

Incorporating Concept Ontology for Hierarchical Video Classification, Annotation, and Visualization

Jianping Fan, Hangzai Luo, Yuli Gao, and Ramesh Jain

Abstract—Most existing content-based video retrieval (CBVR) systems are now amenable to support automatic low-level feature extraction, but they still have limited effectiveness from a user's perspective because of the semantic gap. Automatic video concept detection via semantic classification is one promising solution to bridge the semantic gap. To speed up SVM video classifier training in high-dimensional heterogeneous feature space, a novel *multimodal boosting* algorithm is proposed by incorporating feature hierarchy and boosting to reduce both the training cost and the size of training samples significantly. To avoid the inter-level error transmission problem, a novel *hierarchical boosting* scheme is proposed by incorporating concept ontology and multitask learning to boost hierarchical video classifier training through exploiting the strong correlations between the video concepts. To bridge the semantic gap between the available video concepts and the users' real needs, a novel hyperbolic visualization framework is seamlessly incorporated to enable intuitive query specification and evaluation by acquainting the users with a good global view of large-scale video collections. Our experiments in one specific domain of *urgery education videos* have also provided very convincing results.

Index Terms—Concept ontology, hierarchical boosting, hyperbolic visualization, multimodal boosting, multitask learning, semantic gap, video classification and annotation.

I. INSTRUCTION

DIGITAL VIDEO now plays an increasingly pervasive role in supporting evidence-based medical education by illustrating the most relevant clinic examples in video to students [1]. To do this, it has become increasingly important to have mechanisms that can classify, summarize, index and search medical video clips at the semantic level. Unfortunately, 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 [3]–[6].

Automatic video concept detection via semantic classification is one promising approach to bridge the semantic gap, but its performance largely depends on two inter-related issues: 1) suitable frameworks for video content representation and

Manuscript received September 28, 2006; revised April 14, 2007. This work was supported by the National Science Foundation under Grants 0601542-IIS and 0208539-IIS and by a grant from AO Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Guus Schreiber.

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

R. Jain is with the School of Information and Computer Science, University of California, Irvine, CA 92623 USA (e-mail: jain@ics.uci.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2007.900143

feature extraction and 2) effective algorithms for video classifier training. To address the first issue, the underlying video patterns for feature extraction should be able to capture the video semantics at the object level effectively and efficiently (i.e., semantics for interpreting the real world physical objects in a video clip) [6]–[10]. Using the segmentation of real world physical objects (i.e., semantic video objects) for feature extraction can significantly enhance the ability of perceptual features on discriminating various video concepts. Unfortunately, automatic detection of large amounts of semantic video objects with diverse perceptual properties is still an open problem in computer vision. To address the second issue, robust video classifier training techniques are needed to tackle both the *intra-concept variations* and *inter-concept similarity* effectively.

Ideally, extracting high-dimensional perceptual features for video content representation and classifier training has more capacity to effectively characterize various perceptual properties of video concepts. Thus, using high-dimensional perceptual features may enhance the classifier's ability on discriminating various video concepts and result in higher classification accuracy. However, learning the video classifiers in high-dimensional feature space requires large amounts of training samples that increase exponentially with the feature dimensions [33], [34]. On the other hand, the computational complexity for classifier training also increases exponentially with the size of training samples. Therefore, it is very expensive to directly learn the video classifiers by using high-dimensional perceptual features.

Another challenge for automatic video concept detection is that one single video clip may contain different meanings at multiple semantic levels [52]–[56], [71], [72], thus the video clips may be similar at different semantic levels. The concept ontology offers an effective way for interpreting a basic vocabulary of domain-dependent video concepts and their contextual relationships, and thus it can be integrated to enable more effective video classifier training [1], [2]. Therefore, the classifiers for the video concepts at the higher levels of the concept ontology (i.e., high-level video concepts with larger within-concept variations) can be learned effectively by combining the classifiers for the relevant video concepts at the lower levels of the concept ontology (i.e., low-level video concepts with smaller within-concept variations) [1], [2]. With the help of the concept ontology, naive users can also specify their queries more precisely and unambiguously, which may further lead to better precision and recall rates in the loop of video retrieval. One major problem for hierarchical video classifier training is that the classification errors may be transmitted among different concept levels, i.e., *inter-level error transmission problem*, thus integrating the concept ontology for hierarchical video classifier training may sometimes lead to worse performance rather than improvement [1].

When hierarchical mixture model is used for video classifier training [1], [2], there is an implicit assumption that the sibling video concepts under the same parent node are characterized by the same set of perceptual features or the perceptual features are assumed to be independent. However, such assumptions may not be true in real applications because different video concepts may have different principal properties and the perceptual features are strongly correlated. Based on these observations, there is a great need to develop new schemes for hierarchical video classifier training, which are able to reduce both the computational complexity and the size of training samples significantly while addressing the inter-level error transmission problem effectively.

In this paper, we have proposed a novel scheme to achieve automatic video concept detection via hierarchical classification. To enable more effective video classifier training in high-dimensional heterogeneous feature space, a *multimodal boosting* algorithm is proposed to reduce the size of training samples dramatically, speed up SVM classifier training significantly, and choose more appropriate kernel functions and select more representative feature subsets for video classifier training. A *hierarchical boosting* algorithm is developed to effectively address the inter-level error transmission problem, while significantly reducing the computational complexity for training the classifiers for large amounts of video concepts through exploiting their strong inter-concept correlations. To bridge the semantic gap between the available video concepts (which are detected automatically by using our hierarchical video classification scheme) and the users' real needs, a novel *hyperbolic visualization* scheme is proposed to visually acquaint the users with a good global view of large-scale video collections and enable intuitive query specification and evaluation.

The paper is organized as follows. Section II briefly introduces our framework for hierarchical video concept organization. Section III presents our work on using salient objects and concept ontology to bridge the semantic gap hierarchically. Section IV describes our new scheme for hierarchical video classifier training and multimodal feature subset selection. Section V presents our new algorithm to achieve hierarchical video classification and automatic multilevel video annotation. Section VI introduces our hyperbolic visualization framework for intuitive query specification and evaluation. Section VII introduces our results on algorithm and system evaluation. Section VIII points out the scalability and generalizability of our proposed algorithms. We conclude in Section IX.

II. CONCEPT ONTOLOGY FOR HIERARCHICAL VIDEO CONCEPT ORGANIZATION

As large-scale video collections come into view, there is a growing need to enable computational interpretation of video semantics and achieve automatic video annotation. Because one single video clip may contain different meanings at multiple semantic levels, a successful implementation of such automatic video annotation systems also requires a good structure to enable hierarchical video concept organization [52]–[56], [71], [72]. The artificial intelligence community has incorporated the concept ontology for high-level knowledge representation [11], [12], [40]–[44]. The concept ontologies have also been employed to achieve better precision and recall in text clas-

sification and retrieval systems [45]–[51]. All these existing concept ontologies are simply characterized by single-modal parameter, i.e., text terms (keywords). On the other hand, the video concepts should be characterized by multimodal parameters because the keywords (text terms) may be too abstract to interpret the details of video semantics effectively and efficiently. Therefore, building multimodal concept ontology for hierarchical video concept organization is nontrivial [52]–[56], [71], [72]. Projects dealing with some aspects of these themes on videos include the Informedia project at Carnegie Mellon University [13]–[16], the Advent project at Columbia University [17]–[19], works done at the University of South California [20], IBM research [23]–[26], Dublin Core [27], University of Amsterdam [28]–[31], and the University of Illinois [21], [22]. Most of these projects focus on the video domains of broadcast news, sports and films which have rich production metadata and editing structures.

Our proposed work significantly differ 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 metadata. Because large amounts of real clinical examples in video are illustrated for student training [1], surgery education videos are significantly different from traditional lecture videos [32]. b) Salient objects and their cumulative volumetric features are used to effectively characterize the video semantics at the object level, while significantly reducing the computational complexity for video analysis [6]. c) A novel multimodal boosting algorithm is proposed to speed up SVM video classifier training and generalize the video classifiers from fewer training samples. d) A novel hierarchical boosting scheme is developed to boost hierarchical video classifier training and scale up our statistical learning techniques to large amounts of video concepts through exploiting their strong inter-concept correlations. e) A new top-down scheme is proposed to achieve hierarchical video classification with automatic error recovery. f) A hyperbolic visualization scheme is proposed to acquaint users with a good global view of large-scale video collections and enable intuitive query specification and evaluation.

Our concept ontology consists of three key issues: 1) *video concept nodes*; 2) *multimodal concept properties*; and 3) *contextual and logical relationships* between an upper concept node and its children concept nodes. Our concept ontology is used to interpret a basic vocabulary of domain-dependent video concepts and their contextual and logical relationships, where the multimodal properties for each video concept are further characterized by multimodal parameters. The deeper the level of the concept ontology, the narrower the coverage of semantic subjects. Thus, the video concepts at the deeper level of the concept ontology can represent more specific semantic subjects with smaller within-concept variations. On the other hand, the video concepts at the upper level of the concept ontology can cover more general semantic subjects with larger within-concept variations. Therefore, it is very expensive to directly train the classifiers for detecting the high-level video concepts with larger within-concept variations. The deepest level of the concept ontology (i.e., leaf nodes) is named as *atomic video concepts*, which are used to interpret the most specific semantic subjects with the smallest within-concept variations. The perceptual

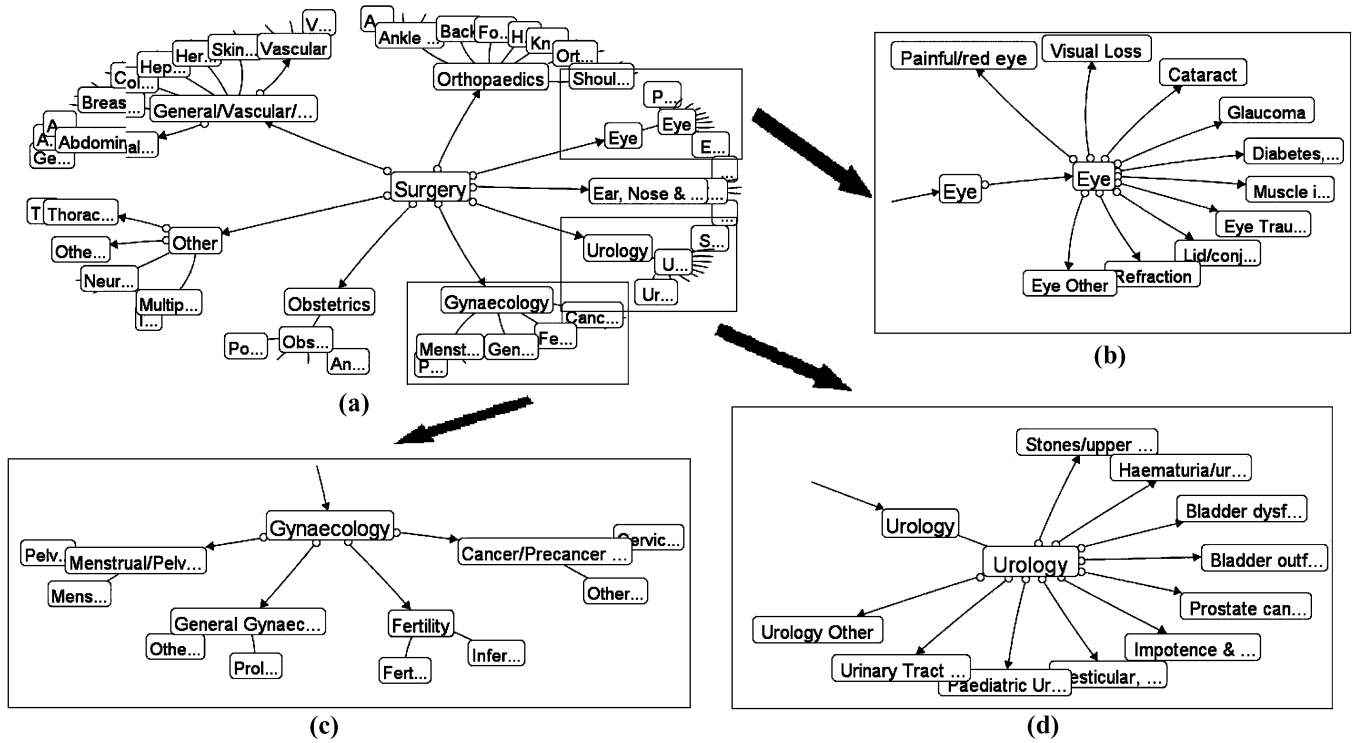


Fig. 1. Concept ontology for hierarchical video concept organization: (a) global view and (b)–(d) details for the specific regions (in box) on the concept ontology.

properties of the atomic video concepts can be characterized effectively by using the relevant salient objects and their cumulative volumetric features. Two kinds of such contextual and logical relationships are considered in our current work: a) *hypernymy/hyponymy relationship* between the concept nodes and b) *holonymy/meronymy relationship* between the atomic video concept nodes and the relevant salient objects for video content representation.

