Incorporating Feature Hierarchy and Boosting to Achieve More Effective Classifier Training and Concept-Oriented Video Summarization and Skimming

HANGZAI LUO
Software Engineering Institute, East China Normal University, China
YULI GAO
University of North Carolina, Charlotte
XIANGYANG XUE
Fudan University, China
JINYE PENG
Northwestern Polytechnical University, China
and
JIANPING FAN
University of North Carolina, Charlotte

For online medical education purposes, we have developed a novel scheme to incorporate the results of semantic video classification to select the most representative video shots for generating concept-oriented summarization and skimming of *surgery education videos*. First, salient objects are used as the video patterns for feature extraction to achieve a good representation of the intermediate video semantics. The salient objects are defined as the salient video compounds that can be used to characterize the most significant perceptual properties of the corresponding real world physical objects in a video, and thus the appearances of such salient objects can be used to predict the appearances of the relevant semantic video concepts in a specific video domain. Second, a novel *multi-modal boosting* algorithm is developed to achieve more reliable video classifier training by incorporating feature hierarchy and boosting to dramatically reduce both the training cost and the size of training samples, thus it can significantly speed up SVM (support vector machine) classifier training. In addition, the unlabeled samples are integrated to reduce the human efforts on labeling large amount of training samples. Finally, the results of semantic video classification are incorporated to enable concept-oriented video summarization and skimming. Experimental results in a specific domain of *surgery education videos* are provided.

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning—*Concept learning*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms: Algorithms, Experimentation

This work was supported by National Science Foundation (NSF) under 0601542-IIS, 0208539-IIS, and by a grant from the AO Foundation.

Authors' addresses: H. Luo, Software Engineering Institute, East China Normal University, China; Y. Gao and J. Fan, Department of Computer Science, University of North Carolina, Charlotte, NC; X. Xue, Department of Computer Science, Fudan University, China; J. Peng, School of Electronics and Information, Northwestern Polytechnical University, China; Corresponding author's e-mail: jfan@uncc.edu.

Additional Key Words and Phrases: Semantic video classification, concept-oriented video skimming, salient objects, unlabeled samples, feature hierarchy, multi-modal boosting

ACM Reference Format:

Luo, H., Gao, Y., Xue, X., Peng, J., and Fan, J. 2008. Incorporating feature hierarchy and boosting to achieve more effective classifier training and concept-oriented video summarization and skimming. ACM Trans. Multimedia Comput. Commun. Appl. 4, 1, Article 1 (January 2008), 25 pages. DOI = 10.1145/1324287.1324288 http://doi.acm.org/10.1145.1324287.1324288

1. INTRODUCTION

Content-based video retrieval (CBVR) techniques are very attractive for supporting more cost-efficient online medical education by searching and illustrating the most relevant clinical examples in a video to medical students over Internet [Ebadollahi et al. 2002]. To incorporate CBVR techniques for online medical education purposes, the following inter-related issues should be addressed simultaneously: (a) Developing more effective techniques for semantic video classification, so that the medical students can search large-scale archives of medical video clips at the semantic level by using keywords. (b) Enabling concept-oriented video summarization and skimming, so that the medical students can quickly browse the query results (i.e., the medical video clips returned by the same keyword-based query) and judge the relevance between the query results and their real needs. (c) Considering the influence of users' preferences, so that *subjective video summarization and skimming* can be achieved for online medical education purposes. For example, when a medical professor is giving a lecture presentation for one particular medical concept such as "trauma surgery", she/he may want to summarize the query results subjectively according to the particular medical concept "trauma surgery", so that she/he can illustrate the query results to the students more objectively and precisely.

To achieve concept-oriented video summarization and skimming, there is an urgent need to develop more effective schemes for classifying the medical video clips into the most relevant semantic video concepts. However, the CBVR community has long struggled to *bridge the semantic gap* from successful low-level feature extraction to high-level human interpretation of video semantics, thus bridging the semantic gap is of crucial importance for achieving more effective video classification. In order to bridge the semantic gap more effectively, the following inter-related issues should be addressed jointly: (1) What are the suitable video patterns for feature extraction? The basic requirement for such video patterns is that they should be able to achieve a good representation of the intermediate video semantics effectively and efficiently. (2) What is the basic vocabulary of semantic video concepts of particular interest in a specific medical video domain? (3) Because the medical experts are highly paid, it is very expensive to obtain large amount of labeled training samples for reliable video classifier training. How can we learn the concept models (i.e., video classifiers) accurately when only a small number of labeled samples are available? (4) How can we automatically incorporate the results of semantic video classification to select the most representative video shots for generating the concept-oriented video summarization and skimming? (5) How can we consider the influence of the users' preferences to achieve *subjective video summarization and skimming*?

Bridging the semantic gap in general is very hard if not impossible, thus narrowing the video domain plays an important role on achieving more effective video classification [Pfeiffer et al. 1999; Kender and Yeo 1998; Sundaram and Chang 2002a; Adames et al. 2002; Haering et al. 2000; Zhou et al. 2000; Fischer et al. 1995; Hanjalic et al. 1999; Liu et al. 1998; Ekin et al. 2003; Xie et al. 2003; Dimitrova et al. 2000; Adams et al. 2003; Naphade and Huang 2001; Jaimes and Chang 2001; Greenspan et al. 2004; Qi et al. 2003; Gatica-Perez et al. 2003]. Based on this observation, our research focuses on one specific domain of *surgery education videos*, where the contextual relationships between the semantic video concepts and the relevant video patterns and their low-level perceptual features are better defined.

1.1 Related Works and Our Contributions

Many techniques have been developed to achieve low-level video content analysis automatically [Sebe et al. 2003; Djeraba 2002; Correia and Pereira 2004; Freund and Schapire 1996; Tieu and Viola 2000; Zhang et al. 1993; Chang et al. 1998; Alatan et al. 1998; Lew 2001; Snoek and Morring 2003]. Without understanding the video semantics, some computable perceptual features are used to characterize the principal properties of video contents. The advantage of such computational approach is that the perceptual features are easy to compute, but the shortcoming is that these computable perceptual features may not have direct correspondence with the underlying video semantics (i.e., semantic gap between the low-level perceptual features and the high-level human interpretation of video semantics). Thus, using such computable perceptual features to generate video summarization and skimming may not provide acceptable results. In addition, the quality of such computable perceptual features (i.e., the effectiveness of the perceptual features for characterizing the underlying video semantics) largely depends on the video patterns that are used for feature extraction.

Thus high-level video semantics such as semantic video events are needed to be detected for achieving more accurate video summarization and skimming at the concept level, and the rule-based approach is widely used for semantic video event detection by using both the structure elements of video contents and the underlying video making rules [Pfeiffer et al. 1999; Kender and Yeo 1998]. One advantage of such rule-based approach is the ease to insert, delete, and modify the existing rules when the nature of video classes changes. Since the underlying video making rules are used for detecting the semantic video events, the rule-based approach is only attractive for some specific domains such as news video and film which have well-defined story structure for the semantic video events [Adams et al. 2003; Naphade and Huang 2001; Jaimes and Chang 2001; Greenspan et al. 2004; Qi et al. 2003; Gatica-Perez et al. 2003; Ma et al. 2002; Smith and Kanade 1995; Arman et al. 1994; He et al. 1999; Sundaram et al. 2002; Chang 2002; Sundaram and Chang 2002b]. On the other hand, surgery education videos do not have well-defined video making rules that can be used to generate the semantic video events, thus it is very important to develop new techniques that are able to achieve more effective video classification and concept-oriented video summarization and skimming. For online medical education purposes, it is also very important to achieve *subjective video summarization and skimming* by considering the influence of the users' preferences.

By considering both the users' information needs and the video structure elements, Sundaram et al. and others have proposed a novel algorithm to achieve more effective video summarization and skimming [Sundaram et al. 2002; Chang 2002; Djeraba 2000; Fan et al. 2001; Cristianini and Shawe-Taylor 2000; Sundaram and Chang 2002b]. He et al. [1999] have also incorporated audio-visual information to summarize the audio-video presentations. To integrate the high-level video semantics for video summarization and skimming, hidden Markov models and Gaussian mixture models have been investigated widely for detecting semantic video scenes [Liu et al. 1998; Ekin et al. 2003; Xie et al. 2003; Dimitrova et al. 2000; Adams et al. 2003; Naphade and Huang 2001; Jaimes and Chang 2001; Greenspan et al. 2004; Qi et al. 2003; Gatica-Perez et al. 2003]. Naphade and Huang [2001] and Adams et al. [2003] have also investigated the techniques for hierarchical video classification. For distance education purposes, some researchers have also developed multiple techniques to achieve task-driven lecture video analysis, summarization and retrieval [Zhang and Nunamaker 2004; Liu and Kender 2004; Deshpande and Hwang 2001; Li et al. 2006].

Our proposed scheme significantly differs from all these earlier works in multiple respects: (a) We focus on one specific domain of *surgery education videos* with less editing structures and production

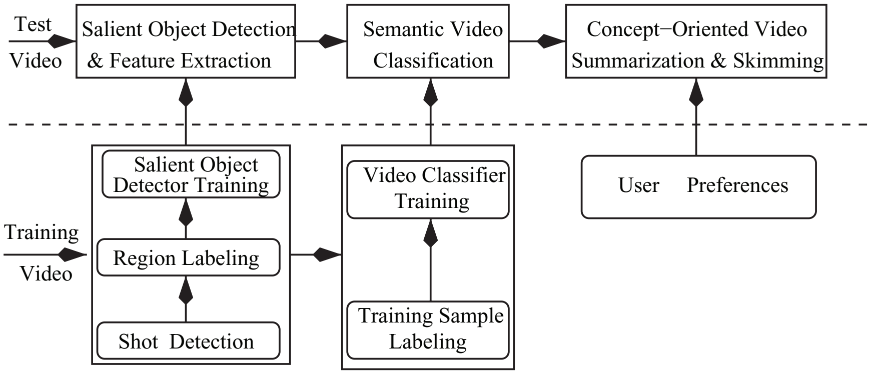


Fig. 1. The flowchart for our system.

metadata. Because large amount of real clinical examples in a video are illustrated for student training, surgery education videos are significantly different from traditional lecture videos [Zhang and Nunamaker 2004; Liu and Kender 2004; Deshpande and Hwang 2001; Li et al. 2006]. (b) A new scheme for video content representation is proposed by using the salient objects to achieve a good interpretation of the intermediate video semantics. The salient objects are defined as the salient video compounds that can be used to effectively characterize the most significant perceptual properties of the corresponding real world physical objects in a video. Thus, the appearances of the salient objects can be used to effectively predict the appearances of the relevant semantic video concepts in a specific video domain. (c) A novel multi-modal boosting algorithm is proposed to significantly speed up SVM video classifier training and generalize the video classifiers from fewer training samples by incorporating feature hierarchy and boosting for ensemble classifier training and feature subset selection. (d) A new scheme is developed to incorporate unlabeled samples for classifier training and reduce the human efforts on labeling large amounts of training samples, where the outlying unlabeled samples are distinguished from the informative unlabeled samples and the certain unlabeled samples accurately. (e) Finally, the results of semantic video classification are seamlessly incorporated to enable concept-oriented video summarization and skimming.

This article is organized as follows: Section 2 gives a brief description of our system flowchart; Section 3 presents a new scheme for video content representation and presents a novel scheme for automatic salient object detection; Section 4 introduces a novel algorithm to speed up SVM video classifier training; Section 5 proposes a new technique to enable concept-oriented video summarization and skimming; Section 6 shows the benchmark environment for evaluating our techniques for semantic video classification and concept-oriented video summarization and skimming; We conclude in Section 7.

