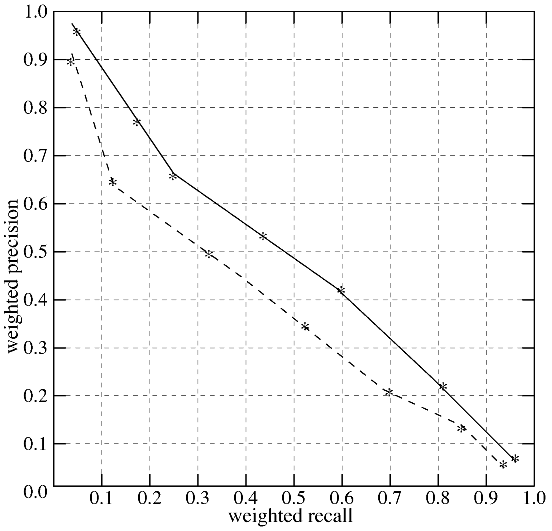


Fig. 20. Average performance of our query evaluation technique with different values of λ and β.

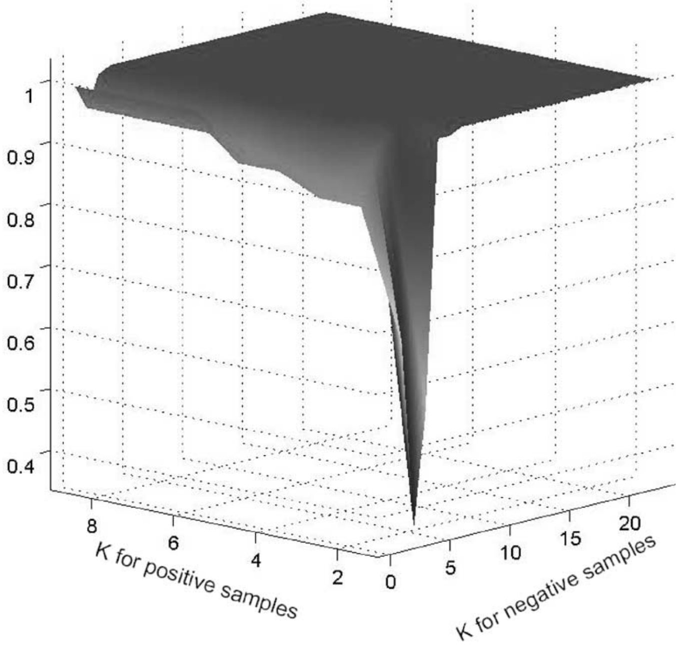


Fig. 21. Surface of classification accuracy for the semantic medical concept "lecture presentation" (with KLT) by using different κ for positive and negative training samples.

Moreover, only the discriminating features are selected for video representation and indexing, and thus the basic time for calculating the feature-based similarity distance is also reduced ($T_c, T_{sc}, T_s, T_o \leq T_m$ because only the discriminating features are used). The total retrieval time for our CBVR system is

$$T_c = N_c \cdot T_c + N_{sc} \cdot T_{sc} + N_s \cdot T_s + N_o \cdot T_o + O(N_o \log N_o) \tag{20}$$

where N_c, N_{sc}, N_s are the numbers of the nodes at the semantic concept cluster and the most relevant subclusters and elementary semantic medical concept levels, N_o is the number of principal video shots that reside in the most relevant elementary semantic medical concept node, T_c, T_{sc}, T_s, T_o are the basic times for calculating the similarity distances in the corresponding feature subspace, and $O(N_o \log N_o)$ is the total time for ranking the relevant principal video shots residing in the corresponding elementary semantic medical concept node. Since $(N_c+N_{sc}+N_s+N_o) \ll N_T, (T_c, T_{sc}, T_s, T_o) \leq T_m$, thus $T_c \ll T_e$. The average performance of our query-evaluation technique is given in Fig. 20.