There have been a lot of efforts to construct the concept ontology to cover a large set of useful concepts [11], [12], [40]–[51]. We have incorporated WordNet with the existing ontology alignment techniques [42]–[44] to construct the concept ontology for hierarchical video concept organization, where both the joint probability and the contextual dependency of the video concepts are seamlessly integrated to formulate a new measurement for determining their associations effectively. The joint probability $\rho(C_i, C_j)$, between the text terms for interpreting the video concept C_i and C_j , is directly obtained from a corpus of annotated videos

$$\rho(C_i, C_j) = \frac{P(C_i, C_j)^2}{\sum_{h=1}^n P(C_i, C_h)^2} \quad (1)$$

where $P(C_i, C_j)$ is the co-occurrence frequency between the relevant text terms for interpreting the video concepts C_i and C_j , n is the total number of such co-occurrence appearances between C_i and all the other text terms in the basic vocabulary.

WordNet is used as the priority set to accurately quantify the contextual dependency $\pi(C_i, C_j)$ between the text terms for interpreting the video concepts C_i and C_j [11]

$$\pi(C_i, C_j) = -\log \frac{\text{length}(C_i, C_j)}{2D} \quad (2)$$

where $\text{length}(C_i, C_j)$ is the length of the shortest path between the text terms for interpreting the video concepts C_i and C_j on the WordNet, and D is the maximum depth of the WordNet.

The association between the given video concept C_i and the most relevant video concept C_h is then determined by

$$\phi(C_i, C_h) = \max \{ \rho(C_i, C_j) \pi(C_i, C_j) | j = 1, \dots, n, i \neq j \}. \quad (3)$$

Thus, the given video concept C_i is automatically linked with the most relevant video concept C_h with the highest value of the association $\phi(\cdot, \cdot)$. The multimodal conceptual properties are further determined by our hierarchical video classifier training algorithm. Our concept ontology construction results which cover 176 video concepts (medical concepts) [66] in a specific domain of surgery education videos are given in Figs. 1 and 2.

III. HIERARCHICAL APPROACH FOR BRIDGING THE SEMANTIC GAP

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 retrieval [3]–[6]. Our essential goal for video analysis is to provide more precise video content representation that allows more accurate solutions for video classification, indexing and retrieval by bridging the semantic gap. In this paper, we have developed a number of comprehensive techniques to bridge the semantic gap by: a) narrowing the video domain to a specific domain of surgery education videos, so that the contextual relationships (holonymy/meronymy relationship) between the atomic video concepts and the relevant salient objects are well defined, and thus more reliable video concept detection

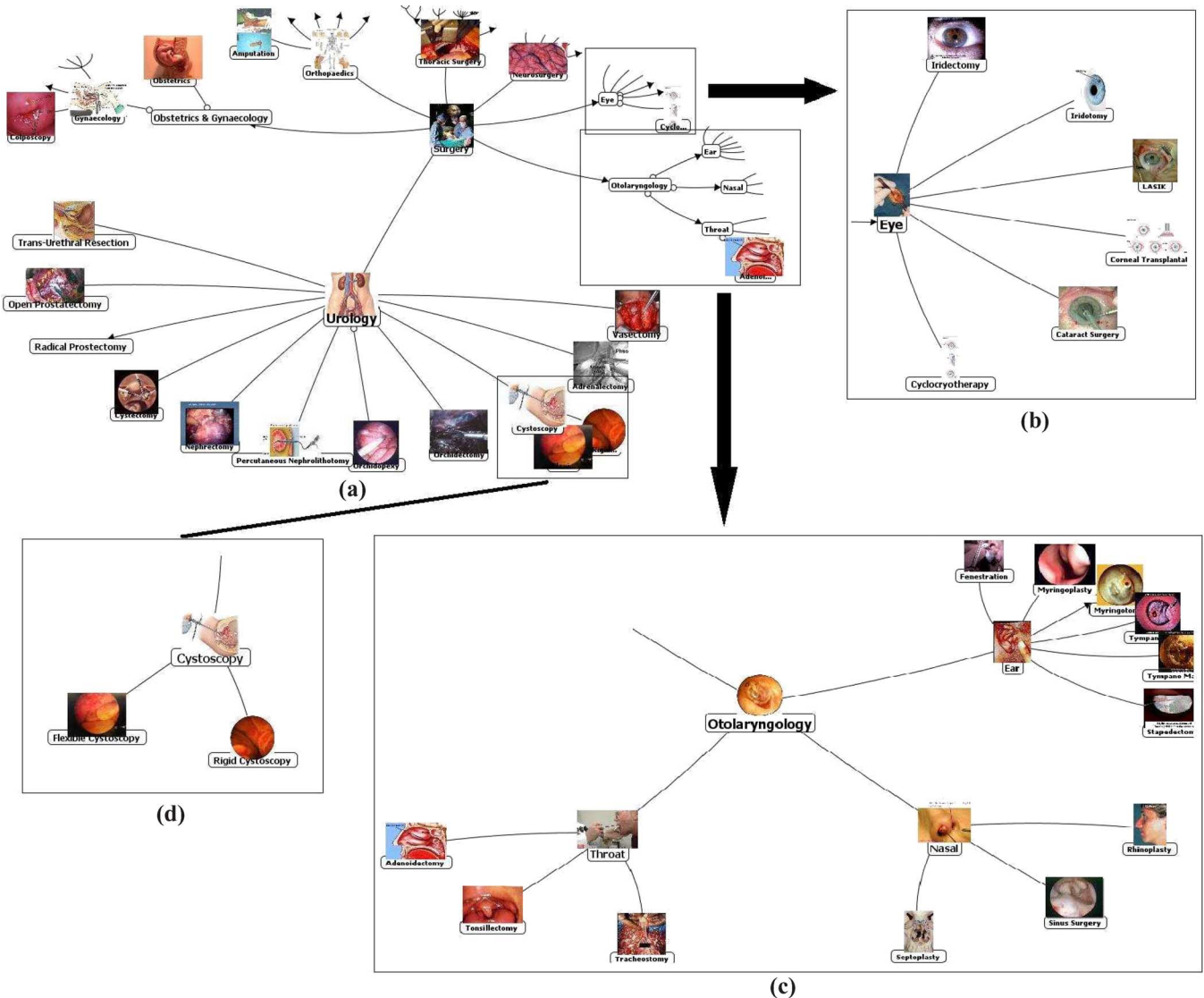


Fig. 2. Concept ontology with icons for hierarchical video concept organization: (a) global view and (b)–(d) details for the specific regions (in box) on the concept ontology.

can be achieved; b) using salient objects and their multimodal cumulative volumetric features to achieve more precise video content representation, and the salient objects are defined as the salient video components that are roughly related to the real world physical objects in a video clip [6]; c) developing new machine learning tools to incorporate concept ontology and multi-task learning for exploiting the strong correlations between the video concepts to boost hierarchical video classifier training; and d) incorporating hyperbolic visualization to bridge the semantic gap between the available video concepts and the users' real needs by visually acquainting the users with a good global view of large-scale video collections.

To enable computational interpretation of video semantics, we have developed a hierarchical scheme to bridge the semantic gap in four steps as shown in Figs. 3 and 4: 1) The semantic gap between the salient video components (i.e., real world physical objects in a video) and the low-level video signals (i.e., **Gap 1**) is bridged by using salient objects [6] and their multimodal cumulative volumetric features for video content rep-

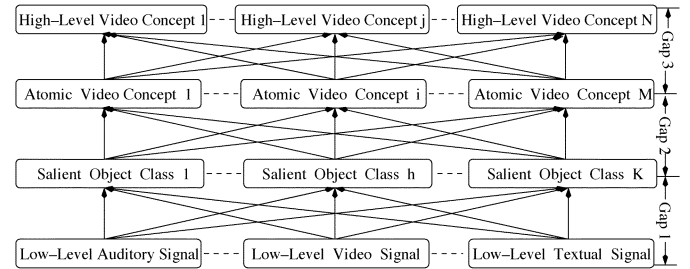


Fig. 3. Flowchart for bridging the semantic gap hierarchically.

resentation. The salient objects are defined as the salient video components that capture the most significant perceptual properties linked to the semantic meaning of the corresponding physical objects in a video clip. Thus, a salient object can describe the most significant perceptual properties of the corresponding physical video object without having to have precise segmentation. Using salient objects for video content representation can

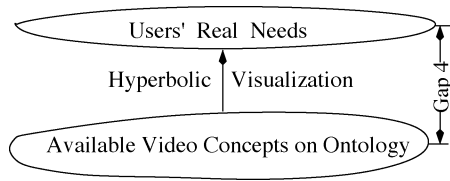


Fig. 4. Hyperbolic visualization for bridging the semantic gap between the available video concepts and the users' real needs.

provide at least four significant benefits: a) Comparison with the video shots, the salient objects can effectively characterize the most significant perceptual properties of the relevant real world physical objects in a video clip [6]. b) The salient objects are not necessarily the accurate segmentation of the real world physical objects in a video clip, thus both the computational cost and the detection error rate are reduced significantly. c) It is able to achieve a good balance between the computational complexity, the detection accuracy, and the effectiveness for interpreting the video semantics at the object level. d) Similar video clips are not necessarily similar in all their salient video components, thus partitioning the original video streams into a set of salient objects can also support partial matching and achieve more effective video classification and retrieval. 2) The semantic gap between the atomic video concepts and the salient objects (i.e., **Gap 2**) is bridged by using *multimodal boosting* to exploit the strong correlations (i.e., contextual relationships) between the appearances of the atomic video concepts and the relevant salient objects. For example, the appearance of the atomic video concept “colonic surgery” is strongly related to the appearances of the salient objects, such as “blue cloth,” “doctor gloves,” and “colonic regions.” 3) The semantic gap between the high-level video concepts and the atomic video concepts (i.e., **Gap 3**) is bridged by incorporating concept ontology and multitask learning to exploit their strong inter-concept correlations to boost hierarchical video classifier training. 4) The semantic gap between the available video concepts for semantics interpretation and the users' real needs (i.e., **Gap 4**) is bridged by using hyperbolic visualization to visually acquaint the users with a good global view of large-scale video collections.