2. SYSTEM OVERVIEW

In this article, we focus on supporting concept-oriented video summarization and skimming in one specific domain of *surgery education videos*. Our system consists of the following major components as shown in Figure 1:

(a) *Automatic Salient Object Detectors*. To achieve a good representation of the intermediate video semantics, salient objects are extracted automatically and treated as the underlying video patterns for feature extraction and video content representation. Detecting the salient objects are able to effectively predict the appearances of the relevant semantic video concepts in a specific domain of *surgery education videos*. For examples, human faces, human voices, lecture slides are the salient objects that can be used to interpret the appearance of the semantic video concept “lecture presentation”; The appearance of the semantic video concept “colon surgery” is highly related to the appearances of the salient objects, “blue cloth,” “doctor gloves,” and “colon regions.” Thus using the salient objects for video content representation has at least three significant benefits: (a) Comparison with the video shots, the salient objects can characterize the most significant perceptual properties of the corresponding real world physical objects in a video effectively. Thus, using the salient objects for feature extraction and video content representation will enhance the quality of features and allow more accurate video classification, indexing and retrieval. (b) The salient objects are not necessarily the accurate segmentation of the real world physical objects in a video, thus both the computational cost and the detection error rate can be reduced significantly. (c) It is able to achieve a good balance between the computational complexity, the detection accuracy, and the effectiveness for video semantics interpretation.

(b) *SVM Video Classifiers*. The SVM classifiers have better generalization ability in high-dimensional feature space, thus we focus on training the SVM video classifiers for detecting large amount of semantic video concepts in a specific domain of surgery education videos. However, the computational complexity for SVM video classifier training in high-dimensional feature space is very expensive and large amount of training samples are needed to achieve reliable classifier training. To reduce the computational complexity for SVM video classifier training, a novel *multi-modal boosting* algorithm is proposed to speed up SVM classifier training and enable multi-modal feature subset selection by incorporating feature hierarchy and boosting for ensemble classifier training. In addition, a novel algorithm is proposed to integrate unlabeled samples for SVM video classifier training for reducing the human efforts on labeling large amount of training samples. After the SVM video classifiers are available, the medical video clips in database are classified into the most relevant semantic video concepts automatically.

(c) *Video Summarization and Skimming Generators*. The results of semantic video classification are further incorporated to select the most representative video shots to generate the concept-oriented video summarization and skimming for each video clip in database. The principal video shots for each medical video clip are sorted according to their importance scores, and the most representative video shots with higher importance scores are selected automatically for generating the concept-oriented video summarization and skimming. In addition, we have also designed an interactive visualization interface for users to select the most relevant semantic video concepts for video summarization generation according to their personal preferences.

3. AUTOMATIC SALIENT OBJECT DETECTION FOR VIDEO CONTENT REPRESENTATION

By using the salient objects for feature extraction and video content representation [Luo et al. 2004; Fan et al. 2004], the semantic gap, between the low-level multi-modal video signals (i.e., perceptual features for video content representation) and the semantic video concepts (i.e., high-level human interpretation of video semantics), is partitioned into two “bridgable” gaps: (a) *Gap 1*: The bridgable gap between the low-level multi-modal perceptual features and the salient objects (i.e., intermediate video semantics); (b) *Gap 2*: The bridgable gap between the salient objects and the relevant semantic video concepts. The Gap 1, between the salient video compounds (i.e., real world physical objects in a video) and the low-level video signals, is bridged by using the salient objects and their multi-modal perceptual features to achieve a good representation of the intermediate video semantics. The Gap 2, between the semantic video concepts and the salient objects, is bridged by using *multi-modal boosting* to exploit the strong correlations (i.e., contextual relationships) between the appearances of the semantic video concepts and the appearances of the relevant salient objects (see Section 4). For a specific domain of surgery education videos, our assumption is that the appearance of the semantic video concepts may depend on the co-appearances of the relevant salient objects.

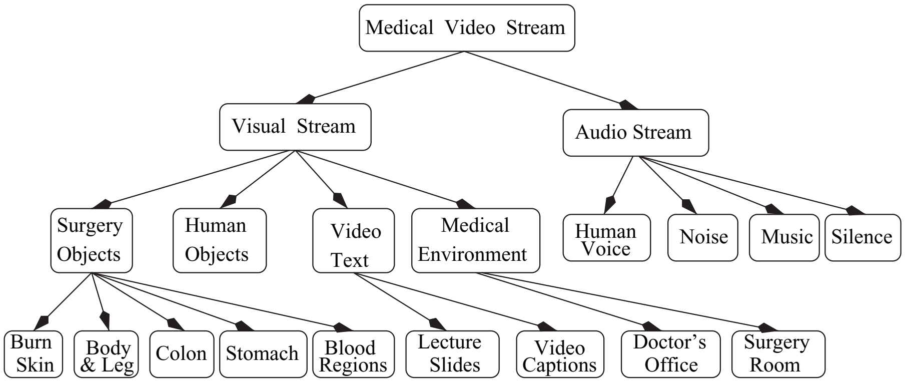


Fig. 2. Taxonomy of salient video compounds in a specific domain of surgery education videos.

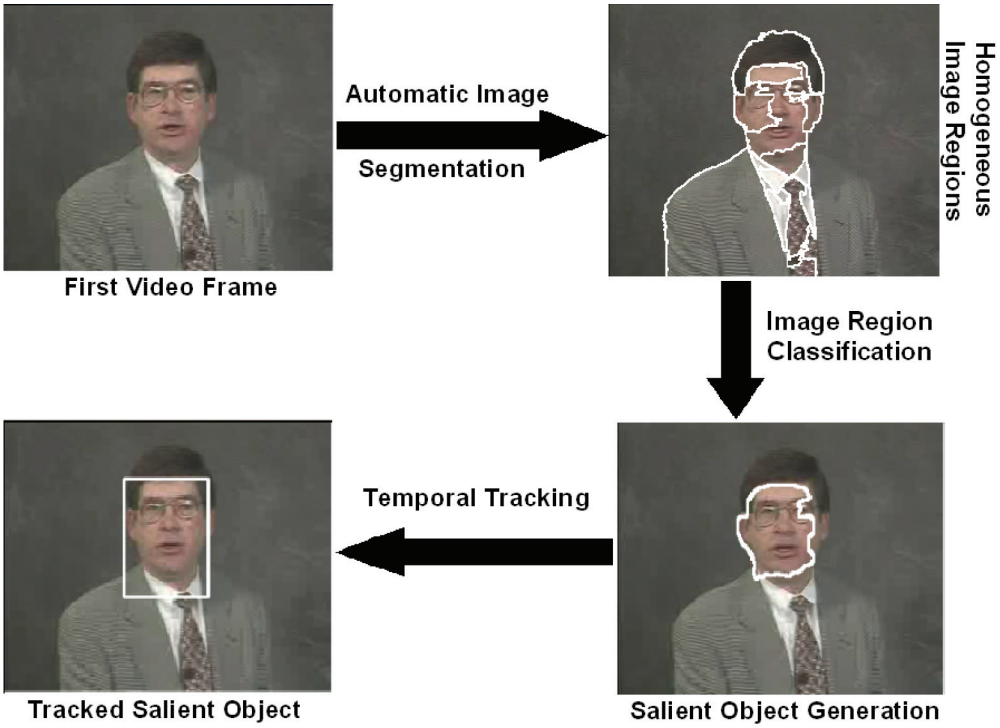

First Video Frame
Automatic Image Segmentation
Homogeneous Image Regions
Image Region Classification
Salient Object Generation
Temporal Tracking
Tracked Salient Object

Fig. 3. The major steps for salient object detection.

The basic vocabulary of the salient objects in a specific domain of surgery education videos can be obtained by using the taxonomy of the salient video compounds as shown in Figure 2. We have designed a set of detection functions to detect these salient objects, and each function is designed for detecting one particular type of the salient objects in the basic vocabulary. We use our detection function of the salient object "human face" as an example to illustrate how we can design our salient object detection functions. As shown in Figure 3, image regions with homogeneous color or texture are first obtained by using our automatic image segmentation technique [Fan et al. 2001]. Since the visual properties for one

Table I. The Average Performances (i.e., Precision ρ versus Recall ϱ) of Our Detection Functions

Salient Objects	Face	Slide	Blood	Music	Speech
ρ	90.3%	96.4%	92.2%	94.6%	89.8%
ϱ	87.5%	95.6%	96.7%	81.2%	82.6%
Salient Objects	Skin	Sketch	Dialog	Noise	Colon
ρ	96.7%	93.3%	89.7%	86.9%	94.3%
ϱ	95.4%	88.5%	83.2%	84.7%	87.5%
Salient Objects	Silence	Legs	Stomach	Captions	Blue Cloth
ρ	96.3%	92.4%	89.7%	91.8%	96.7%
ϱ	94.5%	89.8%	91.2%	93.2%	95.8%

specific salient object may look different at various lighting and capturing conditions, using only one video frame is insufficient to represent its visual properties. Thus, this automatic image segmentation procedure is performed on a set of training video frames which consist of the specific salient object “human face”.

The homogeneous image regions in the training video frames are labeled as the positive samples and the negative samples for the given salient object “human face”. For each labeled image region (i.e., labeled training samples), the following 23-dimensional perceptual features are extracted: 3-dimensional LUV dominant colors, 7-dimensional Tamura texture features, 12-bins color histogram, 1-dimensional coverage ratio. A binary SVM image-region classifier is automatically learned from these labeled image regions [Cristianini and Shawe-Taylor 2000]. After the binary SVM image-region classifier is obtained, it is used to classify the homogeneous image regions into two classes: face regions *versus* non-face regions. The connected homogeneous image regions with the same labels are further merged and aggregated as the corresponding salient object “human face”. To track the salient object “human face” among the video frames, a region-based motion estimation is performed to determine the temporal relationships of the object regions among the video frames [Luo et al. 2004].

After the automatic detection functions are learned, they are used to detect the salient objects from the test medical video clips. Our detection functions take the following major steps for automatic salient object detection: (a) The video shots are first detected by using the scene cut detection technique developed in Fan et al. [2004]; (b) For each video shot, all these salient object detection functions are then performed to determine the underlying salient objects; (c) Because one video shot may contain multiple salient objects and the appearances of salient objects may be uncertain cross the video frames, *confidence map* is calculated to measure the posterior probability for each video region to be classified into the most relevant salient object. The significance of our new video content representation scheme is that the confidence maps are used to tackle the uncertainty and the dynamics of the appearances of the salient objects along the time, and the changes of the confidence maps can also indicate their motion properties effectively.

Precision ρ and recall ϱ are further used to measure the average performance of our detection functions:

$$\rho = \frac{\eta}{\eta + \varepsilon}, \qquad \varrho = \frac{\eta}{\eta + \upsilon}, \tag{1}$$

where η is the set of true positive samples that are related to the given type of salient object and are detected correctly, ε is the set of true negative samples that are irrelevant to the given type of salient object and are detected incorrectly, and υ is the set of false positive samples that are related to the given type of salient object but are mis-detected. The average performance for some detection functions are given in Table I.