The limitation of our semantic video-classification technique is that it necessitates a large size of labeled samples to learn accurately because the dimensions of the multimodal perceptual features for video content representation are normally very large, but labeling sufficient video clips that are required for high-dimensional video classification is very expensive and thus infeasible. If only a limited number of labeled samples are available for classifier training, the learned classifier models are incomplete and suffer from the *overfitting* problem, as shown in Fig. 21.

VII. Conclusion

In a certain medical education video domain, we have proposed a novel framework to support more effective semantic video characterization and classification. Our new semantic-sensitive video content characterization framework and adaptive EM algorithm have improved the classification accuracy significantly. The major contributions of this paper include the following.

- A novel semantic-sensitive video content characterization and representation framework via principal video shots. The multimodal perceptual features, that are extracted from the principal video shots, are more effective to discriminate among various semantic medical concepts.
- Semantic medical concept interpretation via flexible mixture model that can be learned from the training samples automatically.
- Adaptive EM algorithm for model selection, parameter estimation and feature subset selection.

The definition of principal video shots is largely domain dependent, but it can be easily extended to other video domains such as news and films by selecting the suitable domain-dependent semantic concepts and defining the relevant concept-driven and domain-dependent multimodal salient objects. After that, our adaptive EM algorithm will also be very attractive to enable semantic video classification for other video domains.

The major limitation of our semantic video classifier is that its performance largely depends on the limited size of the labeled training data set. To address the problem of the limited number of labeled training samples, we are now working on the following.

- Using unlabeled data to obtain more accurate estimation because the limited number of labeled training samples may lead to large generalization error when the data distribution for these limited labeled training samples is different from that of the large-scale unlabeled samples. Our adaptive EM algorithm is very attractive for integrating large-scale unlabeled training samples with the limited number of labeled training samples to obtain a good classifier because the optimal number of mixture Gaussian components is estimated adaptively.
- More extensional studies on performance comparison between our classifier and SVM because SVM was reported to be effective for high-dimensional data classification.

Acknowledgment

The authors would like to thank the reviewers for their useful comments and suggestions. They would also like to thank Dr. J. Kellam for his efforts in providing domain knowledge and evaluating the experimental results.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] J. Fan, W. G. Aref, A. K. Elmagamid, M.-S. Hacid, M. S. Marzouk, and X. Zhu, "MultiView: Multi-level video content representation and retrieval," *J. Electron. Imaging*, vol. 10, no. 4, pp. 895–908, 2001. special issue on multimedia database.

[6] J. D. Courtney, "Automatic video indexing via object motion analysis," *Pattern Recognit.*, vol. 30, pp. 607–625, 1997.

[7] 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.

[8] 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.

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

[10] 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, 2000.

[11] H. J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic parsing of video," in *ACM Multimedia Syst.*, vol. 1, 1993, pp. 10–28.

[12] P. Bouthemy and E. Francois, "Motion segmentation and qualitative dynamic scene analysis from an image sequence," *Int. J. Comput. Vis.*, vol. 10, pp. 157–182, 1993.

[13] B. L. Yeo and B. Liu, "Rapid scene change detection on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, pp. 533–544, Dec. 1995.

[14] T. Meier and K. N. Ngan, "Automatic segmentation of moving objects for video object plane generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 525–538, Sept. 1998.

[15] A. A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tuncel, and T. Sikora, "Image sequence analysis for emerging interactive multimedia services—The European COST 211 framework," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 802–813, Nov. 1998.

[16] B. Gunsel, A. M. Ferman, and A. M. Tekalp, "Temporal video segmentation using unsupervised clustering and semantic object tracking," *J. Electron. Imaging*, vol. 7, pp. 592–604, 1998.

[17] T. N. Tan and K. D. Baker, "Efficient image gradient based vehicle localization," *IEEE Trans. Image Processing*, vol. 9, pp. 1343–1356, Aug. 2000.

[18] J. Meng and S.-F. Chang, "CVEPS—A compressed video editing and parsing system," in *ACM Multimedia Conf.*, Boston, MA, Nov. 1996.