To detect the salient objects automatically, we have designed a set of detection functions and each function is used to detect one certain type of salient objects [6]. Because one video shot may contain multiple types of 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 and achieve automatic multiclass salient object detection. As shown in Fig. 5, the white color represents the largest confidence value and the black color represents the smallest confidence value for the relevant video regions to be classified into the given 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 salient objects along the time, and the changes of the confidence maps (i.e., color changes from gray to white) can also indicate their motion properties effectively. Some experimental results on automatic salient object detection are shown in

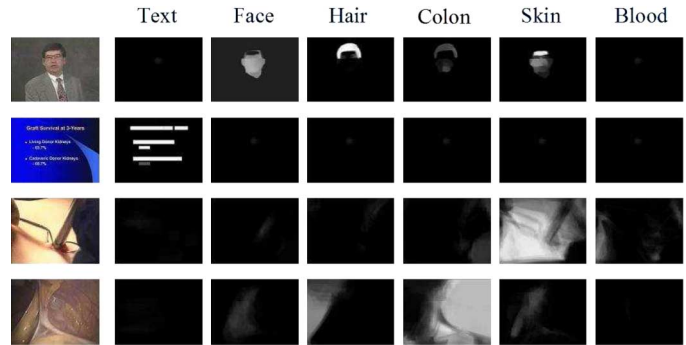


Fig. 5. Confidence maps to enable multiclass salient object detection.

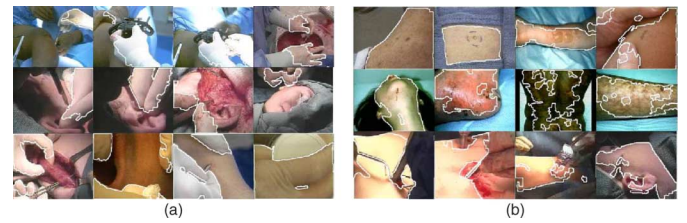


Fig. 6. Experimental results for salient object detection: (a) doctor glove and (b) human skin.

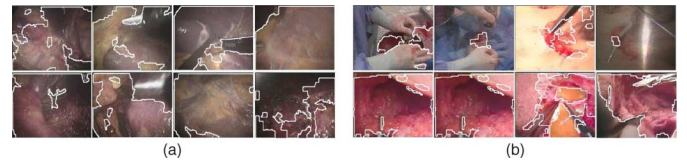


Fig. 7. Experimental results for salient object detection: (a) colonic regions and (b) blood regions.

Figs. 6 and 7. One can observe that not all the relevant video regions for the relevant physical video objects should be detected accurately. Even the salient objects may not achieve precise segmentation of the relevant physical video objects, they can still characterize the appearances of the relevant physical video objects and their most significant perceptual properties effectively.

After the salient objects are extracted, the original video streams are decomposed into a set of 3-D spatio-temporal salient objects with coherent color and motion along the time. Surgery education videos have few scene changes and they also have high spatio-temporal color and motion coherency [1], thus *cumulative volumetric features* are used to characterize such spatio-temporal color and motion coherency of the salient objects and exploit a natural 3-D spatio-temporal volume-based representation of video streams [7]–[10]. Some of our experimental results on volume-based salient object representation are given in Fig. 8. The principal properties of the salient objects can be characterized by a set of multimodal cumulative volumetric features, whose components consist of 1-D coverage ratio (i.e., density ratio) for object shape representation, 6-D object locations (i.e., 2-dimensions for object location center and 4-dimensions to indicate the rectangular box for coarse shape representation of salient object), 7-D LUV dominant colors and color variances, 14-D Tamura texture, 28-D wavelet texture features, and confidence map which is

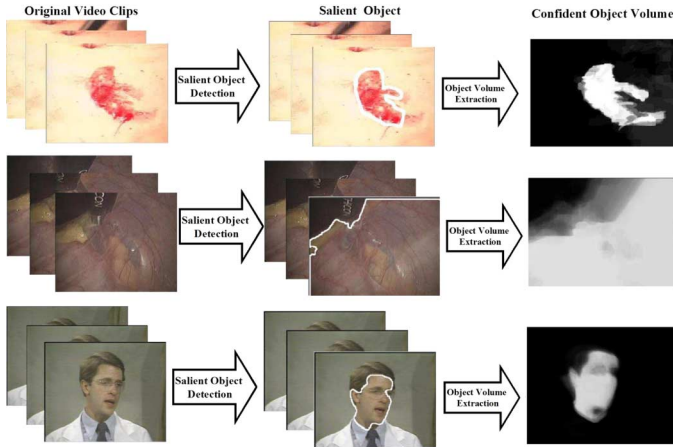


Fig. 8. Our experimental results on volume-based salient object representation.

used to quantify the posterior probability of the relevant video regions to be treated as the corresponding salient object class. The feature set to characterize the auditory salient objects includes 14-D auditory features such as loudness, frequencies, pitch, fundamental frequency, and frequency transition ratio.

Using high-dimensional cumulative volumetric features for salient object representation is able to characterize the diverse perceptual properties of the relevant video concepts more effectively. However, such high-dimensional feature space may bring two challenging problems for subsequent video classifier training and interactive video retrieval: a) learning the video classifiers in such high-dimensional heterogeneous feature space requires large amounts of training samples that increase exponentially with the feature dimensions [33]–[36] and b) high-dimensional database indexing is not available to enable fast and interactive video access. Thus, there is an urgent need to develop new approach to enable multimodal feature subset selection and dynamic classifier combination with right feature modalities.

To reduce the computational complexity for video classifier training, the high-dimensional heterogeneous (multimodal) cumulative volumetric features are automatically partitioned into 11 low-dimensional homogeneous feature subsets according to the principal properties to be characterized, so that the strongly correlated cumulative volumetric features with the same modality can be partitioned into the same homogeneous feature subset. In addition, the high-dimensional cumulative volumetric features are organized more effectively by using a two-level *feature hierarchy* (i.e., each homogeneous feature subset consists of a set of cumulative volumetric features with same modality at the first level, the high-dimensional feature space consists of 11 homogeneous feature subsets at the second level). Each homogeneous feature subset is used to characterize one certain principal property of video concept, thus the geometric property of video data is uniform and can be approximated effectively by using suitable probabilistic kernel function [37], [38].

IV. HIERARCHICAL VIDEO CLASSIFIER TRAINING

We have developed a novel bottom-up scheme by incorporating *concept ontology* and *multitask learning* to boost hierar-

chical video classifier training, and SVM classifiers are trained for automatic video concept detection [37], [38].

A. Multimodal Boosting for Atomic Video Concept Detection

To achieve more effective SVM video classifier training, we focus on addressing the following problems: a) *Kernel Function Selection*: The performance of SVM classifiers is very sensitive to the adequate selection of kernel functions [37], [38], but automatic kernel function selection heavily depends on the underlying geometric property of video data in the high-dimensional feature space. In addition, the high-dimensional feature space is heterogeneous, it is very hard to use one single type of kernel function to accurately approximate the diverse geometric properties of video data. b) *Training Sample Size Reduction and Classifier Generalization*: Learning SVM video classifiers in the high-dimensional feature space requires large amounts of training samples that exponentially increase with the feature dimensions [33]–[36]. On the other hand, learning from limited training samples could result in higher generalization error rate. Therefore, there is an urgent need to develop new classifier training algorithms that can achieve better generalization from a limited number of training samples. c) *Training Complexity Reduction*: The standard techniques for SVM classifier training have $O(m^3)$ time complexity and $O(m^2)$ space complexity, where m is the number of training samples [37], [38]. Because the number of training samples increases exponentially with the feature dimensions [33]–[36], it is too expensive to directly train reliable SVM video classifiers in the high-dimensional feature space.

To reduce the computational complexity for SVM video classifier training, we have developed a novel algorithm to incorporate *feature hierarchy* and *boosting* for video classifier training and feature selection. For a given atomic video concept C_j at the first level of the concept ontology, our multimodal boosting algorithm takes the following steps for classifier training and feature subset selection: a) To reduce the cost for SVM video classifier training, the high-dimensional cumulative volumetric features are automatically partitioned into multiple low-dimensional homogeneous feature subsets according to the perceptual properties to be characterized, where the strongly correlated cumulative volumetric features of the same modality are automatically partitioned into the same homogeneous feature subset. b) To speed up SVM video classifier training, a weak SVM classifier is learned for each homogeneous feature subset. Thus, the number of the required training samples for weak classifier training can be reduced significantly because the feature dimensions for each homogeneous feature subset are relatively low. c) Each homogeneous feature subset is used to characterize certain perceptual property of video concept, thus the underlying geometric property of video data is uniform and can accurately be approximated by using one specific type of probabilistic kernel functions [37], [38]. In addition, different types of probabilistic kernel functions can be used for different homogeneous feature subsets to approximate the diverse geometric properties of video data more accurately. d) To exploit the intra-set feature correlation, principal component analysis (PCA) is performed on each homogeneous feature subset to select the most representative feature components for each homo-

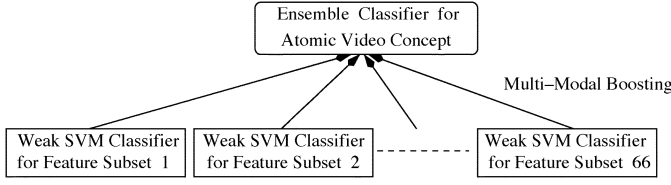


Fig. 9. Flowchart for our multimodal boosting algorithm.

geneous feature subset. e) The homogeneous feature subsets of different modalities are used to characterize different perceptual properties of video concept, where each homogeneous feature subset is responsible for certain perceptual property. Thus, the outputs of the corresponding weak SVM classifiers are diverse and may compensate each other. Based on this observation, a novel **multimodal boosting** algorithm (shown in Fig. 9) has been developed to generate an ensemble classifier by combining the weak SVM classifiers for all these homogeneous feature subsets of different modalities. The *inter-set* and *inter-modality feature correlations* among different homogeneous feature subsets of different modalities are exploited effectively by using pair-wise feature subset combinations. 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 ensemble classifier training and result in higher classification accuracy. f) Feature subset selection is achieved simultaneously by selecting the most effective weak SVM classifiers and the corresponding homogeneous feature subsets for ensemble classifier training. Thus, our feature subset selection algorithm can provide a low-dimensional feature space with better generalization, which can preserve the most discriminating information contained in the original high-dimensional feature space.

It is important to note that the process for *feature subset selection* is also a process for ensemble classifier training (i.e., dynamic weak classifier combination). For the given atomic video concept C_j , the weak classifiers for all these 11 homogeneous feature subsets and their $(11 \times 10/2) = 55$ pair-wise combinations are integrated to boost the ensemble classifier [39]

$$f_{C_j}(X) = \text{sign} \left\{ \sum_{t=1}^T \sum_{j=1}^{66} \alpha_t^j f_t^j(X) \right\}, \quad \sum_{t=1}^T \sum_{j=1}^{66} \alpha_t^j = 1 \quad (4)$$

where $f_t^j(X)$ is the weak classifier for the j th homogeneous feature subset or the pair-wise combination S_j at the t th boosting iteration, and $T = 50$ is the total number of boosting iterations. For each homogeneous feature subset, each boosting iteration learns a weak classifier from the reweighted version of the training samples. The weak classifiers and the corresponding homogeneous feature subsets which have large values of α_t^j play more important role on final prediction. Our multimodal boosting algorithm has employed a “divide and conquer” strategy with different “experts” being used to characterize the diverse perceptual properties under different feature subsets, thus higher prediction accuracy can be obtained.