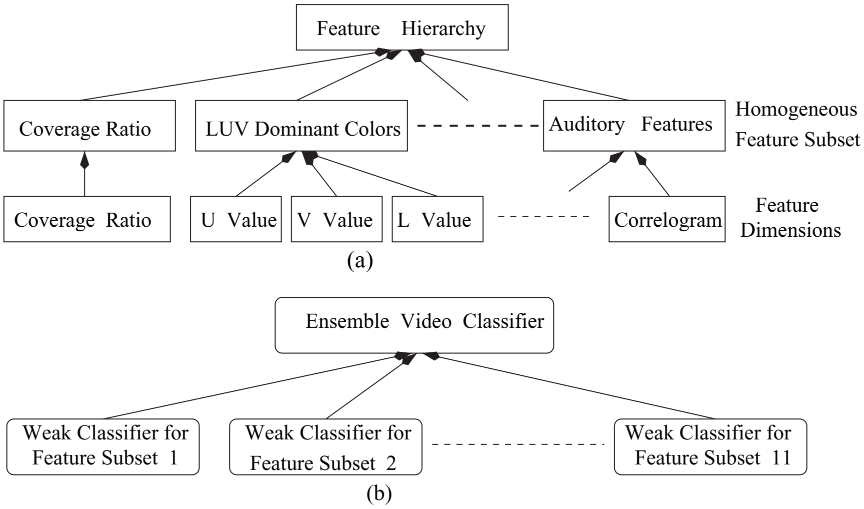

Feature Hierarchy
Coverage Ratio
LUV Dominant Colors
Auditory Features
Homogeneous Feature Subset
Coverage Ratio
U Value
V Value
L Value
Correlogram
Feature Dimensions
(a)
Ensemble Video Classifier
Weak Classifier for Feature Subset 1
Weak Classifier for Feature Subset 2
Weak Classifier for Feature Subset 11
(b)

Fig. 4. (a) Feature hierarchy; (b) Multi-modal boosting for classifier combination.

After the salient objects are detected, a set of multi-modal perceptual features are then extracted to characterize their perceptual properties. The visual features include 1-dimensional coverage ratio (i.e., density ratio) for object shape representation, 6-dimensional object locations (i.e., 2-dimensions for object center and 4-dimensions to indicate the rectangular box for coarse shape representation of salient object), 7-dimensional LUV dominant colors and color variances, 14-dimensional Tamura texture, 28-dimensional wavelet texture features, 14-dimensional feature set for the sampling locations of the object trajectory. The 14-dimensional auditory features include loudness, frequencies, pitch, fundamental frequency, and frequency transition ratio [Liu et al. 1998; Luo et al. 2004].

Using high-dimensional multi-modal perceptual features for salient object representation is able to characterize the principal properties of the relevant real world physical objects in a video more effectively. However, learning the video classifiers on such high-dimensional feature space requires large amount of training samples that increase exponentially with the feature dimensions. In addition, these multi-modal perceptual features are heterogeneous on their statistical and geometrical properties. To address this problem, we incorporate *feature hierarchy* to organize the high-dimensional multi-modal perceptual features more effectively, that is, homogeneous feature subsets at the first level and feature dimensions for each homogeneous feature subset at the second level. In our current implementation, these 76-dimensional multi-modal perceptual features are partitioned into 11 homogeneous feature subsets as shown in Figure 4(a): coverage ratio, object locations, LUV dominant colors, LUV color variances, Tamura texture 1, Tamura textue 2, average energy wavelet features, variance wavelet features, object trajectory, volume-based auditory features, ratio-based auditory features. Each homogeneous feature subset has different feature dimensions, for examples, the homogeneous feature subset "coverage ratio" has one feature dimension: coevrage ratio; and the homogeneous feature subset "LUV dominant colors" has three feature dimensions: L value, U value, and V value. Each feature subset is used to characterize one particular type of perceptual properties of the salient objects, and thus the underlying statistical and geometrical properties are homogeneous.

4. SVM CLASSIFIER TRAINING FOR SEMANTIC VIDEO CLASSIFICATION

It is worth noting that the video classifiers at the concept level are different from the image-region classifiers which are used for salient object detection. The video classifiers are used to classify the video shots into the most relevant semantic video concepts, and the image-region classifiers described in Section 3 are used to classified the image regions into the most relevant salient objects. They are trained by using different training samples with different features, that is, the training samples for the video classifiers are the labeled video shots and the training samples for the image-region classifiers are the labeled image regions.

To learn the SVM video classifiers, we use *one-against-all* rule to organize the labeled samples (i.e., video shots) $\Omega_{c_j} = \{X_l, C_j(S_l)|l = 1, \ldots, N_L\}$ into: *positive samples* for one given semantic video concept C_j and *negative samples*. Each labeled sample is a pair $(X_l, C_j(S_l))$ that consists of a set of 76-dimensional features X_l and the semantic label $C_j(S_l)$ for the corresponding sample S_l. The unlabeled samples $\overline{\Omega_{c_j}} = \{X_k, S_k|k = 1, \ldots, N_u\}$ can also be used to improve the classification accuracy. For the given semantic video concept C_j, we then define the mixture training sample set as $\Omega = \Omega_{c_j} \bigcup \overline{\Omega_{c_j}}$.

To reduce the cost for SVM video classifier training, the high-dimensional multi-modal perceptual features are automatically partitioned into multiple low-dimensional homogeneous feature subsets according to the underlying perceptual properties to be characterized, where the strongly correlated perceptual features of the same modality are automatically partitioned into the same homogeneous feature subset. Each homogeneous feature subset is used to characterize one particular perceptual property of the video data, thus the underlying statistical and geometric property of the video data is uniform and can accurately be approximated by using one particular type of kernel functions. In addition, different types of kernel functions can be used for different homogeneous feature subsets to approximate the diverse statistical and geometric properties of the video data more accurately.

Because the feature dimensions for each homogenous feature subset are relatively low, a weak SVM classifier is trained for each homogeneous feature subset to speed up SVM classifier training and generalize the classifier by using fewer training samples. To exploit the *intra-set* and *intra-modality feature correlations* among different feature dimensions of the same modality, principal component analysis (PCA) is performed on each homogeneous feature subset. The *inter-set* and *inter-modality feature correlations* among different homogeneous feature subsets and the *output correlations* between the weak classifiers are effectively exploited in the following classifier combination (decision fusion) procedure. Exploiting both the feature correlations and the output correlations between the weak classifiers can achieve more reliable training of ensemble classifier and result in higher classification accuracy.

The homogeneous feature subsets of different modalities are used to characterize different perceptual properties of video data, thus the outputs of the corresponding weak SVM classifiers are diverse and compensable. For one specific semantic video concept, a novel *multi-modal boosting* algorithm is developed to generate an ensemble classifier by combining the weak SVM classifiers for all these homogeneous feature subsets of different modalities as shown in Figure 4(b).

4.1 Weak SVM Video Classifier Training

For each homogeneous feature subset, a weak classifier is trained for the given semantic video concept. For the positive samples X_l with $Y_l = +1$, there exists the transformation parameters W and b such that $f(X_l) = W \cdot \Phi(X_l) + b \geq +1$. Similarly, for the negative samples X_l with $Y_l = -1$, we have $f(X_l) = W \cdot \Phi(X_l) + b \leq -1$. $\Phi(X_l)$ is the function that maps X_l into higher-dimensional space and the kernel function is defined as $\kappa(X_i, X_j) = \Phi(X_i)^T \Phi(X_j)$. In our current implementation, the radial basis function (RBF) is selected, $\kappa(X_i, X_j) = \exp(-\gamma||X_i - X_j||^2)$, $\gamma > 0$. The margin between these two supporting planes will be $2/||W||^2$. The weak SVM classifier is then designed for maximizing the margin with the constraints $f(X_l) = W \cdot \Phi(X_l) + b \geq +1$ for the positive samples and $f(X_l) = W \cdot \Phi(X_l) + b \leq -1$ for the negative samples.

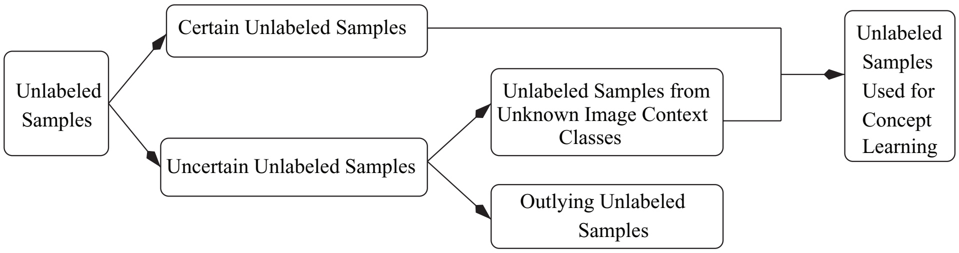

Unlabeled Samples
Certain Unlabeled Samples
Uncertain Unlabeled Samples
Unlabeled Samples from Unknown Image Context Classes
Outlying Unlabeled Samples
Unlabeled Samples Used for Concept Learning

Fig. 5. The flowchart for our algorithm to eliminate the misleading effects of the outlying unlabeled samples.

Given the set of the labeled training samples $\Omega_{c_j} = \{X_l, Y_l | l = 1, \ldots, N_L\}$, the margin maximization procedure is then transformed into the following optimization problem:

$$min \left\{ \frac{1}{2} \|W\|^2 + C \sum_{l=1}^{N_L} \xi_l \right\} \tag{2}$$

subject to:

$$\forall_{l=1}^{N_L} : Y_l(W \cdot \Phi(X_l) + b) \geq 1 - \xi_l$$

where $\xi_l \geq 0$ represents the training error rate, $C > 0$ is the penalty parameter to adjust the training error rate and the regularization term $\frac{1}{2}\|W\|^2$.

We have also incorporated a *hierarchical search* algorithm to determine the optimal model parameters $(\bar{C}, \bar{\gamma})$ for the weak SVM classifier by using moving search window: (a) Instead of using a fixed-size partition of parameter space [Fan et al. 2005], the parameter search procedure is first performed on the parameter space for C and γ by using a large-size moving search window. (b) After the parameter pair (C, γ) at the coarse level is available, the search procedure is further performed on the nearest neighborhood of the given parameter pair (C, γ) by using a small-size moving search window to determine the optimal parameter pair $(\bar{C}, \bar{\gamma})$. (c) With the optimal parameter pair $(\bar{C}, \bar{\gamma})$, the optimal weak SVM classifier (i.e., support vectors) is trained again by using the given training sample set. Compared with the traditional SVM classifier training techniques [Vapnik 1998; Platt 1999; Joachims 1999; Chang et al. 2002; Cristianini and Shawe-Taylor 2000], using multi-resolution moving search window for optimal model parameter estimation is able to avoid the potential pitfalls in the parameter space and result in more accurate SVM classifiers.

4.2 Unlabeled Samples for Incremental Weak SVM Classifier Training

To incorporate the unlabeled samples for training the weak SVM classifiers, we have developed an *incremental algorithm* to predict the labels of the unlabeled samples and estimate a hyperplane $< W, b >$, so that this hyperplane can separate both the labeled samples and the unlabeled samples with a maximum margin [Joachims 1999].

Our incremental SVM classifier training algorithm takes the following major steps by partitioning the unlabeled samples into three groups as shown in Figure 5: (1) For one given semantic video concept, a weak SVM classifier is first learned for each homogeneous feature subset by using the available labeled samples. (2) The weak SVM classifier is used to predict the labels for the unlabeled samples. In addition, the confidence score for the predicted label of the unlabeled sample X_j^* is calculated by applying an additional sigmoid function [Platt 1999]:

$$P(X_j^*) = \frac{1}{1 + e^{\alpha f(X_j^*) + \beta}} \tag{3}$$

where $f(X_j^*)$ is the output of the weak SVM classifier for the unlabeled sample X_j^*. The parameters α and β are determined by minimizing the negative log-likelihood (NLL) function on the validation sample set [Platt 1999]. (3) By incorporating the high-confident unlabeled samples (i.e., certain unlabeled samples) for SVM classifier training, a new SVM video classifier is learned incrementally.

Incorporating the high-confident unlabeled samples for incremental SVM classifier training may cause a small shift of the hyperplane of the SVM classifier. Thus, the new SVM classifier is used to predict the new labels for these unlabeled samples that have low confidence scores. All these unlabeled samples with significant changes of their confidence scores (i.e., informative unlabeled samples) are incorporated to enable more effective SVM classifier training. The unlabed samples without significant changes of their confidence scores (i.e., uncertain unalbeled samples) are eliminated from the training sample set because they may mislead the SVM classifier. By integrating both the certain unlabeled samples and the informative unlabeled samples for classifier training, our algorithm performs this incremental SVM classifier training procedure repeatedly until it converges.