[19] B. Erol and F. Kossentini, "Automatic key video object plane selection using the shape information in the MPEG-4 compressed domain," *IEEE Trans. Multimedia*, vol. 2, pp. 129–138, June 2000.

[20] S.-F. Chang, W. Chen, and H. Sundaram, "Semantic visual templates: Linking visual features to semantics," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998.

[21] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrival," *IEEE Trans. Multimedia*, vol. 3, pp. 141–151, Mar. 2001.

[22] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref, "Image segmentation by integrating color edge detection and seeded region growing," *IEEE Trans. Image Processing*, vol. 10, pp. 1454–1466, Oct. 2001.

[23] J. Fan, X. Zhu, and L. Wu, "An automatic model-based semantic object extraction algorithm," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 1073–1084, Oct. 2001.

[24] Y. Deng, B. S. Manjunath, C. Kenny, M. S. Moore, and H. Shin, "An efficient color representation for image retrieval," *IEEE Trans. Image Processing*, vol. 10, pp. 140–147, 2001.

[25] C. Gu and M. C. Lee, "Semantic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 572–584, Sept. 1998.

[26] J. Guo, J. Kim, and C.-C. J. Kuo, "SIVOG: Smart interactive video object generation system," in *ACM Multimedia Conf.*, Orlando, FL, 1999, pp. 13–16.

[27] D. A. Forsyth and M. Fleck, "Body plan," in *Proc. IEEE Computer Vision and Pattern Recognition*, 1997, pp. 678–683.

[28] M. Kass, A. Wikim, and D. Terzopoulos, "Snakes: Active contour models," in *Proc. 1st Int. Conf. Computer Vision*, June 1987, pp. 259–268.

[29] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis," *IEEE Signal Processing Mag.*, pp. 12–36, Nov. 2000.

[30] C. Snoek and M. Morring, "Multimodal video indexing: A state of the art review," *Multimedia Tools Applic.*, vol. 18, pp. 231–256, 2003.

[31] A. G. Hauptmann and M. A. Smith, "Text, speech, and vision for video segmentation: The informedia project," in *AAAI Fall Symp. Computational Models for Language and Vision*, Orlando, FL, 1995, pp. 123–132.

[32] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," in *EURASIP JASP*, vol. 2, 2003, pp. 170–185.

[33] Z. Liu, J. Huang, and Y. Wang, "Classification of TV programs based on audio information using hidden Markov model," in *IEEE Workshop on Multimedia Signal Processing*, 1998, pp. 27–32.

[34] T. Liu and J. R. Kender, "A hidden Markove model approach to the structure of documents," in *Proc. CAIVD*, vol. 18, 2000, pp. 112–132.

[35] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K. Wong, "Integration of multimodal features for video classification based on HMM," in *IEEE Workshop on Multimedia Signal Processing*, vol. 18, 1999, pp. 132–140.

[36] G. Sudhir, J. Lee, and A. K. Jain, "Automatic classification for tennis video for high-level content-based retrieval," in *Proc. CAIVD*, 1998.

[37] S. Fisher, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *ACM Multimedia*, vol. 18, 1995, pp. 145–154.

[38] H. Sundaram and S. F. Chang, "Computable scenes and structures in films," *IEEE Trans. Multimedia*, vol. 4, pp. 482–491, 2002.

[39] B. Adames, C. Dorai, and S. Venkatesh, "Toward automatic extraction of expressive elements of motion pictures: Tempo," *IEEE Trans. Multimedia*, 2002.

[40] W. Zhou, A. Vellaikal, and C. Kuo, "Rule-based video classification system for basketball video indexing," in *ACM Multimedia*, vol. 18, 2000, pp. 128–132.

[41] A. Alatan, A. Akasu, and W. Wolf, "Multimodal dialog scene detection using hidden markov models for content-based multimedia indexing," *Multimedia Tools Applic.*, vol. 14, pp. 137–151, 2001.