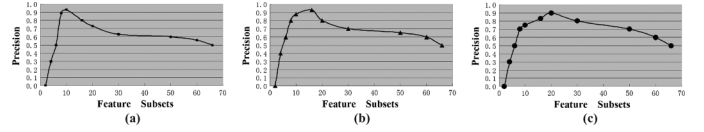


Fig. 10. Relationship between the classification accuracy (precision) of the ensemble video classifier and the number of homogeneous feature subsets and their pair-wise combinations: (a) eye trauma surgery, (b) knee injury surgery, and (c) ankle/foot surgery.

The importance factor α_t^j is updated as

$$\alpha_t^j = \frac{1}{2} \log \frac{1 - \epsilon^t(S_j)}{\epsilon^t(S_j)} \quad (5)$$

where $\epsilon^t(S_j)$ is the error rate for the weak classifier of the j th homogeneous feature subset or the pair-wise combination S_j at the t th boosting iteration. Thus, the importance factor α_t^j is updated according to the error rate of the relevant weak classifier in the current iteration. The error rate is updated as [39]

$$\epsilon^{t+1}(S_j) = \frac{1}{Z_t} \epsilon^t(S_j) e^{-\alpha_t^j Y_n f_t^j(X_n)} \quad (6)$$

where $Z_t = 2\sqrt{\epsilon^t(S_j)(1 - \epsilon^t(S_j))}$ is a normalization factor. The importance factor α_t^j decreases with the error rate $\epsilon^t(S_j)$, and thus more effective weak classifiers have more influence on the final prediction.

By selecting the most effective weak classifiers to boost an ensemble video classifier, our multimodal boosting algorithm for ensemble classifier training has jointly provided a novel approach for automatic selection of more suitable kernel functions and feature subsets. While most existing classifier training methods suffer from the problem of *curse of dimensionality*, our multimodal boosting algorithm can take advantage of high feature dimensionality. Thus, our multimodal boosting algorithm is scalable to the feature dimensions effectively.

To illustrate the evidence of the correction of our idea for feature subset selection, the optimal number of homogeneous feature subsets for the atomic video concepts “eye trauma surgery,” “knee injury surgery,” and “ankle/foot surgery” are given in Fig. 10. For the atomic video concept “eye trauma surgery,” one can conclude that only the top three homogeneous feature subsets and their pair-wise combinations may boost the classifier’s performance significantly, thus only these top three homogeneous feature subsets and their pair-wise combinations and the corresponding weak classifiers are selected to generate the ensemble classifier. From Fig. 10, one can also observe that the existence of redundant perceptual features can overwhelm the most discriminating perceptual features and lead the classifiers to a wrong way and result in lower classification accuracy, thus supporting multimodal feature selection can achieve more reliable video classifier training but also reduce the training cost significantly.

B. Hierarchical Boosting for High-Level Video Concept Detection

Because of the inherent complexity of the task (i.e., the difficulty of using the low-level perceptual features to interpret the high-level video concepts), automatic detection of the high-level video concepts with larger within-concept variations

is still beyond the ability of the state-of-the-art techniques. Thus, there is an urgent need to develop new scheme for hierarchical video classifier training by addressing the following issues simultaneously:

- a) *Computational Complexity Reduction*: The high-level video concepts cover more general semantics with larger within-concept variations, thus the hypothesis spaces for training their classifiers are also very large and result in higher computational complexity. The perceptual properties for such high-level video concepts may have huge diversities, and thus large amounts of training samples are needed to achieve reliable classifier training. Because the cost for SVM video classifier training largely depends on the size of the training samples [33]–[36], it is very expensive to directly train the classifiers for the high-level video concepts.
- b) *Inter-Level Error Transmission Avoidance*: By incorporating the concept ontology for hierarchical video classifier training, the classifiers for the high-level video concepts can be learned hierarchically by combining the classifiers of their children video concepts which have smaller within-concept variations [1], [2]. Learning the classifiers for the high-level video concepts hierarchically can reduce the training cost significantly by partitioning the hypothesis space into multiple smaller ones for their children video concepts. However, such hierarchical approach may seriously suffer from the inter-level error transmission problem [1].
- c) *Knowledge Transferability and Task Relatedness Exploitation*: The video concepts are dependent and such dependencies can be characterized effectively by the concept ontology. Thus, what is already learned for one specific video concept can be transferred to improve the classifier training for its parent video concept on the concept ontology and its sibling video concepts under the same parent node. Therefore, isolating the video concepts and learning their classifiers independently are not appropriate. Multitask learning is one promising solution to this problem [60]–[65], but the success of multitask learning largely depends on the relatedness of multiple learning tasks. One of the most important open problems for multitask learning is to better characterize what the related tasks are [59].

Based on these observations, we have developed a novel **hierarchical boosting** algorithm that is able to combine the classifiers trained under different tasks to boost an ensemble classifier for a new task. First, the concept ontology is used to identify the related tasks, e.g., training the classifiers for the sibling video concepts under the same parent node. Second, such task relatedness is further used to determine the transferable knowledge and common features among the classifiers for the sibling video concepts to generalize their classifiers significantly from fewer training samples. Because the classifiers for the sibling video concepts under the same parent node are used to characterize both their *individual perceptual properties* and the *common perceptual properties* for their parent node, they can compensate each other and their outputs are strongly correlated according

to the new task (i.e., learning a biased classifier for their parent node).

For a given second-level video concept C_k , its children video concepts (i.e., the sibling atomic video concepts under C_k) are strongly correlated and share some common perceptual properties for their parent node, thus *multitask learning* can be used to train their classifiers simultaneously. Because the related tasks are characterized accurately by the underlying concept ontology, our hierarchical classifier training algorithm can provide a good environment to enable more effective multitask learning.

To integrate multitask learning for SVM video classifier training, a *common regularization term* W_0 is used to represent and quantify the transferable knowledge and common features among the SVM video classifiers for the sibling video concepts under the same parent node. The SVM classifier for the atomic video concept C_j can be defined as [37], [38]

$$f_{C_j}(X) = W_j^T X + b \quad (7)$$

where $W_j = W_0 + V_j$, W_0 is the common regularization term shared between the sibling atomic video concepts under the same parent node C_k , and V_j is the specific regularization term for the atomic video concept C_j . Therefore, the regularization terms $\{W_j | j = 1, \dots, L\}$ of these L SVM video classifiers are able to simultaneously characterize both the commonness and the individuality among multiple learning tasks (i.e., training the classifiers for L sibling atomic video concepts).

Given the labeled training samples for these L sibling atomic video concepts under the same parent node $C_k : \Omega = \{X_{ij}, Y_{ij} | i = 1, \dots, N; j = 1, \dots, L\}$, the margin maximization procedure is to search a hypothesis space that is appropriate for all these L atomic video concepts [65]

$$\min \left\{ \sum_{j=1}^L \sum_{i=1}^N \xi_{ij} + \frac{\beta_1}{L} \sum_{j=1}^L \|V_j\|^2 + \beta_2 L \|W_0\|^2 \right\} \quad (8)$$

subject to:

$$\forall_{i=1}^N \text{ and } \forall_{j=1}^L : Y_{ij}(W_0 + V_j) \cdot X_{ij} + b \geq 1 - \xi_{ij}, \quad \xi_{ij} \geq 0$$

where $\xi_{ij} \geq 0$ represents the training error rate, $L > 1$ is the total number of atomic video concepts under the same parent node C_k , β_1 and β_2 are positive regularization parameters. The dual optimization problem for (8) is to determine the optimal α_{ij}^* by maximizing

$$\max \left\{ \sum_{j=1}^L \sum_{i=1}^N \alpha_{ij} - \frac{1}{2} \sum_{j=1}^L \sum_{i=1}^N \sum_{h=1}^L \sum_{l=1}^N \alpha_{ih} Y_{ih} \alpha_{jl} Y_{jl} K_{jh}(X_{ih}, X_{jl}) \right\} \quad (9)$$

subject to:

$$\forall_{i=1}^N \text{ and } \forall_{j=1}^L : 0 \leq \alpha_{ij} \leq C, \quad \sum_{j=1}^L \sum_{i=1}^N \alpha_{ij} Y_{ij} = 0$$

where $K_{jh}(\cdot, \cdot)$ is the underlying kernel function. If α_{ij}^* is a solution of (9), the optimal SVM classifier for the j th atomic video concept C_j is given by

$$f_{C_j}(X) = \sum_{j=1}^L \sum_{i=1}^N \alpha_{ij}^* K_{ij}(X_{ij}, X). \quad (10)$$

Because the common regularization term W_0 is used to represent the transferable knowledge and common features between the sibling atomic video concepts, it can further be treated as a prior regularization term to bias the training of the SVM classifier for their parent node C_k . Setting such prior regularization term is able to generalize the classifier from fewer training samples and reduce the training cost significantly. Based on such prior regularization term, a *biased classifier* for their parent node C_k is trained effectively from a restricted class of hypotheses by using few new training samples. Thus, the biased classifier for their parent node C_k is determined by minimizing

$$\min \left\{ \frac{1}{2} \|W - W_0\|^2 + C \sum_{l=1}^m [1 - Y_l(W^T \cdot X_l + b)] \right\} \quad (11)$$

where W_0 is the common regularization term shared between the sibling atomic video concepts under C_k . (X_l, Y_l) , $l = 1, \dots, m$ are the new training samples for learning the biased classifier for C_k . The dual problem for (11) is solved by minimizing

$$\min \left\{ \frac{1}{2} \sum_{l=1}^m \sum_{h=1}^m \alpha_l \alpha_h Y_l Y_h X_l^T X_h - \sum_{l=1}^m \alpha_l (1 - Y_l W_0^T X_l) \right\} \quad (12)$$

subject to:

$$\forall_{l=1}^m : 0 \leq \alpha_l \leq C, \quad \sum_{l=1}^m \alpha_l Y_l = 0.$$

The optimal solution of (12) satisfies

$$W = W_0 + \sum_{l=1}^m \alpha_l Y_l X_l. \quad (13)$$

By using the prior regularization term W_0 , the hypothesis space of the biased classifier is chosen automatically to be large enough to contain an optimal solution of the new task and yet be small enough to ensure reliable generalization from reasonably-sized training sample set. Knowing the right bias term W_0 also makes the problem for training the biased classifier much easier. The bias classifier for the given second-level video concept C_k is defined as

$$f_{C_k}(X) = W^T X + b, \quad W = W_0 + \sum_{l=1}^m \alpha_l Y_l X_l. \quad (14)$$

To learn the ensemble classifier for the given second-level video concept C_k , its biased classifier should be combined with the classifiers for its children video concepts. Unfortunately, all the existing boosting techniques can only combine the weak classifiers that are learned in different ways (i.e., different input spaces) but for the same task [39], and they did not include the

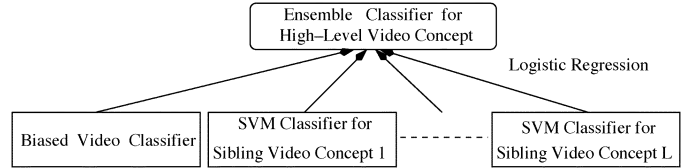


Fig. 11. Flowchart for our hierarchical boosting algorithm.

regularization between different tasks which is very essential for hierarchical video classifier training.