4.3 Multi-Modal Boosting for Ensemble Classifier Training

By taking advantage of the feature hierarchy (i.e., the homogeneous feature subsets at the first level and the feature dimensions at the second level as shown in Figure 4(a)), we perform feature selection at two different levels simultaneously and learn the video classifier by using a small number of training samples. For each homogeneous feature subset, we perform PCA for feature selection and thus the intra-set feature correlation is exploited effectively. To exploit the inter-set feature correlations between various homogeneous feature subsets, a new boosting technique is developed to generate an ensemble classifier for the given semantic video concept by combining the weak SVM classifiers for different feature subsets. Thus, *inter-set feature selection* is achieved by selecting more effective weak SVM classifiers and their corresponding homogeneous feature subsets. Our proposed multi-modal boosting technique can achieve more effective ensemble classifier training and feature selection by exploiting the strong correlations between the weak classifiers for different homogeneous feature subsets of different modalities.

The ensemble classifier $G_t(X)$ for C_t is determined as:

$$G_t(X) = sign\left\{\sum_{j=1}^{11}\sum_{i=1}^{T}\alpha_j^i f_j^i(X)\right\}, \qquad \sum_{j=1}^{11}\sum_{i=1}^{T}\alpha_j^i = 1 \tag{4}$$

where the total number of homogeneous feature subsets is 11, $T = 50$ is the total number of boosting iterations, $f_j^i(X)$ is the weak SVM classifier for the jth homogeneous feature subset S_j at the ith boosting iteration. The importance factor α_j^i will be determined by $\alpha_j^i = \frac{1}{2}\log\frac{1-\epsilon_j^i}{\epsilon_j^i}$, and ϵ_j^i is the error rate for the weak SVM classifier $f_j^i(X)$ for the jth homogeneous feature subset S_j at the ith iteration. For the boosting iteration, sample weighting is performed as AdaBoost does [Freund and Schapire 1996; Tieu and Viola 2000]. For feature subset iteration, feature weighting is performed as FeatureBoost does [O'Sullivan et al. 2000].

The most representative feature set can be determined by selecting the optimal combination of the homogeneous feature subsets such that the corresponding weak SVM classifiers yield the lowest classification error rate. By selecting more effective weak SVM classifiers to boost an optimal ensemble video classifier, our proposed multi-modal boosting technique can jointly provide a novel approach for automatic feature subset selection. While most existing video classification methods suffer from the problem of *curse of dimensionality*, our proposed multi-modal boosting algorithm can take advantage of high dimensionality by incorporating the feature hierarchy to enable more effective feature selection and video classifier training.

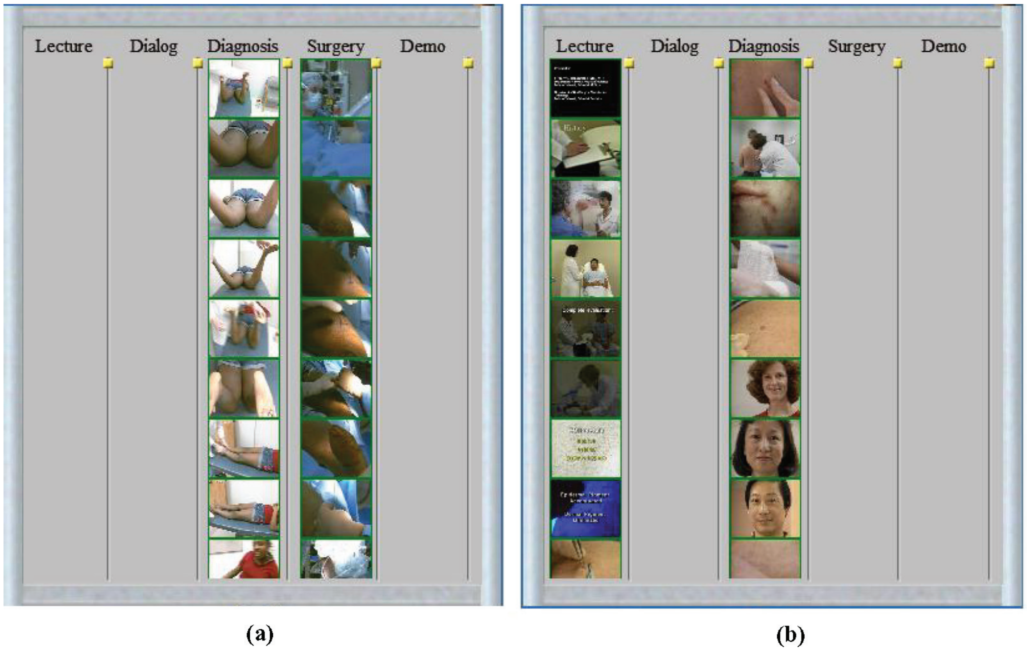


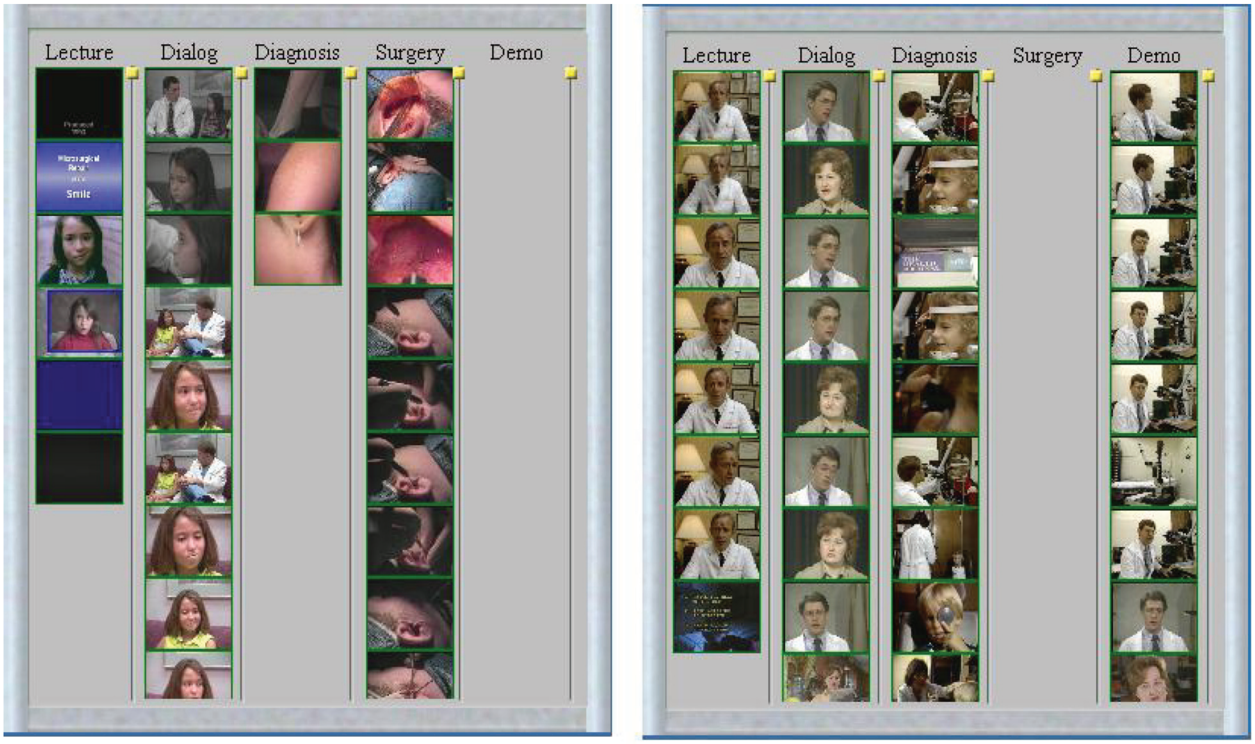
Fig. 6. The semantic video classification results: (a) one 2.12 hours medical video clip; (b) one 1.35 hours medical video clip.

The *advantages* of our proposed multi-modal boosting algorithm include: (a) Incorporating the feature hierarchy and boosting for SVM classifier training and feature subset selection can speed SVM classifier training significantly and generalize the classifiers effectively from fewer training samples, and both the training cost and the size of the training samples are reduced dramatically; (b) Partitioning the heterogeneous feature space into a set of homogeneous feature subsets can support more effective kernel function selection because the underlying statistical and geometric property of the video data can be approximated more effectively by using RBF functions; (c) Incorporating the unlabeled samples for incremental SVM classifier training can significantly reduce human efforts on labeling large-scale training samples and achieve higher classification accuracy; (d) Our multi-modal boosting algorithm is able to simultaneously boost both the training samples and the feature subsets, thus higher classification accuracy can be obtained; (e) Our proposed video classifier training scheme can have good scalability with the training sample sizes and the feature dimensions effectively.

Our current experiments focus on generating 28 semantic video concepts in a specific domain of surgery education video, such as “lecture presentation”, “gastrointestinal surgery”, “traumatic surgery”, “burn surgery”, “dialog”, “diagnosis” and “demo of medical equipment”, which are widely distributed in surgery education video. Some semantic video classification results are given in Figure 6 and Figure 7. The relevant outliers (i.e., video shots relevant to new concepts) are given in Figure 8. When the size of these outliers become large, cross-validation by medical experts may be involved to annotate them as the relevant semantic video concepts of particular interest.

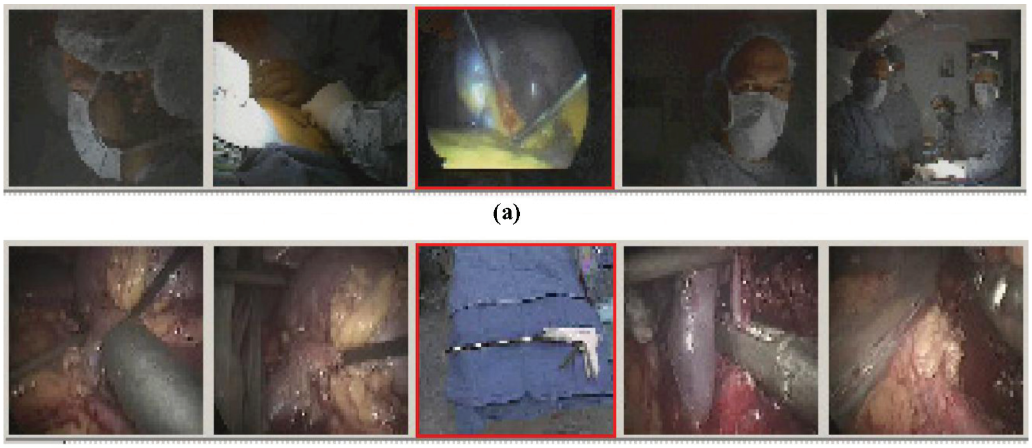
5. CONCEPT-ORIENTED VIDEO SUMMARIZATION AND SKIMMING

The medical video clips in our database are first segmented into a set of video shots and automatic salient object detection is then performed on each video shot. The video shots, that consist of the salient objects of particular interest, are defined as *principal video shots* [Fan et al. 2004]. The principal video



(a) **(b)**

Fig. 7. The semantic video classification results: (a) one 1.15 hours medical video clip; (b) one 1.5 hours medical video clip.