[42] Y. Liu, F. Dellaert, and W. E. Rothfus, "Classification Driven Semantic Based Medical Image Indexing and Retrieval,", CMU-RI-TR-98-25, 1998.

[43] J.-H. Lim, "Learnable visual keywords for image classification," in *ACM Conf. Digital Library*, Berkeley, CA, 1999.

[44] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantic-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Machine Intell.*, 2001.

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

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

[47] A. Vailaya, M. Figueiredo, A. K. Jain, and H. J. Zhang, "A Bayesian framework for semantic classification of outdoor vacation images," in *Proc. SPIE*, vol. 3656, 1998, pp. 231–242.

[48] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[49] W. Liu and A. Hauptmann, "News video classification using SVM-based multimodal classifiers and combination strategies," in *ACM Multimedia*, vol. 18, 2002, pp. 148–152.

[50] N. Vasconcelos and A. Lippman, "A Bayesian framework for semantic content characterization," in *Proc. CVPR*, vol. 18, 1998, pp. 154–162.

[51] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, 2002.

[52] M. Weber, M. Welling, and P. Perona, "Toward automatic discovery of object categories," in *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 18, 2000, pp. 128–136.

[53] P. Lipson, E. Grimson, and P. Sinha, "Configuration based scene and image indexing," in *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 18, 1997, pp. 121–129.

[54] D.-R. Liu, C.-H. Lin, and J.-J. Hwang, "Classifying video documents by hierarchical structure of video contents," *Comput. J.*, vol. 43, no. 5, pp. 396–410, 2000.

[55] W.-S. Li, K. S. Candan, K. Hirata, and Y. Hara, "Hierarchical image modeling for object-based media retrieval," *Data Knowl. Eng.*, vol. 27, pp. 139–176, 1998.

[56] A. Baraani-Dastjerdi, J. Pieprzyk, and R. Safavi-Naini, "A multi-level view model for secure object-oriented databases," *Data Knowl. Eng.*, vol. 23, pp. 97–117, 1997.

[57] A. Benitez, S.-F. Chang, and J. R. Smith, "IMKA: A multimedia organization system combining perceptual and semantic knowledge," in *ACM Multimedia*, vol. 18, 2001, pp. 121–129.

[58] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation," in *Proc. SPIE*, vol. 4210, 2000, pp. 129–140.

[59] C. Meilhac and C. Nastar, "Relevance feedback and category search in image databases," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, Italy, 1999.

[60] T. P. Minka and R. W. Picard, "Interactive learning with a society of models," *Pattern Recognit.*, vol. 30, no. 4, pp. 565–581, 1997.

[61] 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, Sept. 1998.

[62] Y. Ishikawa, R. Subramanya, and C. Faloutsos, "Mindreader: Querying databases through multiple examples," in *Proc. VLDB*, vol. 18, 1998, pp. 210–220.

[63] I. J. Cox, M. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The bayesian image retrieval system, PicHunter: Theory, implementation and psychophysical experiments," *IEEE Trans. Image Processing*, vol. 9, pp. 20–37, Jan. 2000.

[64] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *ACM Multimedia Conf.*, 2001, pp. 107–118.

[65] P. Wu and B. S. Manjunath, "Adaptive nearest neighbor search for relevance feedback in large image database," in *ACM Multimedia Conf.*, 2001, pp. 89–97.

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

[67] D. B. Lomet and B. Salzberg, "The hB-tree: A multiattribute indexing method with good guaranteed performance," *ACM Trans. Database Syst.*, vol. 15, no. 4, pp. 625–658, 1990.

[68] K. Lin, H. V. Jagadish, and C. Faloutsos, "The TV-tree: An index structure for high dimensional data," *VLDB J.*, vol. 18, pp. 120–130, 1994.

[69] N. Katayama and S. Satoh, "The SR-tree: An index structure for high dimensional nearest neighbor queries," in *ACM SIGMOD*, vol. 18, 1997, pp. 125–134.