We have developed a novel *hierarchical boosting* algorithm for multitask classifier combination, which is able to integrate the classifiers trained for multiple tasks and leverage their distinct strengths and exploit the strong correlations of their outputs according to the new task. Our hierarchical boosting algorithm can search an optimal combination of these multitask classifiers by sharing their transferable knowledge and common features according to the new task (i.e., learning the ensemble classifier for their parent node C_k), and thus it is able to generalize the ensemble classifier significantly while reducing the computational complexity dramatically. For the given second-level video concept C_k , the final prediction of its ensemble classifier can be obtained by a *logistic regression* of the predictions of its biased classifier and the classifiers for its children video concepts (shown in Fig. 11) [58]. Thus, the ensemble classifier for the second-level video concept C_k is defined as

$$H_{C_k}(X) = \sum_{j=1}^{L+1} p_j(C_j|X) f_{C_j}(X) \quad (15)$$

where $p_j(C_j|X)$ is the posterior distribution of the j th classifier $f_{C_j}(X)$ to be combined, and $p_j(C_j|X)$ is automatically determined by

$$p_j(C_j|X) = \frac{\exp(f_{C_j}(X))}{\sum_{j=1}^{L+1} \exp(f_{C_j}(X))}. \quad (16)$$

After the classifiers for the sibling second-level video concepts are available, they can further be integrated to boost the classifier for their parent node at the third level of the concept ontology. Through such hierarchical approach, the classifiers for the video concepts at the higher levels of the concept ontology can be obtained automatically.

V. HIERARCHICAL CLASSIFICATION FOR AUTOMATIC MULTILEVEL VIDEO ANNOTATION

After our hierarchical video classifiers are available, a top-down approach is used to classify the video shots into the most relevant video concepts at different semantic levels. Our hierarchical video classification scheme takes the following steps for automatic video concept detection and annotation. a) The video shots are automatically detected from the test video clips. b) For each video shot, the underlying salient objects are automatically detected and tracked along the time for extracting their cumulative volumetric features. c) The video shots are classified into the most relevant video concepts hierarchically according to their perceptual properties which are characterized by the cumulative volumetric features extracted from the relevant salient objects. d) The keywords for interpreting the relevant video concepts are

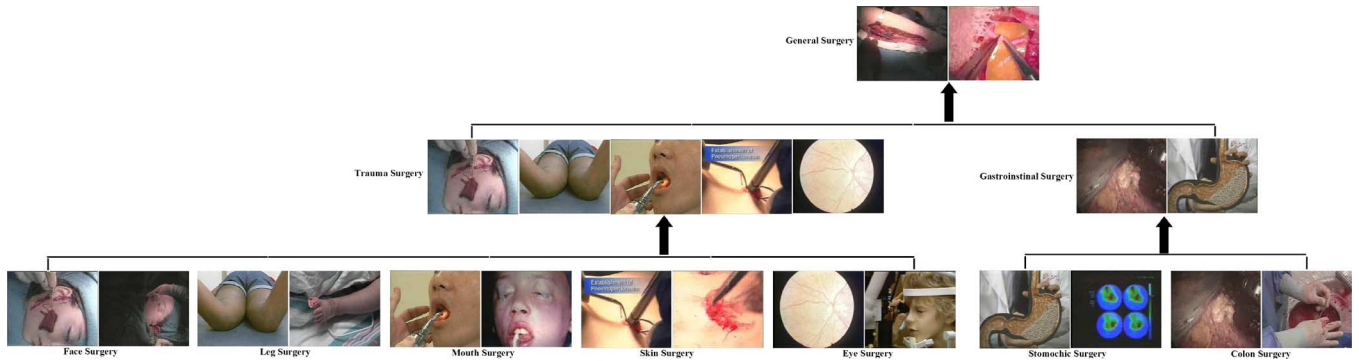


Fig. 12. Hierarchical organization of semantic video classification results.

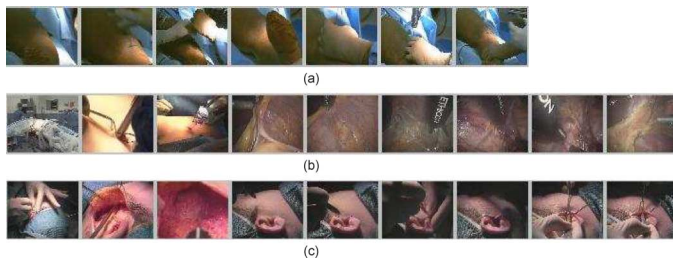


Fig. 13. Second-level video concept “otolaryngology surgery” may consist of multiple atomic video concepts such as “ankle/foot injury surgery,” “colonic surgery,” and “face trauma surgery.”

automatically assigned to the test video shots to achieve multilevel video annotation. Some video classification results are given in Figs. 12 and 13.

In our hierarchical video classification algorithm, the initial classification of a test video clip at the higher levels of the concept ontology is critical, because the classifiers at the subsequent levels cannot recover from the misclassification of the test video clip that may occur at a higher concept level. In addition, this misclassification can be propagated to the terminal node (i.e., inter-level error transmission). To address such inter-level error transmission problem, we have integrated two innovative solutions: 1) enhancing the classifiers for the video concepts at the higher levels of the concept ontology, so that they can have higher discrimination power and 2) integrating a novel A^* search algorithm that is able to detect such misclassification path early and take appropriate actions for automatic error recovery [70].

Three significant respects of our hierarchical video classification algorithm are able to address the inter-level error transmission problem effectively. a) The transferable knowledge and common features can be shared among the classifiers for the video concepts at the same semantic level of the concept ontology to maximize their margins and enhance their discrimination power significantly. Therefore, the test video shots can be classified more accurately at the beginning, i.e., video concepts at the higher levels of the concept ontology. By exploiting the strong correlations between the outputs of the classifiers for the children video concepts, our hierarchical boosting algorithm is able to learn more accurate ensemble classifiers for the high-level video concepts. b) The classification decision for the

test video shot is determined by a voting from multiple multi-task classifiers at the same semantic level to make their errors to be transparent. c) An *overall probability* is calculated to determine the best path for hierarchical video classification. For a given test video shot, an optimal classification path should provide maximum value of the overall probability among all the possible classification paths. The overall probability $h(C_k)$ for one certain classification path (from higher level concept node to the relevant lower level concept nodes) is defined as

$$\begin{aligned} h(C_k) &= p(C_k) + g(C_j) \\ g(C_j) &= \max \{p(C_i) | i = 1, \dots, L\} \end{aligned} \quad (17)$$

where $p(C_k)$ is the posterior probability for the given test video shot to be classified into the current video concept C_k at the higher level of the concept ontology, $p(C_i)$ is the posterior probability for the given test video shot to be classified into the children video concept C_i of C_k , $g(C_j)$ is the maximum posterior probability for the given test video shot to be classified into the most relevant children concept node C_j . Thus, a good path should achieve higher classification accuracy for both the current high-level concept node and the relevant children concept node. By using the overall probability, it is able for us to detect the incorrect classification path early and take appropriate actions for automatic error recovery.

It is important to note that once a test video clip is classified, the keywords for interpreting the underlying salient object classes and the relevant video concepts at different levels of the concept ontology become the keywords for interpreting its multilevel semantics. Our scheme for *automatic multilevel video annotation* is very attractive to enable more effective video retrieval with more sufficient and precise keywords. Thus, the naive users can have more flexibility to specify their queries via sufficient keywords at different semantic levels.

VI. ONTOLOGY VISUALIZATION FOR VIDEO NAVIGATION AND INTUITIVE QUERY SPECIFICATION

For naive users to harvest the research achievements of CBVR community, it is very important to develop more comprehensive framework for intuitive query specification and evaluation, but it is also a problem without a good solution so far. The problem, in essence, is also about how to present

a good global view of large-scale video collections to users [73]–[76], so that users can easily specify their queries for video retrieval. Therefore, there is a great need to generate the overall information of large-scale video collections conceptually and incorporate the concept ontology to organize and visualize such concept-oriented overall information more effectively and intuitively.

To achieve multimodal representation of concept ontology, each concept node on the concept ontology is jointly characterized by: *keyword* to interpret its semantics, *most representative video shots* to display its concept-oriented summary, *decision principles* (i.e., support vectors and importance factors for classifier combination) to characterize its feature-based principal properties, and contextual relationships between the relevant video concepts.

We have developed a novel scheme to generate the concept-oriented summarization of large-scale video collection. For one given video concept on the concept ontology, three types of video shots are automatically selected to generate its concept-oriented summary (i.e., most representative video shots). a) Video shots which locate on the decision boundaries of the SVM video classifier; b) Video shots which have higher confidence scores in the classification procedure; and c) video shots which locate at the centers of dense areas and can be used to represent large amounts of semantically similar video shots for the given video concept. To obtain such representative video shots, multidimensional scaling is used to cluster the semantically similar video shots for the given video concept into multiple significant groups [74], [75].

Our multimodal approach for concept ontology representation can provide not only a basic vocabulary of keywords for users to specify their queries precisely and objectively, but also the concept-oriented video summaries to enable hierarchical video navigation and the feature-based decision principles to support query-by-example effectively. Therefore, query interfaces that contain a graphical representation of the concept ontology can leverage the benefits of the concept ontology and make the task of query specification much easier and more intuitive. By navigating the concept ontology interactively, users can specify their queries easily with a better global view of large-scale video collections.

Visualizing large-scale concept ontology in two-dimensional system interface is not a trivial task [77], [78]. We have developed multiple innovative solutions to tackle this issue effectively. a) A tree-based approach is incorporated to visualize our concept ontology in a nested graph view, where each video concept node is displayed along with its parent node, its children nodes, and its multimodal representation parameters (i.e., keyword, concept-oriented summary, feature-based decision principles). b) The geometric closeness of the concept nodes on the visualization tree is related to the semantic relatedness of the video concepts, so that our graphical presentation can reveal a great deal about how these video concepts are organized and how they are intended to be used. c) Both geometric zooming and semantic zooming are integrated to adjust the level of visible detail automatically according to the discerning constraint on the number of concept nodes that can be displayed per visualization view.

Our approach for concept ontology visualization exploits hyperbolic geometry [77], [78]. The hyperbolic geometry is particularly well suited to graph-based layout of large-scale concept ontology, and it has “more space” than Euclidean geometry. The essence of our approach is to project the concept ontology onto a hyperbolic plane according to the contextual relationships between the video concepts, and layout the concept ontology by mapping the relevant concept nodes onto a circular display region. Thus, our concept ontology visualization framework takes the following steps. a) The video concept nodes on the concept ontology are projected to a hyperbolic plane according to their contextual relationships, and such projection can usually preserve the original contextual relationships between the video concept nodes. b) After we obtain such context-preserving projection of the video concept nodes, we can then use Poincaré disk model [77], [78] to map the concept nodes on the hyperbolic plane to a 2-D display coordinate. Poincaré disk model maps the entire hyperbolic space onto an open unit circle, and produces a nonuniform mapping of the video concept nodes to the 2-D display coordinate. The Poincaré disk model preserves the angles, but distorts the lines. The Poincaré disk model also compresses the display space slightly less at the edges, which in some cases can have the advantage of allowing a better view of the context around the center of projection.

Our implementation relies on the representation of the hyperbolic plane, rigid transformations of the hyperbolic plane and mappings of the concept nodes from the hyperbolic plane to the unit disk. Internally, each concept node on the graph is assigned a location $z = (x, y)$ within the unit disk, which represents the Poincaré coordinates of the corresponding video concept node. By treating the location of the video concept node as a complex number, we can define such a mapping as the linear fractional transformation [77], [78]

$$z_t = \frac{\theta z + P}{1 + \bar{P}\theta z} \quad (18)$$

where P and θ are the complex numbers, $|P| < 1$ and $|\theta| = 1$, and \bar{P} is the complex conjugate of P . This transformation indicates a rotation by θ around the origin following by moving the origin to P (and $-P$ to the origin).