(a)

(b)

Fig. 8. The outliers for semantic video classification: (a) outliers for one 1.15 hours medical video clip; (b) outliers for one 1.5 hours medical video clip.

shots are classified into the relevant semantic video concepts automatically. After all the principal video shots for one specific medical video clip are mapped into the most relevant semantic video concepts, the most representative principal video shots can be selected to generate the dynamic summary (i.e., skimming) for the specific medical video clip. In addition, selecting different number of these representative principal video shots can provide the summarization and skimming results with multi-level details.

Our work on video summarization and skimming includes two parts: (a) general concept-oriented video summarization and skimming; (b) subjective video summarization and skimming. We discuss these two parts separately in the following two subsections.

5.1 General Concept-Oriented Video Summarization and Skimming

There are *two conflicting goals* for video summarization and skimming: (1) reducing the number of video frames according to the available network bandwidth; and (2) keeping the most important video contents by using a small number of video frames. In addition, the optimal number of the selected video frames, that are needed to represent the most important video contents, largely depends on the underlying semantics of the given medical video clip. To achieve concept-oriented video summarization and skimming, it is very important to assign a suitable weight for each principal video shot to define its importance according to the video semantics.

Given the maximum length of video frames for the final summarization and skimming result, we have developed a novel algorithm to achieve concept-oriented video summarization and skimming and it takes the following steps: (a) The principal video shots from a given medical video clip are first classified into the most relevant semantic video concepts automatically. (b) The importance of each principal video shot (i.e., described by using a weight) is automatically determined by analyzing the properties of the given medical video clip. (c) The principal video shots are then sorted in a descendant order of their importance (i.e. descendant order of the values of the weights). For example, the sorting result for the given medical video clip with n principal video shots is $V = \{V_1, \dots, V_n\}$, where V_1 is the most important principal video shot with the maximum value of the weights. (d) The set for the selected video shots S is first treated as an empty set $S = \emptyset$. L_{max} is defined as the maximum length of video frames for the final summarization and skimming result. (e) The summarization and skimming result for the given medical video clip is then obtained by selecting the most important principal video shots sequentially according to their values of the weights and assigning them into S. In addition, an *in-shot video summarization and skimming* technique as described below is used to obtain the summarization and skimming result for each selected principal video shot. (f) The step (e) is performed repeatedly until the length of the selected video frames is close to L_{max}. (g) Re-organize the principal video shots in S with their original display orders.

To implement this scheme for concept-oriented video summarization and skimming, the following problems should be addressed effectively: (1) Automatic weight assignment algorithm to define the importances of various semantic video concepts according to their properties. (2) In-shot summarization and skimming algorithm to prevent the very long principal video shots from dominating the final summarization and skimming results.

To assign a weight with each principal video shot in the given medical video clip, w_i^c, we use multiple independent factors of importances:

$$\begin{aligned} w_i^c &= \left\{\alpha_1 w_i^{structure} + \alpha_2 w_i^{event} + \alpha_3 w_i^{concept} + \alpha_4 w_i^{object} + \alpha_5 w_i^{length}\right\} \\ \sum_{j=1}^{5} \alpha_j &= 1. \end{aligned} \tag{5}$$

We obtain these weights independently as described below.

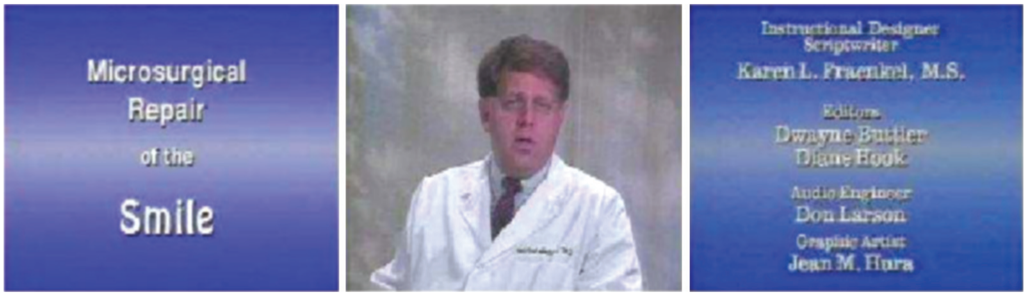


Fig. 9. Example video frames for title, introduction and final copyright notice.

Table II. Weight Assignment for Some Semantic Video Concepts

Concept	$w_i^{concept}$	Concept	$w_i^{concept}$
Outlier	1.0	Lecture	0.9
Dialog	0.8	Demo	0.8
Surgery	0.8	Diagnosis	0.8
Unknown	0.5		

Structure Weight $w_i^{structure}$. The medical video clips, that consist of some specific semantic video concepts of particular interest, may have the structure elements like title, subsection title, introduction, etc. With the help of the available salient objects and semantic video concepts, the structure elements can be effectively detected by using some statistical rules. As shown in Figure 9, one statistical rule for detecting the structure elements from the lecture video is that the principal video shot for "Title" with the salient object "Text" may always follow the principal video shot for "Lecture" with the salient object "Human Face". Thus, the structure weight for one specific principal video shot V_i is assigned as:

$$w_i^{structure} = \begin{cases} 1.0, & if\ V_i\ is\ a\ video\ structure\ element \\ 0.9, & else \end{cases} \tag{6}$$

Semantic Video Concept Weight $w_i^{concept}$. By classifying the principal video shots into the most relevant semantic video concepts, the existences of the semantic video concepts of particular interest should be selected as one of the most representative measurement to define the relative importances of video contents. In addition, the outlying principal video shots (i.e., outliers), that consist of different semantic video concepts compared with their adjacent principal video shots, should prior be selected because they generally provide additional and unusual information. Thus, we first detect the outliers that contain different semantic video concepts and a higher weight is assigned to the outliers as shown in Table II.

Semantic Medical Event Weight w_i^{event}. After the principal video shots for a given medical video clip are classified into the relevant semantic video concepts, the temporal coherence on the video semantics among the adjacent principal video shots can also be obtained. The semantic video events are then defined as the union of these adjacent principal video shots that are coherent on their semantics. Each semantic video event acts as the basic unit to characterize the semantics of the given medical video clip. Within the same semantic video event, an equal weight is assigned to each principal video shot without considering its duration time. If the weight w_i^{event} for one specific semantic video event e_j is determined by its principal video shot V_i with the shortest length t_{min}:

$$w_i^{event} = \frac{t_{min}}{t_j}, \tag{7}$$

where $V_i \in e_j$, and t_j is the length of e_j (i.e. total number of video frames in e_j).

Salient Object Weight w_i^{object}. Some specific salient objects such as "Text Area" may contain more information. Thus, a specific weight w_i^{object} is assigned to take care this. The object weight is assigned as:

$$w_i^{object} = \begin{cases} 1.0, & \text{if } V_i \text{ has "text area" object} \\ 0.9, & \text{else} \end{cases} \tag{8}$$

Shot Length Weight w_i^{length}. A very long principal video shot (e.g., more than 20 seconds) probably contains more details and a very short principal video shot (e.g. less than 1 second) may contain unimportant content in general. Thus, the principal video shots with moderate length need to be assigned a higher weight. If the length of principal video shot V_i is t_i, the shot length weight for V_i is assigned by using a normalized Maxwell distribution:

$$w_i^{length} = ct_i^2 \exp\left(1 - ct_i^2\right), \tag{9}$$

where $c = 250^{-2}$ is a constant in our experiments. It means that the principal video shots with the length near 250 will have the maximum weight.

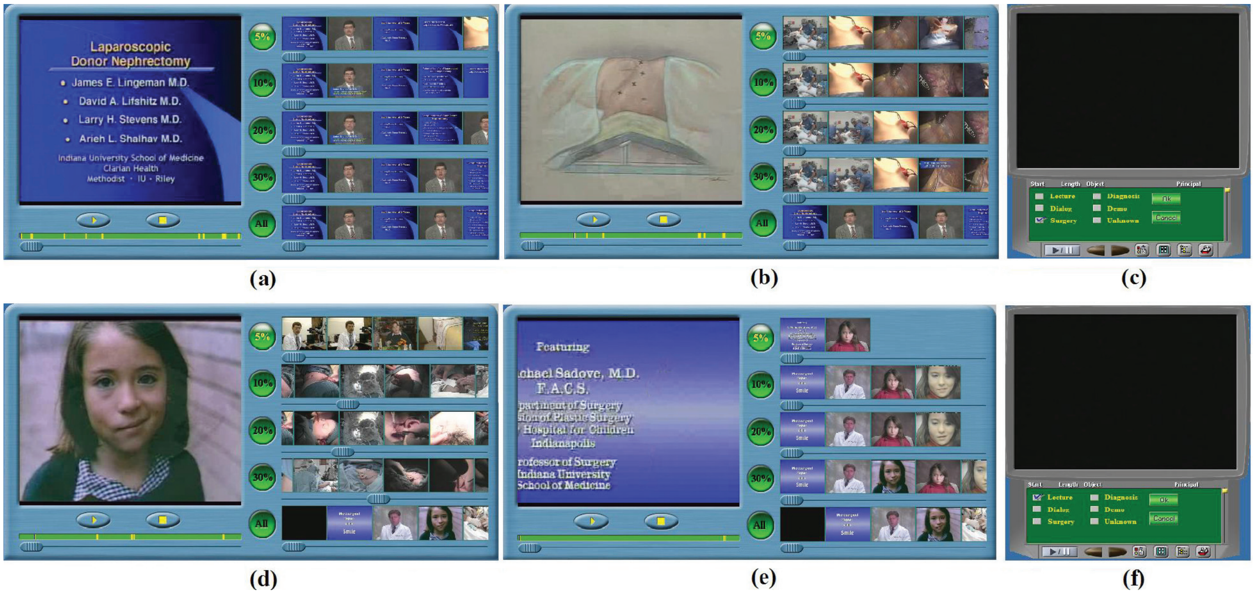
After these weights are available, the principal video shots for the given medical video clip are sorted according to their weights, and thus the most important ones are selected to generate the summarization and skimming results with multi-level details.

In order to prevent the very long principal video shots from dominating the final summarization and skimming results, our in-shot video summarization and skimming technique is performed on each selected principal vidoe shot and it is implemented by using the following statistical rules: (1) Surgery education videos have strong spatio-temporal color and motion coherency, thus many long principal video shots may apear. If the length of one specific principal video shot is less than a threshold $T = 200$, all its video frames are selected to generate the summarization and skimming. If the length for one specific principal video shot is bigger than the threshold $T = 200$, only the video frames that consist of the underlying salient objects are selected as the summarization and skimming result for the given principal video shot. (2) For online surgery education pruposes, the principal video shots for lecture presentation should have the first priority to be selected as the final summarization and skimming result. If one certain principal video shot consists of the salient object "Text Area", the video frames that consist of the salient object "Text Area" are selected as the summarization and skimming result for this principal video shot.

To evaluate the effectiveness and correction of the settings of these statistical rules (weights), we have involved medical experts to evaluate our video summarization and skimming results which are obtained by using these statistical rules. The conclusions from our two medical experts are very positive. Obviously, more user study should be performed with IRB approval.