[70] S. Berchtold, D. A. Keim, and H. P. Kriegel, "The X-tree: An index structure for high-dimensional data," in *Proc. Int. Conf. Very Large Databases*, vol. 18, 1996, pp. 134–145.

[71] C. Li, E. Chang, H. Garcia-Molina, J. Z. Wang, and G. Wiederhold, "Clindex: Clustering for similarity queries in high-dimensional spaces," Dept. Comput. Sci., Stanford Univ., Stanford, CA, Tech. Rep., 2000.

[72] M. A. Carreira-Perpinan, "A review of dimension reduction techniques," University of Sheffield, Sheffield, U.K., Tech. Rep. CS-96-09, 1997.

[73] A. Thomasian, V. Castelli, and C.-S. Li, "Clustering and singular value decomposition for approximate indexing in high dimensional space," in *Proc. Int. Conf. Information and Knowledge Management (CIKM)*, Bethesda, MD, 1998, pp. 201–207.

[74] K.V.R. Kanth, D. Agrawal, and A. Singh, "Dimensionality reduction for similarity searching in dynamic databases," in *ACM SIGMOD*, 1998, pp. 166–176.

[75] Z. Su, S. Li, and H. Zhang, "Extraction of feature subspace for content-based retrieval using relevance feedback," in *ACM Multimedia Conf.*, Ottawa, ON, Canada, 2001, pp. 98–106.

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

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

[78] 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.

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

[80] 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*, vol. 3846, Boston, MA, Sept. 1999, pp. 148–164.

[81] J. R. Smith, "VideoZoom spatial-temporal video browsing," *IEEE Trans. Multimedia*, vol. 1, pp. 151–171, June 1999.

[82] X. Zhu, J. Fan, A. K. Elmagarmid, and W. G. Aref, "Hierarchical video summarization for medical data," in *Proc. SPIE: Storage and Retrieval for Media Databases*, San Jose, CA, Jan. 23–26, 2002.

[83] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Mag.*, vol. 11, pp. 47–60, 1996.

[84] L. Xu and M. I. Jordan, "On the convergence properties of the EM algorithm for Gaussian mixtures," *Neural Computat.*, vol. 8, pp. 129–136, 1996.

[85] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, pp. 2664–2677, 1994.

[86] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 2000.

[87] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant-EM algorithm with application to image retrieval," in *Proc. CVPR*, 2000, pp. 222–227.

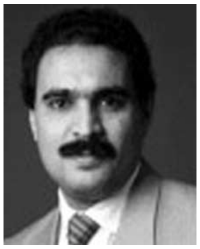
[88] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Mathemat. Stat.*, vol. 22, pp. 76–86, 1951.

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, China, in 1997.

He was a Researcher at Fudan University, Shanghai, China, 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 a Researcher in the 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, Charlotte. His research interests include nonlinear systems, error correction codes, image processing, video coding, semantic video computing, and content-based video indexing and retrieval.

Hangzai Luo received the B.S. degree in computer science from Fudan University, Shanghai, China, in 1998. He is currently pursuing the Ph.D. degree in information technology at University of North Carolina, Charlotte, NC.

From 1998 to 2002, he was a Lecturer in Department of Computer Science, Fudan University. His research interests includes video analysis and content-based video retrieval.



Ahmed K. Elmagarmid (M'88–SM'93) received the M.S. and Ph.D. degrees in computer and information sciences from Ohio State University, 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.

Dr. Elmagarmid has served on the Editorial Board of IEEE TRANSACTIONS ON Computers and is now the Associate Editor for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. He is the Founding Editor-in-Chief of *International Journal on Distributed and Parallel Databases*. He serves as an Editor for *Information Science Journal, International Journal of Communication Systems,* and the book series *Advanced Database Systems* Kluwer). 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 as Program Chair and General Chair. He received a National Science Foundational PYI Award in 1988 and was named a "Distinguished Alumnus" of the Ohio State University in 1993 and the University of Dayton in 1995. He is a Member of the ACM.