After the hyperbolic visualization of the concept ontology is available, it can be used to enable interactive exploration and navigation of large-scale video collections at the concept level via *change of focus*. The *change of focus* is implemented by changing the mapping of the video concept nodes from the hyperbolic plane to the unit disk for display, and the positions of the video concept nodes in the hyperbolic plane need not to be altered during focus manipulation. Users can change their focus of video concepts by clicking on any visible video concept node to bring it into focus at the center, or by dragging any visible video concept node interactively to any other location without losing the contextual relationships between the video concept nodes, where the rest of the layout of the concept ontology transforms appropriately. Thus, our hyperbolic framework for concept ontology visualization has demonstrated the remarkable capabilities for interactively exploring large-scale video collections. By supporting change of focus, our hyperbolic visualization frame-

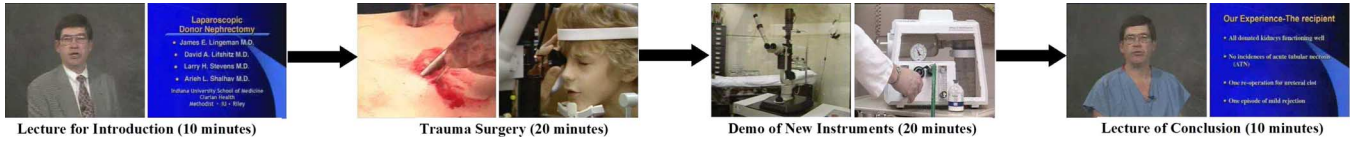


Fig. 16. Temporal context network for one surgery education video clip.

VII. ALGORITHM AND SYSTEM EVALUATION

We carry out our experimental studies of our proposed algorithms by using our large-scale collections of surgery education videos. We currently have collected more than 300 h MPEG surgery education videos, where 175 h MPEG videos are currently used for algorithm and system evaluation. 125 h MPEG videos are used for classifier training for 176 video concepts [66] and 50 h MPEG videos are used as the test samples for algorithm evaluation.

Our works on algorithm evaluation focus on comparing our proposed algorithms with other existing techniques for the same video classification task. a) By using the same sets of object-based cumulative volumetric features and training samples for video classifier training, we have compared the performance differences between our proposed multimodal boosting algorithm, Adaboost [39], linear SVM classifier combination [37], [38]. b) By using salient objects for feature extraction, we have compared the performance differences between two approaches for video classifier training by using the same sets of cumulative volumetric features and training samples: our hierarchical boosting algorithm *versus* the flat approach (i.e., the classifier for each video concept is learned independently). c) By using the same sets of cumulative volumetric features and training samples for video classifier training, we have also compared the performance differences between three approaches: our hierarchical boosting algorithm, multiclass boosting [57], [58], multitask boosting techniques [59].

The *benchmark metric* for classifier evaluation includes *precision* ρ and *recall* q . They are defined as

$$\rho = \frac{\vartheta}{\vartheta + \xi}, \quad q = \frac{\vartheta}{\vartheta + \nu} \quad (19)$$

where ϑ is the set of true positive samples that are related to the corresponding video concept and are classified correctly, ξ is the set of true negative samples that are irrelevant to the corresponding video concept and are classified incorrectly, and ν is the set of false positive samples that are related to the corresponding video concept but are misclassified.

To assess the statistical difference between multiple algorithms for video classifier training, we have also performed *t-test* and *t-value* is used to compare their mean precision and recall. The *t-value* is defined as

$$T_p = \frac{|\mu_T^p - \mu_C^p|}{\sqrt{(\sigma_T^p + \sigma_C^p)/n}}, \quad T_r = \frac{|\mu_T^r - \mu_C^r|}{\sqrt{(\sigma_T^r + \sigma_C^r)/n}} \quad (20)$$

where n is the total number of video concepts used in the test, T_p and T_r indicate the *t-values* for the precision and recall, μ_T^p and σ_T^p are the mean and variance of the precision for the target algorithm, μ_C^p and σ_C^p are the mean and variance of the precision for the comparison algorithm, μ_T^r , σ_T^r , μ_C^r , σ_C^r are the mean and variance of the recall for the target and comparison

algorithms. The larger *t-value* indicates that the target algorithm significantly outperforms the comparison algorithm.

A. Hierarchical Boosting versus Flat Approach

By extracting the cumulative volumetric features for video classifier training, we have compared the performance differences between two approaches by using same sets of training samples and cumulative volumetric features: flat approach (i.e., the classifier for each video concept is learned independently) *versus* our hierarchical boosting scheme. Table I gives the precision and recall of the classifiers for some atomic video concepts and high-level video concepts, and the statistical difference between our hierarchical boosting algorithm and the flat approach is characterized effectively by the *t-values*: $T_p = 11.8$ and $T_r = 8.9$.

From these experimental results, one can observe that our hierarchical video classifier training scheme can improve the detection accuracy for the high-level video concepts significantly. Such significant improvement on the detection accuracy benefits from three components. a) The classifiers for the high-level video concepts with larger within-concept variations are trained hierarchically by combining the classifiers for the relevant low-level video concepts with smaller within-concept variations, and thus the hypothesis space for classifier training is reduced significantly which can generalize the classifiers effectively from fewer training samples. b) For a given high-level video concept, the strong correlations between its children video concepts are exploited effectively by sharing their transferable knowledge and common features. Thus, our hierarchical boosting algorithm can learn not only the reliable video classifiers but also the bias, i.e., learn how to generalize. c) The final prediction results for the classifiers of the high-level video concepts are obtained by a voting procedure according to the classifiers at the same semantic level to make their prediction errors to be transparent, and thus the inter-level error transmission problem can be addressed effectively.

For the atomic video concepts at the first level of the concept ontology, our proposed hierarchical classifier training scheme can also obtain higher detection accuracy, because the strong correlations between the sibling atomic video concepts under the same parent node are exploited by sharing their transferable knowledge and common features via multitask learning. In addition, our hierarchical video classifier training scheme has provided a good environment to enable more effective multitask learning, i.e., training the classifiers simultaneously for the strongly correlated and sibling video concepts under the same parent node. Through multitask learning, the risk of overfitting the shared part is significantly reduced and the problem of inter-concept similarity can be addressed more effectively, which can result in higher classification accuracy.

TABLE I
PERFORMANCE DIFFERENCES BY USING OUR HIERARCHICAL BOOSTING ALGORITHM AND THE FLAT APPROACH TO TRAIN THE CLASSIFIERS FOR THE SAME VIDEO CONCEPTS BY USING SAME SETS OF TRAINING SAMPLES AND CUMULATIVE VOLUMETRIC FEATURES

atomic video concepts	red eye	visual loss	diabetes	muscle imbalance	urinary infection	postnatal
hierarchical boosting	90.2% /92.8%	90.4% /90.6%	89.5% /89.3%	92.3% /93.8%	91.3% /92.6%	92.7% /92.9%
flat approach	81.3% /82.6%	79.3% /92.2%	71.8% /72.4%	81.6% /82.7%	80.2% /82.1%	80.5% /81.3%
atomic video concepts	eye trauma	refraction	shoulder injury	shoulder arthritis	menstrual disorders	fertility fail
hierarchical boosting	92.8% /93.6%	91.2% /91.8%	93.2% /91.6%	93.5% /93.3%	88.8% /89.3%	87.8% /90.2%
flat approach	80.5% /81.2%	80.3% /82.1%	80.6% /82.3%	81.3% /81.6%	78.2% /79.3%	77.5% /78.3%
atomic video concepts	knee arthritis	knee injury	forearm/hand	forearm fracture	head injury	pelvic pain
hierarchical boosting	93.7% /92.9%	93.5% /91.8%	94.2% /92.6%	93.2% /93.5%	93.9% /94.2%	93.6% /93.7%
flat approach	80.2% /80.4%	78.4% /82.1%	76.9% /78.2%	80.3% /81.5%	81.6% /82.1%	81.8% /82.4%
atomic video concepts	ankle/foot	ankle fracture	arteria	vericose veins	anaemia pain	neurosurgery
hierarchical boosting	93.5% /94.2%	92.8% /91.2%	90.2% /89.6%	91.6% /92.1%	89.7% /90.2%	89.3% /90.2%
flat approach	80.2% /81.3%	80.5% /82.3%	81.2% /82.1%	79.3% /80.5%	78.3% /77.9%	76.8% /77.4%
atomic video concepts	lesions	ventral hernia	inguinal hernia	hepatobiliary	breast lump	Anaemia & lypmhold
hierarchical boosting	92.6% /92.3%	91.6% /92.3%	92.7% /91.8%	93.6% /94.3%	91.2% /92.2%	93.5% /93.2%
flat approach	80.6% /81.5%	80.2% /82.1%	81.5% /82.0%	79.8% /80.3%	80.2% /82.1%	80.5% /81.3%
atomic video concepts	upper-GI	perianal	colonic	colon neoposia	obesity	endocrinic
hierarchical boosting	92.3% /92.6%	92.3% /91.8%	90.6% /91.6%	92.8% /93.2%	92.9% /91.6%	92.1% /93.3%
flat approach	80.6% /81.2%	80.6% /81.6%	81.2% /82.2%	80.2% /81.5%	80.2% /81.3%	78.5% /79.1%
second-level concepts	breast-endocrine	abdominal	thoracic	multiple injury & NOS	vascular	colorectal
hierarchical boosting	94.3% /94.8%	93.6% /93.8%	95.2% /94.5%	93.7% /93.8%	94.2% /94.8%	93.6% /93.4%
flat approach	71.2% /72.3%	78.3% /75.8%	71.3% /72.3%	73.8% /75.6%	78.5% /77.9%	80.1% /81.6%
second-level concepts	eye	gynaecology	urology	menstrual/pelvic pain	fertility	hip arthritis
hierarchical boosting	93.6% /94.3%	93.6% /94.2%	93.4% /94.7%	94.8% /95.6%	93.8% /93.9%	95.2% /95.6%
flat approach	74.2% /73.8%	75.4% /76.2%	74.3% /75.3%	76.5% /76.3%	78.2% /79.1%	77.4% /78.6%

B. Comparing Multiple Boosting Approaches

By using the same sets of training samples and cumulative volumetric features for video classifier training, we have also compared the performance differences between our hierarchical boosting algorithm, multiclass boosting [57], [58] and multitask boosting [59]. The multiclass boosting techniques learn multiple classifiers by optimizing a joint objective function [57], [58]. The multitask boosting algorithm has recently been proposed to enable multiclass object detection by sharing the common features among the classifiers [59]. Rather than incorporating the transferable knowledge and common features to learn a biased classifier, the ensemble classifier for each object class is simply composed by the classifiers that are trained for all the pair-wise object class combinations [59]. For video classification applications, pair-wise concept combinations are used to exploit the *transferable knowledge and common features* between the sibling atomic video concepts.