5.2 Subjective Video Summarization and Skimming

For online medical education applications, it is very attractive to achieve subjective video summarization and skimming because medical students may have precise ideas of what they really want to see from a given medical video clip. To incorporate the users' preferences for achieving concept-oriented video summarization and skimming, the correlation w_i^u, between the users' preferences and the semantic video concepts in the given medical video clip, can be used to define the relative importances of different



(c)
□ Lecture □ Diagnosis
□ Dialog □ Demo
☑ Surgery □ Unknown

(f)
☑ Lecture □ Diagnosis
□ Dialog □ Demo
□ Surgery □ Unknown

Fig. 10. The subjective concept-oriented summarization and skimming results with multi-level details for two video clips: (a) results for one specific semantic video concept “lecture presentation”; (b) results for one specific semantic video concept “trauma surgery”; (c) system interface for users to define their preference; (d) results for one specific semantic video concept “trauma surgery”; (e) results for one specific semantic video concept “lecture presentation”; (f) system interface for users to define their preference.

principal video shots. In Section 5.1, we have already assigned the weight, w_i^c, to the principal video shots according to the video semantics. Thus, the overall weight of a given principal video shot, w_i, can be defined by:

$$w_i = w_i^c \times w_i^u. \tag{10}$$

If the medical students do not have any specific preference, the preference weights w_i^u for all these principlal video shots are assigned to 1.0. If the medical students have specific preferences, the subjective video summarization and skimming for the specific semantic video concept can be achieved automatically by assigning different preference weights to the principal video shots according to their classification results:

$$w_i^u = \begin{cases} 1.0, & interesting\ \ concept \\ 0.0, & else. \end{cases} \tag{11}$$

After the weights of the importances for all the principal video shots are available, our proposed video summarization and skimming algorithm can be used to select the principal video shot with non-zero value of w_i and produce subjective video summarization and skimming according to users’ preferences.

Our system interface for capturing the users’ prefereces to support subjective concept-oriented video summarization and skimming is given in Figure 10, where the users have full control on the interestness of the skimming results.

6. ALGORITHM EVALUATION AND EXPERIMENTAL RESULTS

We have tested our algorithms on a video database including more than 7000 principal video shots that are obtained from 25 hours of MPEG surgery education videos. For evaluating the effectiveness

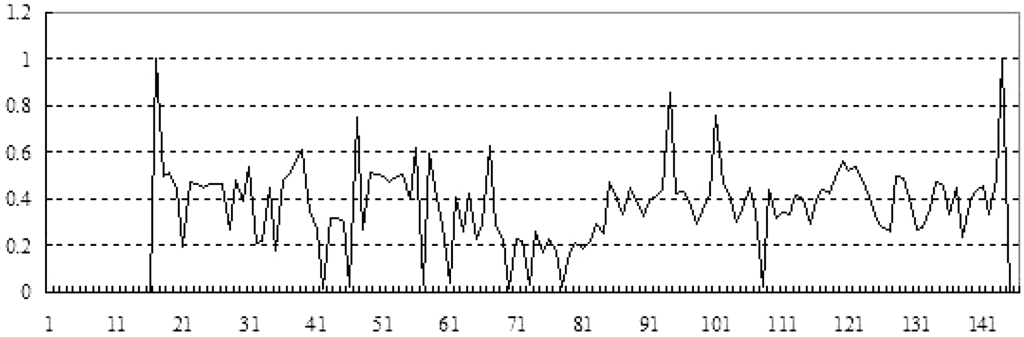


Fig. 11. w_i^c for a given surgery education video with structure.

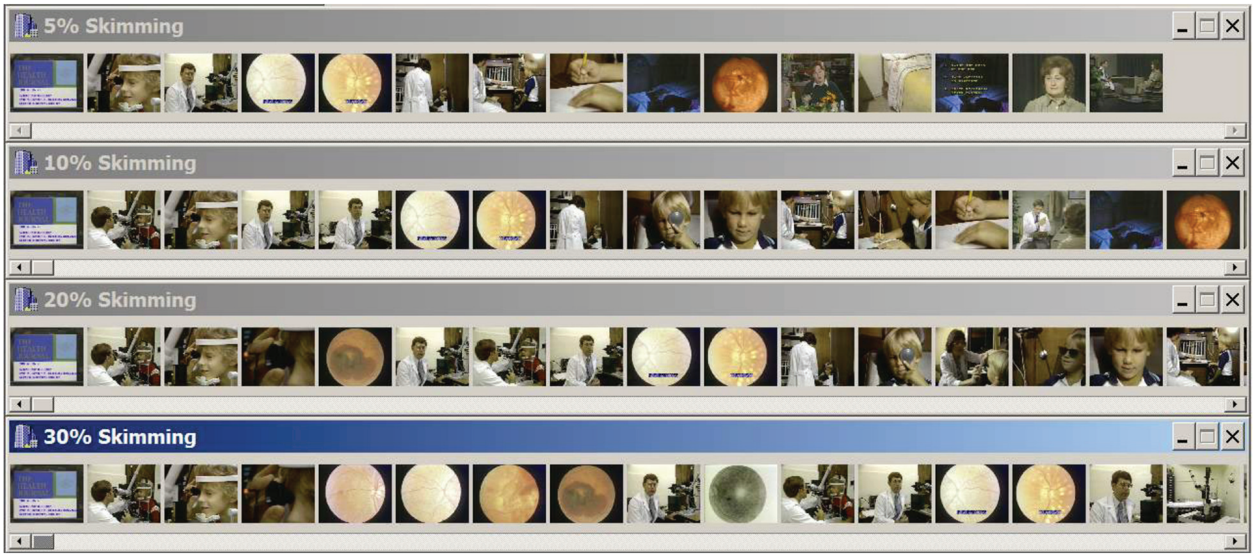


Fig. 12. The general concept-oriented summarization and skimming results with multi-level details.

of our classifier training technique, 4200 principal video shots are used as the training samples and 2800 principal video shots are used as the test samples. In addition, the summarization and skimming results for these 2800 test video clips are automatically generated and evaluated by our medical experts. Due to the limitation of pages, only part of results are given below to show the effectiveness of our algorithms.

6.1 Concept-Oriented Video Summarization and Skimming

To enable concept-oriented video summarization and skimming, the first step is to assign the importance (weight) for each principal video shot. Figure 11 shows the automatic weight assignment for one test surgery education video with structure element. For surgery education videos, the weight for the first principal video shot is assigned with a value close to 1.0 because it is a principal video shot with lecture title, and the weight for the last principal video shot is assigned with a value close to 1.0 because it shows the doctor saying good-bye to the medical students. Other principal video shots with very high weights are the outliers. Obviously, all these weights are obtained automatically by using

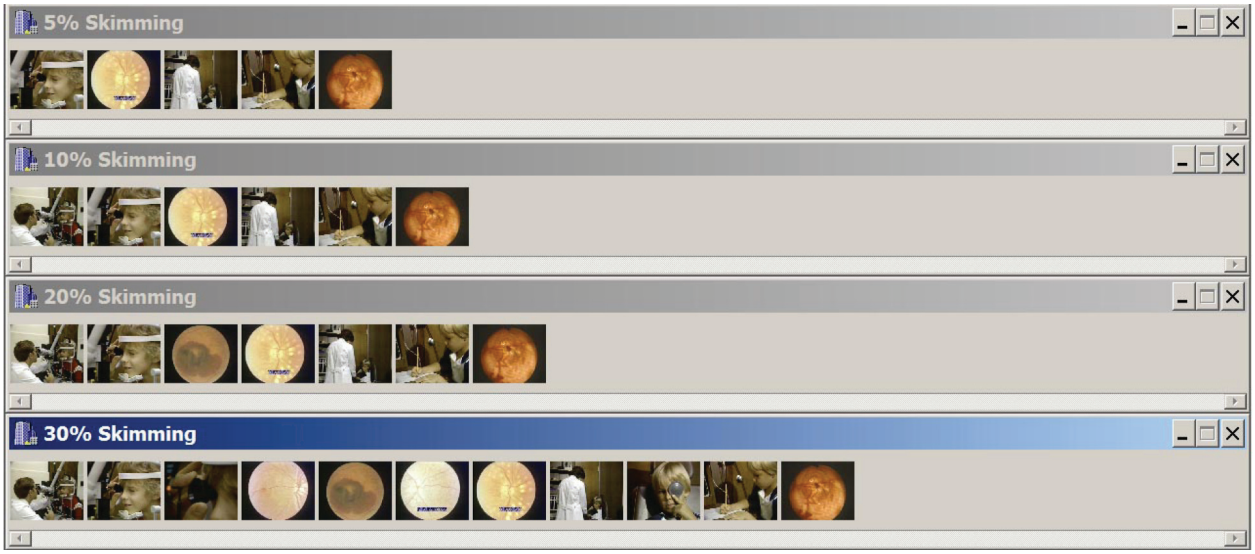


Fig. 13. The subjective concept-oriented summarization and skimming results with multi-level details.

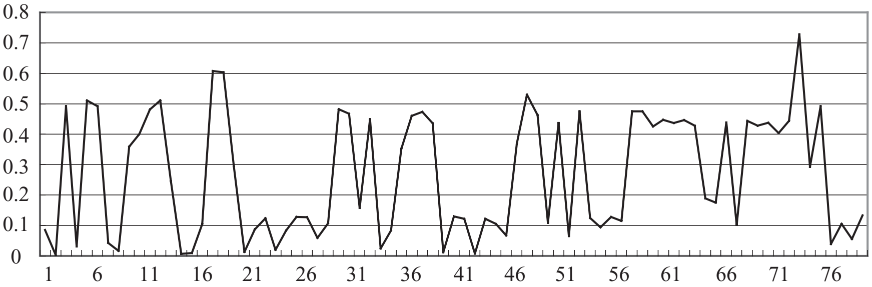


Fig. 14. w_i^c for a given surgery education video without structure.

our proposed technique. Under these weights, the general concept-oriented summarization and skimming results with multi-level details are obtained for the given surgery education video clip (shown in Figure 12). One can find that our concept-oriented video summarization and skimming technique is able to preserve both the structural elements and the important semantic video concepts with different details. Figure 13 shows the subjective concept-oriented summarization and skimming results for the same video clip, where the student is interested in only the semantic video concept "Diagnosis". One can observe that our subjective concept-oriented video summarization and skimming technique is very attractive for the users who have specific preferences on the final summarization results.

Figure 14 gives the weight assignment for another test surgery education video clip without structure element. The general concept-oriented summarization and skimming results for this test medical video clip are given in Figure 15. Figure 16 is the subjective concept-oriented summarization and skimming results for the same video clip where the studdent is only interested in the semantic video concept "Surgery".

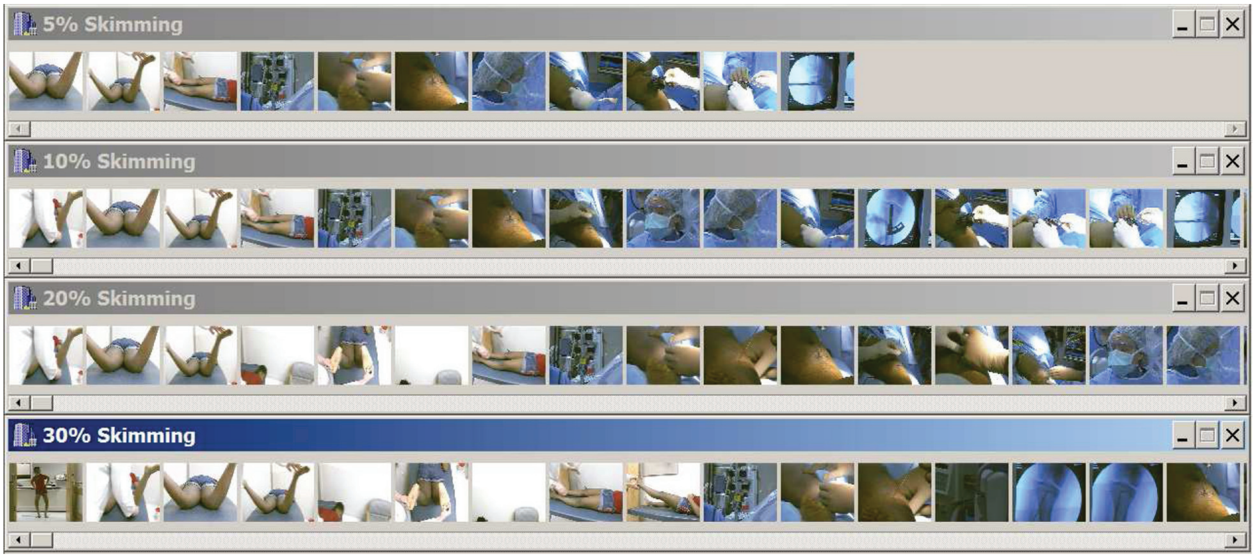


Fig. 15. The general concept-oriented summarization and skimming results with multi-level details.