As shown in Fig. 17, there are $(L(L-1)/2) + 1$ inter-concept combinations for L atomic video concepts under the same parent node C_k (i.e., $(L(L-1)/2)$ for all possible pair-wise combinations of the atomic video concepts and 1 for combining all these L atomic video concepts). The relevant training samples are integrated to learn $(L(L-1)/2) + 1$ combined classifiers, and these combined classifiers are used to characterize the common principal properties for these L sibling atomic video concepts under the same parent node C_k . Thus, all these $\bar{L} = (L(L-1)/2) + 1 + L = (L(L+1)/2) + 1$ classifiers are integrated to generate the ensemble classifier $H_{C_k}(X)$ for the second-level video concept C_k . The optimal ensemble classifier for the given second-level video concept C_k is determined by

$$H_{C_k}(X) = \text{sign} \left\{ \sum_{j=1}^{\bar{L}} \beta_j H_j(X) \right\}, \quad \sum_{j=1}^{\bar{L}} \beta_j = 1 \quad (21)$$

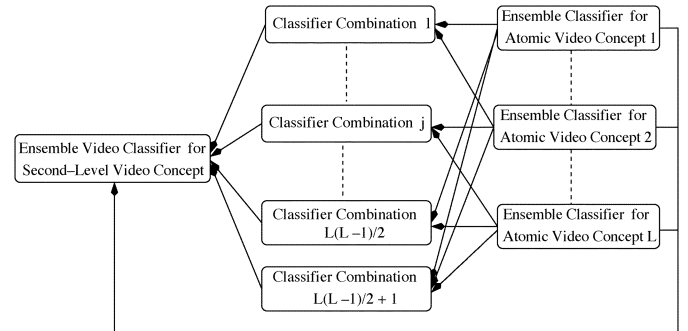


Fig. 17. Flowchart for multitask boosting via pair-wise concept combination.

where $H_j(X)$ is the j th classifier in the basic vocabulary for \bar{L} classifier combinations, \bar{L} is the total number of potential classifiers in the basic vocabulary, and β_j is the relative importance factor for the j th classifier. For the given second-level video concept C_k , the final prediction of its ensemble classifier is obtained by voting the predictions of all these \bar{L} pair-wise classifiers.

As shown in Table II, our hierarchical boosting algorithm can significantly outperform both the multiclass boosting and the multitask boosting techniques [57]–[59]. The statistical difference between our hierarchical boosting algorithm and multiclass boosting algorithm is characterized effectively by the t-values: $T_p = 12.6$ and $T_r = 7.98$. The t-values for comparing our hierarchical boosting algorithm with multitask boosting are: $T_p = 9.28$ and $T_r = 6.29$.

The multiclass boosting techniques do not explicitly exploit the transferable knowledge and common features among the classifiers to enhance their classification performance [57], [58]. The multitask boosting algorithm via pair-wise concept combinations may seriously suffer from the following problems. a) The decision boundaries of these pair-wise classifiers may not

TABLE II

PERFORMANCE DIFFERENCES BY USING OUR HIERARCHICAL BOOSTING ALGORITHM, MULTICLASS BOOSTING AND MULTITASK BOOSTING TO TRAIN THE CLASSIFIERS FOR THE SAME VIDEO CONCEPTS BY USING SAME SETS OF TRAINING SAMPLES AND CUMULATIVE VOLUMETRIC FEATURES

atomic video concepts	red eye	visual loss	diabetes	muscle imbalance	head injury	pelvic pain
hierarchical boosting	90.2% /92.8%	90.4% /90.6%	89.5% /89.3%	92.3% /93.8%	93.9% /94.2%	93.6% /93.7%
multi-task boosting	83.2% /81.8%	80.5% /82.3%	78.6% /76.9%	81.2% /82.4%	84.8% /85.3%	84.6% /85.6%
multi-class boosting	78.3% /77.2%	74.6% /75.6%	74.2% /73.5%	74.6% /77.3%	80.2% /82.6%	79.4% /80.2%
atomic video concepts	eye trauma	refraction	shoulder injury	shoulder arthritis	anaemia pain	neurosurgery
hierarchical boosting	92.8% /93.6%	91.2% /91.8%	93.2% /91.6%	93.5% /93.3%	89.7% /90.2%	89.3% /90.2%
multi-task boosting	81.6% /81.8%	82.4% /82.6%	81.4% /83.5%	84.6% /85.2%	81.6% /80.9%	80.4% /81.3%
multi-class boosting	76.3% /78.2%	73.5% /76.2%	75.6% /74.8%	77.7% /75.8%	78.3% /78.5%	76.7% /78.6%
atomic video concepts	knee arthritis	knee injury	forearm/hand	forearm fracture	breast lump	Anaemia & lymphoid
hierarchical boosting	93.7% /92.9%	93.5% /91.8%	94.2% /92.6%	93.2% /93.5%	91.2% /92.2%	93.5% /93.2%
multi-task boosting	82.6% /83.3%	84.7% /85.2%	86.3% /86.5%	82.5% /83.1%	82.6% /82.3%	83.9% /84.2%
multi-class boosting	76.9% /77.2%	76.9% /78.3%	79.8% /80.2%	79.2% /80.3%	79.5% /80.5%	78.6% /79.8%
atomic video concepts	ankle/foot	ankle fracture	arteria	vericose veins	obesity	endocrinic
hierarchicak boosting	93.5% /94.2%	92.8% /91.2%	90.2% /89.6%	91.6% /92.1%	92.9% /91.6%	92.1% /93.3%
multi-task boosting	82.6% /81.8%	84.6% /85.6%	82.6% /82.4%	81.2% /81.8%	83.4% /84.2%	81.9% /82.1%
multi-class boosting	78.4% /77.5%	80.2% /81.3%	78.4% /80.2%	75.7% /78.0%	78.9% /80.1%	79.8% /80.4%
atomic video concepts	lesions	ventral hernia	inguinal hernia	hepatobiliary	colonic	colon neoposia
hierarchical boosting	92.6% /92.3%	91.6% /92.3%	92.7% /91.8%	93.6% /94.3%	90.6% /91.6%	92.8% /93.2%
multi-task boosting	83.8% /82.6%	81.8% /81.7%	82.7% /81.9%	83.2% /83.8%	83.6% /83.8%	82.5% /82.5%
multi-class boosting	78.5% /75.8%	76.5% /75.8%	77.9% /78.3%	78.8% /79.2%	78.3% /79.5%	77.4% /78.3%
second-level concepts	breast-endocrine	abdominal	thoracic	multiple injury & NOS	back	shoulder
hierarchical boosting	94.3% /94.8%	93.6% /93.8%	95.2% /94.5%	93.7% /93.8%	96.3% /95.7%	94.9% /93.8%
multi-task boosting	78.6% /80.2%	81.5% /82.3%	80.6% /83.1%	80.3% /81.2%	87.2% /85.8%	84.5% /83.2%
multi-class boosting	74.5% /75.3%	76.3% /77.6%	74.2% /75.3%	76.4% /77.5%	80.5% /80.3%	78.5% /78.6%
second-level concepts	eye	gynaecology	urology	menstrual/pelvic pain	hand/foot	obstetrics
hierarchical boosting	93.6% /94.3%	93.6% /94.2%	93.4% /94.7%	94.8% /95.6%	92.6% /92.4%	94.9% /95.1%
multi-task boosting	80.5% /80.3%	80.4% /80.2%	83.5% /84.2%	86.2% /86.5%	81.2% /82.3%	83.6% /85.3%
multi-class boosting	76.1% /77.3%	78.2% /79.3%	78.4% /79.1%	81.3% /81.5%	75.3% /74.8%	76.3% /75.8%
second-level concepts	fertility	hip arthritis	vascular	colorectal	upper-GI	perianal
hierarchical boosting	93.8% /93.9%	95.2% /95.6%	94.2% /94.8%	93.6% /93.4%	92.3% /92.6%	92.3% /91.8%
multi-task boosting	81.5% /82.3%	83.2% /84.3%	85.6% /87.1%	82.4% /83.5%	82.7% /82.3%	82.8% /83.1%
multi-class boosting	78.3% /78.5%	79.2% /80.1%	80.4% /81.3%	78.7% /79.1%	78.4% /79.2%	76.4% /75.9%
third-level concepts	orthopaedics	vascular/plastics	general Gny	ear, nose & throat		
hierarchical boosting	94.7% /95.3%	95.3% /95.5%	95.6% /95.8%	96.3% /96.2%		
multi-task boosting	86.3% /85.5%	81.6% /82.2%	83.8% /85.1%	84.2% /85.6%		
multi-class boosting	78.5% /75.6%	74.3% /75.2%	76.7% /75.8%	78.6% /80.7%		

exactly be in the same place of the high-dimensional heterogeneous feature space [60]–[65], thus such simple combinations may not be able to achieve a reliable ensemble classifier for the new task. b) The tasks for detecting multiple video concepts are parallel at the same semantic level, thus the strong output correlations between the classifiers cannot be exploited effectively. c) Training the pair-wise classifiers which have larger hypothesis variances may increase the computational complexity dramatically, and large amounts of training samples are needed to achieve reliable classifier training. On the other hand, our hierarchical boosting algorithm can integrate the transferable knowledge and common features to enhance all these single-task classifiers at the same semantic level simultaneously, exploit their strong inter-concept correlations to learn a biased classifier, and generate an ensemble classifier for their parent node with higher discrimination power.

C. Comparing Multiple Approaches for SVM Video Classifier Training

By using the same sets of cumulative volumetric features and training samples for video classifier training, we have also compared the performance differences between our multimodal boosting, AdaBoost and linear SVM combination. AdaBoost

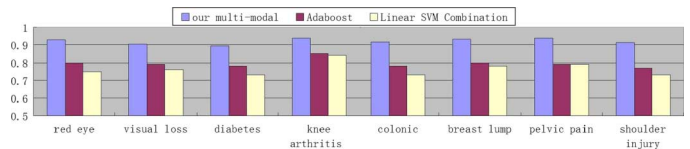


Fig. 18. Comparison results between our multimodal boosting algorithm, linear SVM, and AdaBoost.

and its variants have recently been used for feature selection by training a cascade of linear classifiers [35][36]. However, one weak classifier is learned for each feature dimension independently, and thus the *feature correlations* are ignored. Therefore, the gain on performance improvement may be limited because the feature correlations and the strong correlations between the outputs of the weak classifiers are not exploited effectively. On the other hand, our multimodal boosting algorithm uses multimodal perceptual features and it also exploits the intra-set, the intra-modality, the inter-set and the inter-modality feature correlations, and the output correlations between the weak classifiers, thus it can achieve more reliable classifier training and result in higher classification accuracy. The comparison results for some video concepts are given in Figs. 18 and 19.

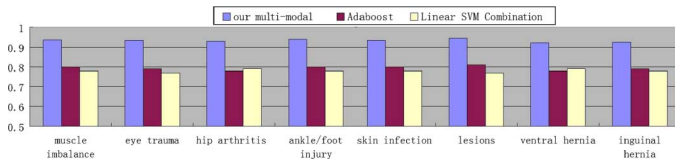


Fig. 19. Comparison results between our multimodal boosting algorithm, linear SVM, and AdaBoost.

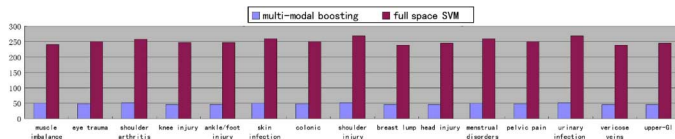


Fig. 20. Comparison results on the size of training samples between our multimodal boosting algorithm and traditional SVM classifier training algorithm (full space SVM) to learn SVM video classifier with the same accuracy rate.