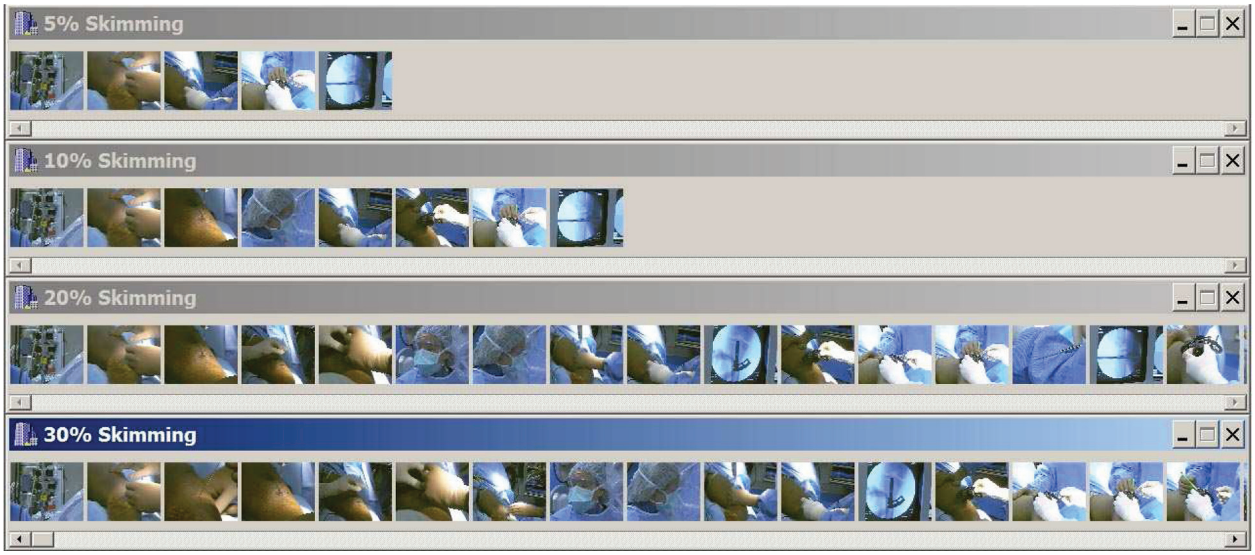


Fig. 16. The subjective concept-oriented summarization and skimming results with multi-level details.

6.2 Video Classifier Evaluation

Since the performance of our new framework for concept-oriented video summarization and skimming largely depends on the performance of our classifiers, our work on classifier evaluation focuses on: (1) Evaluating the performance difference of our video classifiers by using two types of video patterns for feature extraction: *salient objects* versus *video shots*. (2) Evaluating the performance of our SVM classifier training technique. (3) Evaluating the performance difference of our video classifiers by using different sizes of unlabeled samples for classifier training.

Table III. The Average Performance Differences (i.e., ξ versus ζ) for Our Classifiers

Concepts	Lecture	Trauma Surgery	Diagnosis
salient	83.2%	85.4%	81.6%
objects	82.4%	83.1%	90.1%
video	75.4%	76.9%	78.8%
shots	77.3%	73.5%	77.5%
Concepts	Gastrointestinal Surgery	Dialog	Burn Surgery
salient	88.3%	78.9%	82.9%
objects	90.6%	80.3%	84.5%
video	80.3%	73.3%	75.7%
shots	78.2%	75.9%	78.5%

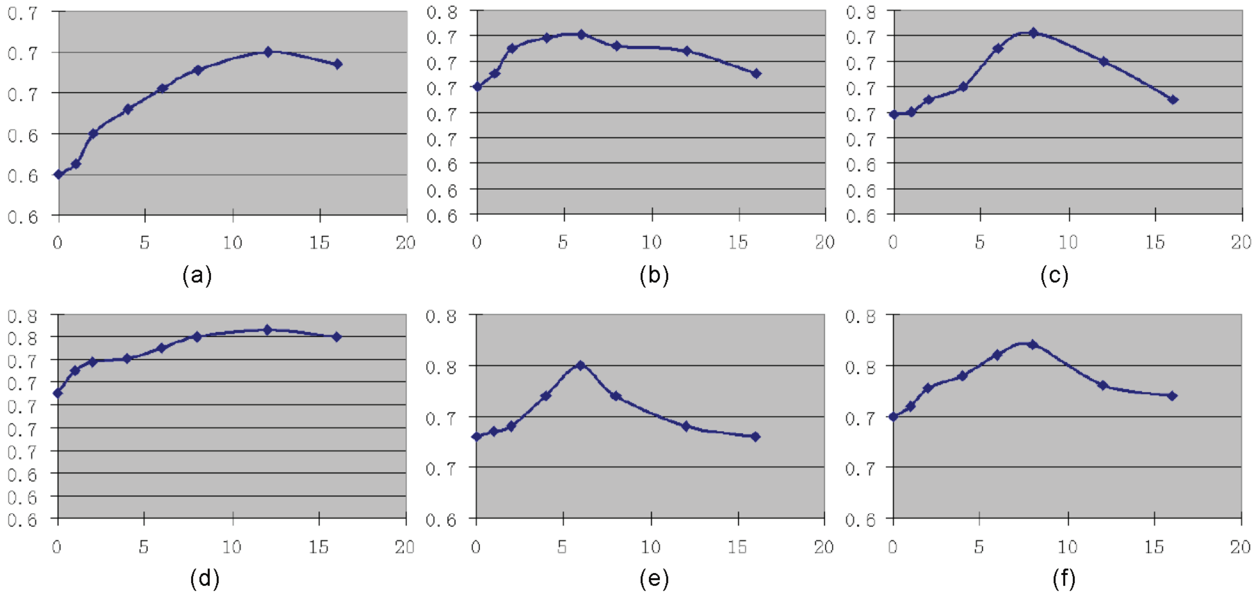


Fig. 17. The classifier performance (i.e., precision ξ) with different ratio $\lambda' = \frac{N_u}{N_L}$ between the unlabeled samples N_u and the labeled samples N_L: (a) lecture; (b) dialog; (c) trauma surgery; (d) gastrointestinal surgery; (e) burn surgery; (f) diagonosis.

The *benchmark metric* for classifier evaluation includes *precision* ξ and *recall* ζ as defined in Eq. (1). The average performance of our classifiers is given in Table III, one can find that using salient objects for video content representation and feature extraction can improve the classifier performance significantly.

In our current implements, we label 120 principal video shots for each atomic video concept node. Given a limited number of labeled samples, we have tested the performance of our classifiers by using different sizes of unlabeled samples for classifier training (i.e., with different size ratios $\lambda = \frac{N_u}{N_L}$ between the unlabeled samples N_u and the labeled samples N_L). The average performance differences for some semantic video concepts are given in Figure 17. One can find that incorporating the unlabeled samples for classifier training is able to improve the classification accuracy significantly, but using more unlabled samples may also decrease the performance [Cohen et al. 2004]. In our experiments, when $\lambda \geq 8$, the performance of the classifiers for most semantic video concepts start to decrease.

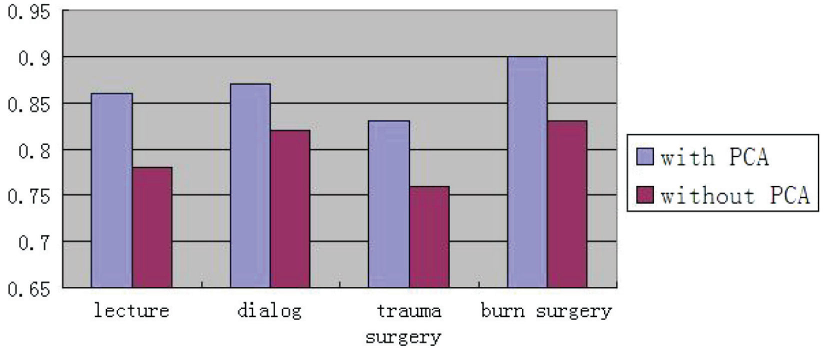


Fig. 18. The performance difference of our video classifiers with and without performing PCA.

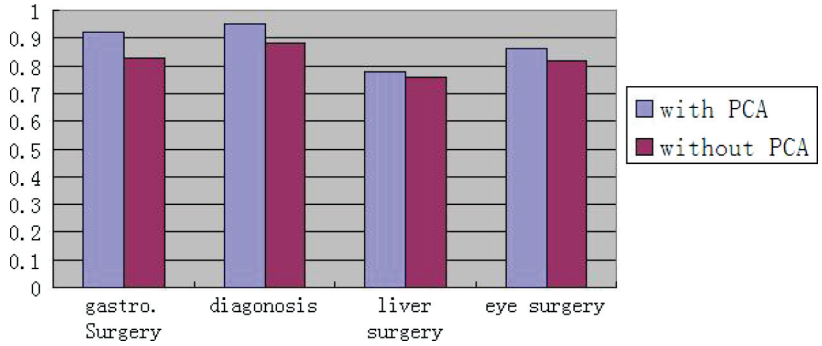


Fig. 19. The performance difference of our video classifiers with and without performing PCA.

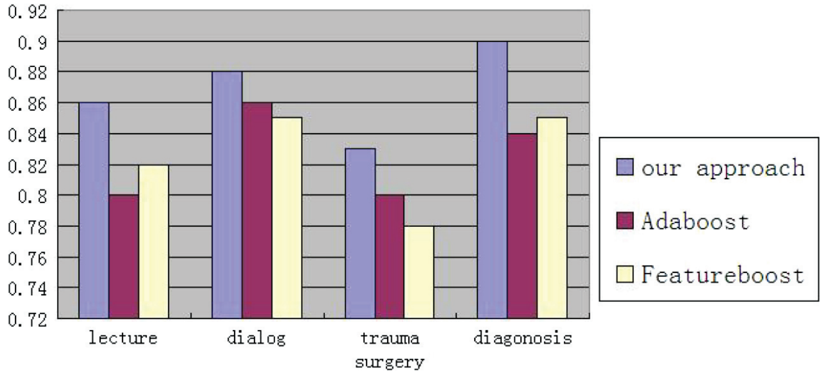


Fig. 20. The performance difference between our proposed boosting technique, Adaboost, and FeatureBoost.

We have also compared the performance difference of our ensemble classifier by performing only the second-level feature selection and performing both the first-level and second-level feature selection. As shown in Figure 18 and Figure 19, one can find that performing both the first-level and the second-level feature selection (i.e., both PCA and feature subset selection) is able to improve the accuracy of the ensemble classifier significantly.

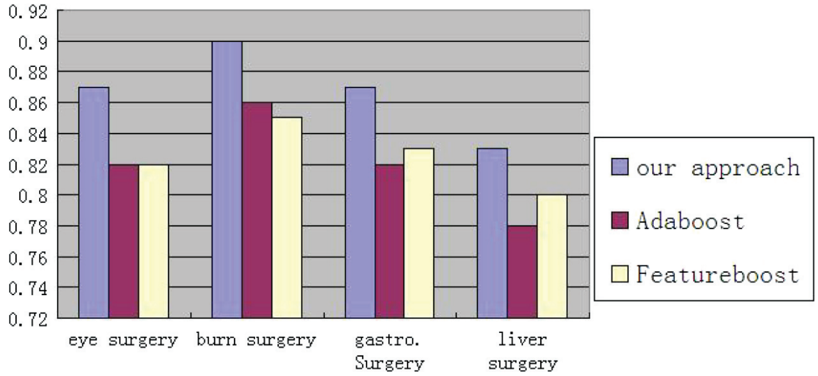


Fig. 21. The performance difference between our proposed boosting technique, Adaboost, and FeatureBoost.

We have also compared the performance differences between multiple approaches for ensemble classifier training: our multi-modal boosting algorithm, AdaBoost [Freund and Schapire 1996; Tieu and Viola 2000], and FeatureBoost [O'Sullivan et al. 2000]. As mentioned above, our multi-modal boosting algorithm has taken the advantages of both AdaBoost and FeatureBoost, and thus higher classification accuracy is expected. As shown in Figure 20 and Figure 21, one can find that our multi-modal boosting algorithm can obtain higher classification accuracy than AdaBoost and FeatureBoost.

7. CONCLUSIONS AND FUTURE WORKS

To incorporate the results of semantic video classification for concept-oriented video summarization and skimming, we have developed a novel scheme to achieve more effective video classifier training by incorporating the feature hierarchy, boosting and the unlabeled samples to generalize the SVM video classifiers from fewer training samples. Our scheme for semantic video classification has also provided a good scheme for achieving semantic video understanding and concept-oriented video summarization and skimming. In addition, our techniques will be very attractive to support semantic video retrieval for *online multimedia medical education*. The online demo is also available at: http://www.cs.uncc.edu/~jfan/video.html.

It is also worth noting that the definitions of semantic video concepts and salient objects are largely domain-dependent, but our proposed algorithms can be easily extended to other video domains such as news and films by selecting the suitable domain-dependent semantic video concepts and the relevant salient objects.

Our proposed algorithm for weight determination is now based on statistical rules, we plan to improve this algorithm by: (a) proposing a weight optimization technique to obtain the relative importance between these weights automatically; (b) developing more generic tools for weight determination. With IRB approval, we will also perform user study at nursing school at UNC-Charlotte.

ACKNOWLEDGMENTS

The authors thank the reviewers for their useful comments and insight suggestions to make this paper more readable. We also want to thank the Associate Editor Prof. Klara Nahrstedt for her comments and handling the review process of our paper.

REFERENCES

Adames, B., Dorai, C., and Venkatesh, S. 2002. Towards automatic extraction of expressive elements of motion pictures: Tempo. *IEEE Trans. Multimedia 4*, 4, 472–481.

Adams, W., Iyengar, G., Lin, C.-Y., Naphade, M., Neti, C., Nock, H., and Smith, J. 2003. Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP J. Appl. Sig. Proc. 2*, 1–16.

Alatan, A., Onural, L., Wollborn, M., Mech, R., Tuncel, E., and Sikora, T. 1998. Image sequence analysis for emerging interactive multimedia services-the european cost 211 framework. *IEEE Trans. Circ. Syst. Video Tech. 8*, 7, 802–813.

Arman, F., Depommier, R., Hsu, A., and Chiu, M. 1994. Content-based browsing of video sequences. In *ACM Multimedia*. ACM, New York, 97–103.

Chang, E., Goh, K., Sychay, G., and Wu, G. 2002. CBSA: Content-based annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. Circ. Syst. Video Tech. 13*, 1, 26–38.

Chang, S.-F. 2002. Optimal video adaptation and skimming using a utility-based framework. In *Proceedings of the International Tyrrhenian Workshop on Digital Communications*.

Chang, S.-F., Chen, W., and Sundaram, H. 1998. Semantic visual templates: linking visual features to semantics. In *Proceedings of the International Conference on Image Processing*. Vol. 3. IEEE Computer Society Press, Los Alamitos, CA, 531–535.

Cohen, I., Sebe, N., Cozman, F., Cirelo, M., and Huang, T. 2004. Semi-supervised learning of classifiers: Theory and algorithms and their applications to human-computer interaction. *IEEE Trans. Patt. Anal. Mach. Intel. 26*, 12, 1553–1567.

Correia, P. and Pereira, F. 2004. Classification of video segmentation application scenarios. *IEEE Trans. Circ. Syst. Video Tech. 14*, 5, 735–741.

Cristianini, N. and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, MA.

Deshpande, S. and Hwang, J.-N. 2001. A real-time interactive virtual classroom multimedia distance learning system. *IEEE Trans. Multimed. 3*, 4, 432–444.

Dimitrova, N., Agnihotri, L., and Wei, G. 2000. Video classification based on hmm using text and faces. In *ACM Multimedia*. ACM, New York, 499–500.

Djeraba, C. 2000. When image indexing meets knowledge discovery. In *MDM/KDD*. ACM, New York, 73–81.

Djeraba, C. 2002. *Multimedia Mining: A Highway to Intelligent Multimedia Documents*. Kluwer.

Ebadollahi, S., Chang, S.-F., and Wu, H. 2002. Echocardiogram videos: Summarization, temporal segmentation and browsing. In *Proceedings of the International Conference on Image Processing*. IEEE Computer Society Press, Los Alamitos, CA, I–613–I–616.

Ekin, A., Tekalp, A., and Mehrotra, R. 2003. Automatic soccer video analysis and summarization. *IEEE Trans. Image Process. 12*, 796–807.

Fan, J., Luo, H., and Elmagarmid, A. 2004. Concept-oriented indexing of video database toward more effective retrieval and browsing. *IEEE Trans. Image Process. 13*, 7, 974–992.

Fan, J., Yau, D., Elmagarmid, A., and Aref, W. 2001. Image segmentation by integrating color edge detection and seeded region growing. *IEEE Trans. Image Process. 10*, 1454–1466.

Fan, R.-E., Chen, P.-H., and Lin, C.-J. 2005. Working set selection using the second order information for training svm. *J. Mach. Learn. Res. 6*, 1889–1918.

Fischer, S., Lienhart, R., and Effelsberg, W. 1995. Automatic recognition of film genres. In *ACM Multimedia*. ACM, New York, 367–368.

Freund, Y. and Schapire, R. 1996. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 148–156.

Gatica-Perez, D., Loui, A., and Sun, M.-T. 2003. Finding structure in home videos by probabilistic hierarchical clustering. *IEEE Trans. Circ. Syst. Video Tech. 13*, 6, 539–548.

Greenspan, H., Goldberger, J., and Mayer, A. 2004. Probabilistic space-time video modeling via piecewise gmm. *IEEE Trans. Patt. Anal. Mach. Intel. 26*, 3, 384–396.

Haering, N., Qian, R., and Sezan, M. 2000. A semantic event-based detection approach and its application to detecting hunts in wildlife video. *IEEE Trans. Circ. Syst. Video Tech. 10*, 6, 857–868.

Hanjalic, A., Lagendijk, R., and Biomond, J. 1999. Automated high-level movie segmentation for advanced video retrieval system. *IEEE Trans. Circ. Syst. Video Tech. 9*, 4, 580–588.

He, L., Sanocki, E., Gupta, A., and Grudin, J. 1999. Auto-summarization of audio-video presentations. In *ACM Multimedia*. ACM, New York, 489–498.

Jaimes, A. and Chang, S. 2001. Learning structured visual detectors from user input at multiple levels. *Int. J. Image Graph. 1*, 3, 415–444.

Joachims, T. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the International Conference on Machine Learning*. Morgan, Kaufmann, San Francisco, CA, 200–209.

KENDER, J. AND YEO, B.-L. 1998. Video scene segmentation via continuous video coherence. In Proceedings of the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society Press, Los Alamitos, CA, 367–373.

LEW, M. 2001. *Principles of Visual Information Retrieval*. Springer-Verlag, New York.

LI, Y., PARK, Y., AND DORAI, C. 2006. Atomic topical segments detection for instructional videos. In *ACM Multimedia*. ACM, New York, 53–56.

LIU, T. AND KENDER, J. 2004. Lecture videos for e-learning: Current research and challenges. In *IEEE International Symposium on Multimedia Software Engineering*. IEEE Computer Society Press, Los Alamitos, CA, 574–578.

LIU, Z., WANG, Y., AND CHEN, T. 1998. Audio feature extraction and analysis for scene segmentation and classification. *J. VLSI Signal Process. Syst. 20*, 1, 61–79.

LUO, H., FAN, J., GAO, Y., AND XU, G. 2004. Multimodal salient objects: General building blocks of semantic video concepts. In *Proceedings of the International Conference on Image and Video Retrieval*. Springer, Berlin / Heidelberg, Germany, 374–383.

MA, Y., LU, L., ZHANG, H., AND LI, M. 2002. A user attention model for video summarization. In *ACM Multimedia*. ACM, New York, 533–542.

NAPHADE, M. AND HUANG, T. 2001. A probabilistic framework for semantic video indexing, filtering, and retrival. *IEEE Trans. Multimed. 3*, 141–151.

O'SULLIVAN, J., LANGFORD, J., AND BLUM, A. 2000. Featureboost: A meta learning algorithm that improves model robustness. In *Proceedings of the International Conference on Machine Learning*. Morgan, Kaufmann, San Francisco, CA, 703–710.

PFEIFFER, S., LIENHART, R., AND EFFELSBERG, W. 1999. Scene determination based on video and audio features. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, Vol. 15. IEEE Computer Society Press, Los Alamitos, CA, 685–690.

PLATT, J. 1999. *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*. Adavances in Large Margin Classifiers, MIT Press, Cambridge, MA.

QI, Y., LIU, T., AND HAUPTMANN, A. 2003. Supervised classification of video shot segmentation. In *International Conference on Multimedia and Expo*. IEEE Computer Society Press, Los Alamitos, CA, II–689–92.

SEBE, N., LEW, M., AND SMEULDERS, A. 2003. Video retrieval and summarization. *Comput. Vision Image Understand. 92*, 2, 146–152.

SMITH, M. AND KANADE, T. 1995. Video skimming for quick browsing based on audio and image characterization. Tech. rep., CMU: TR-CMU-CS-95-186.

SNOEK, C. AND MORRING, M. 2003. Multimodal video indexing: A state of the art review. *Multimed. Tools Appl. 25*, 1, 5–35.

SUNDARAM, H. AND CHANG, S. 2002a. Computable scenes and structures in films. *IEEE Trans. Multimed. 4*, 482–491.

SUNDARAM, H. AND CHANG, S.-F. 2002b. Video skims: Taxonomies and an optimal generation framework. In *Proceedings of the International Conference on Image Processing*. IEEE Computer Society Press, Los Alamitos, CA, II–21–II–24.

SUNDARAM, H., XIE, L., AND CHANG, S.-F. 2002. A unility framework for the automatic generation of audio-visual skims. In *ACM Multimedia*. ACM, New York, 189–198.

TIEU, K. AND VIOLA, P. 2000. Boosting image retrieval. *Int. J. Comput. Vision 56*, 1, 17–36.

VAPNIK, V. 1998. *Statistical Learning Theory*. Wiley-Interscience, New York.

XIE, L., XU, P., CHANG, S., DIVAKARAN, A., AND SUN, H. 2003. Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognition Letters 24*, 767–775.

ZHANG, D. AND NUNAMAKER, J. 2004. A natural language approach to content-based video indexing and retrieval for interactive e-learning. *IEEE Trans. Multimed. 6*, 3, 450–458.

ZHANG, H., KANKANHALLI, A., AND SMOLIAR, S. 1993. Automatic parsing of video. In *International Conference on Multimedia Systems*. Vol. 1. IEEE Computer Society Press, Los Alamitos, CA, 45–54.

ZHOU, W., VELLAIKAL, A., AND KUO, C. 2000. Rule-based video classification system for basketball video indexing. In *ACM Multimedia*. ACM, New York, 213–216.

Received June 2006; revised December 2006, June 2007, September 2007; accepted October 2007