It is well accepted that the number of training samples for achieving reliable classifier training largely depends on the feature dimensions [33]–[36]. To evaluate the scalability of different classifier training algorithms with the feature dimensions, we have compared two approaches on their sizes of training samples that are required for achieving the same classification accuracy rate. a) For each atomic video concept, its SVM classifier is trained by using all the cumulative volumetric features (i.e., full space). b) For each atomic video concept, our multimodal boosting algorithm is used by partitioning all the cumulative volumetric features into 11 homogeneous feature subsets. Because the feature dimensions for each homogeneous feature subset are relatively low, our multimodal boosting algorithm can significantly reduce the number of training samples to learn the classifier for the same video concept with the same classification accuracy rate, i.e., the number of training samples for the full-space approach is at least four times more than that for our multimodal boosting algorithm as shown in Fig. 20. In addition, each homogeneous feature subset is used to characterize one certain perceptual property of video concept, thus the underlying geometric property of video data is uniform and can accurately be approximated by using one specific type of probabilistic kernel functions. Learning the classifiers for the strongly correlated and sibling video concepts simultaneously (multitask learning) can also require fewer training samples per task with better generalization than that if those classifiers are learnt independently.

D. Comparing Two Approaches for Hierarchical Video Classification

There are two approaches to achieve hierarchical video classification: a) top-down approach and b) bottom-up approach. In a bottom-up approach, only the classifiers for the atomic video concepts at the first level of the concept ontology are learned, and the appearances of the relevant high-level video concepts can be predicted automatically according to the concept ontology. Compared with our top-down approach, the bottom-up approach has the following shortcomings. 1) The bottom-up approach has higher computational cost for hierarchical video classification because it has to go through $(M(M-1)/2)$ binary SVM video classifiers for all these M atomic video con-

cepts at the first level of the concept ontology. On the other hand, our top-down approach for hierarchical video classification can ignore large amounts of irrelevant video concepts early and only perform the classifiers for the video concepts on the selected hierarchical classification path, and thus the computational cost for video classification is significantly lower. 2) The contextual relationships between the video concept nodes could be hypernymy or hyponymy, one atomic video concept may be related to multiple video concepts at the second level of the concept ontology. Therefore, simple inference via the concept ontology could not accurately achieve hierarchical video classification and automatic multilevel video annotation, and this misclassification error will be propagated along the concept levels. On the other hand, our top-down approach can tackle this inter-level error transmission problem more effectively with automatic error recovery.

VIII. ALGORITHM SCALABILITY AND GENERALIZABILITY

One problem for automatic video concept detection is the large range of possible within-concept variations because of various viewing and illumination conditions. It is very important to develop new techniques that are able to effectively tackle the changes of viewing and illumination conditions. By treating various viewing conditions or illumination conditions as the additional selection units in our multimodal boosting algorithm, we can further learn the SVM video classifiers for the same video concept under different viewing or illumination conditions. Our multimodal boosting algorithm has provided a natural way to effectively combine the SVM video classifiers for different homogeneous feature subsets, viewing conditions and illumination conditions, and thus it is able to easily generalize across different viewing and illumination conditions and provide a scalable solution for such problems. In addition, our multimodal boosting algorithm is scalable to the dynamic extraction of new homogeneous feature subsets by simply adding the corresponding weak SVM classifiers into the ensemble classifier.

In this paper, we focus on one specific domain of surgery education videos because of the significant application values. Our proposed hierarchical video classifier training scheme does not depend on the video domain, because the domain-dependent concept ontology is only used for determining the task relatedness to enable multitask learning and it is not used for predicting the appearances of high-level video concepts. Therefore, it is worth noting that our hierarchical video classification scheme can easily be extended to more broader video domains such as news, films and sports. In Fig. 21, we have demonstrated some of our preliminary results on the news video domain, where the concept ontology for hierarchical video concept organization is obtained automatically by using the textual terms extracted from news closed captions [67]. The concept ontology is then used to determine the related tasks (i.e., learning the classifiers for the sibling video concepts on the concept ontology), and our proposed hierarchical video classifier training scheme can directly be extended to train the classifiers for detecting large amounts of video concepts automatically. For more broader video domains such as news, the within-concept variations may be larger and thus more training samples are needed to achieve reliable video classifier training. On the other hand, the rich production meta-

- [6] H. Luo, J. Fan, and G. Xu, "Multi-modal salient objects: General building blocks of semantic video concepts," in *Proc. ACM CIVR*, 2004, pp. 374–383.
- [7] D. Dementhon and D. Doermann, "Video retrieval of near-duplicates using k-nearest neighbor retrieval of spatiotemporal descriptors," *Multimedia Tools Appl.*, 2005.
- [8] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proc. IEEE ICCV*, 2005.
- [9] Y. Deng and B. S. Manjunath, "Netra-V: Toward an object-based video representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 616–627, 1998.
- [10] S.-F. Chang and H. Sundaram, "Semantic visual template-linking features to semantics," in *Proc. IEEE ICIP*, 1998, pp. 531–535.
- [11] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Boston, MA: MIT Press, 1998.
- [12] R. Fikes, A. Farquhar, and J. Rice, *Tools for Assembling Modular Ontologies in Ontolingua*. New York: AAAI/IAAI, 1997, pp. 436–441.
- [13] S. Satoh and T. Kanada, "Name-It: Association of face and name in video," in *Proc. CVPR*, 1997.
- [14] H. Waclar, M. Christel, Y. Gong, and A. Hauptmann, "Lessons learned from the creation and deployment of a terabyte digital video library," *IEEE Computer*, vol. 32, pp. 66–73, 1999.
- [15] A. G. Hauptmann, "Towards a large scale concept ontology for broadcast video," in *Proc. CIVR*, 2004, pp. 674–675.
- [16] M. Christel and A. Hauptmann, "The use and utility of high-level semantic features in video retrieval," in *Proc. CIVR*, 2005.
- [17] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: A multimedia information network for knowledge representation," *Proc. SPIE*, vol. 4210, 2000.
- [18] C. Jorgensen, A. Jaimes, A. B. Benitez, and S.-F. Chang, "A conceptual framework and research for classifying visual descriptors," *J. Amer. Soc. Information Science (JASIS)*, vol. 52, no. 11, pp. 938–947, 2001.
- [19] A. B. Benitez, S.-F. Chang, and J. R. Smith, "IMKA: A multimedia organization system combining perceptual and semantic knowledge," in *ACM Multimedia*, 2001.
- [20] R. Nevatia, T. Zhao, and S. Hengeng, "Hierarchical language-based representation of events in video streams," in *Proc. IEEE CVPR Workshop on Event Mining*, 2003.
- [21] M. R. Naphade, I. Kozintsev, and T. S. Huang, "Factor graph framework for semantic video indexing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 1, pp. 40–52, 2002.
- [22] M. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering and retrieval," *IEEE Trans. Multimedia*, vol. 3, no. 1, pp. 141–151, 2001.
- [23] 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 *Proc. EURASIP JASP*, 2003, vol. 2, pp. 170–185.
- [24] M. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. ACM Multimedia*, 2004.
- [25] A. Jaimes, B. L. Tseng, and J. R. Smith, "Modal keywords, ontologies, and reasoning for video understanding," in *Proc. CIVR*, 2003.
- [26] Y. Wu, B. Tseng, and J. R. Smith, "Ontology-based multi-classification learning for video concept detection," in *Proc. IEEE ICME*, 2004.
- [27] [Online]. Available: <http://dublincore.org/>
- [28] C. G. M. Snoek, M. Worring, and A. G. Hauptmann, "Learning rich semantics from news video archives by style analysis," *ACM Trans. Multimedia Comput., Commun., Applicat.*, vol. 2, no. 2, 2006.
- [29] C. G. M. Snoek, M. Worring, J. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, pp. 1678–1689, Oct. 2006.
- [30] L. Hollink, M. Worring, and G. Schreiber, "Building a visual ontology for video retrieval," in *Proc. ACM Multimedia*, 2005.
- [31] G. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, "Ontology-based photo annotation," *IEEE Intell. Syst.*, 2001.
- [32] T. Liu and J. R. Kender, "Lecture videos for e-learning: Current research and challenges," in *Proc. IEEE ISMSE*, 2004.
- [33] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, pp. 273–324, 1997.
- [34] T. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
- [35] K. Tieu and P. Viola, "Boosting image retrieval," in *Proc. IEEE CVPR*, 2000.
- [36] J. O'Sullivan, J. Langford, R. Caruana, and A. Blum, "FeatureBoost: A meta learning algorithm that improves model robustness," in *Proc. ICML*, 2000, pp. 703–710.
- [37] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [38] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods-Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.
- [39] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML*, 1996, pp. 148–156.
- [40] LookSmart [Online]. Available: <http://www.looksmart.com/>
- [41] Open Project [Online]. Available: <http://dmoz.org/>
- [42] Ontology Alignment [Online]. Available: <http://oaei.ontology-matching.org/>
- [43] A. D. Maedche, *Ontology Learning for the Semantic Web*. New York: Springer-Verlag, 2002.
- [44] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology Learning from Text: Methods, Evaluation, and Applications*. New York: IOS, 2005.
- [45] M. Sanderson and W. B. Croft, "Deriving concept hierarchies from text," in *Proc. ACM SIGIR*, 1999, pp. 206–213.
- [46] K. Punera, S. Rajan, and J. Ghosh, "Automatically learning document taxonomies for hierarchical classification," WWW pp. 1010–1011, 2005.
- [47] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Ng, "Improving text classification by shrinkage in a hierarchy of classes," in *Proc. ICML*, 1998, pp. 359–367.
- [48] S. T. Dumais and H. Chen, "Hierarchical classification of Web content," in *Proc. ACM SIGIR*, 2000, pp. 256–263.
- [49] M. Ciaramita, T. Hofmann, and M. Johnson, "Hierarchical semantic classification: Word sense disambiguation with world knowledge," in *Proc. IJCAI*, 2003.
- [50] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan, "Using taxonomy, discriminants, and signatures for navigating in text databases," in *Proc. VLDB*, 1997.
- [51] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," in *Proc. ICML*, 1997.
- [52] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, 2006.
- [53] A. Kojima, T. Tamura, and K. Fukunaga, "Natural language description of human activities from video images based on concept hierarchy of actions," *Int. J. Comput. Vis.*, vol. 50, no. 2, pp. 171–184, 2002.
- [54] A. Jaimes and J. R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies," in *Proc. IEEE ICME*, 2003.
- [55] C. A. Lindley, "A multiple-interpretation framework for modelling video semantics," in *ER-97 Workshop on Conceptual Modeling in Multimedia Information Seeking*, 1997.
- [56] J. Hunter, "Enhancing the semantic interoperability of multimedia through a core ontology," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 13, pp. 49–58, Jan. 2003.
- [57] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, pp. 135–168, 2000.
- [58] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–374, 2000.
- [59] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: Efficient boosting procedures for multiclass object detection," in *Proc. CVPR*, 2004.
- [60] S. Ben-David and R. Schuller, "Exploiting task relatedness for multilple task learning," in *Proc. COLT*, 2003, pp. 567–580.
- [61] J. Baxter, "A model for inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, pp. 149–198, 2000.
- [62] K. Yu, A. Schwaighofer, V. Tresp, W.-Y. Ma, and H. J. Zhang, "Collaborative ensemble learning: Combining content-based information filtering via hierarchical Bayes," in *Proc. Int. Conf. Uncertainty in Artificial Intelligence (UAI)*, 2003.
- [63] S. Thrun and L. Pratt, *Learning to Learn*. Norwell, MA: Kluwer, 1997.
- [64] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, pp. 615–637, 2005.
- [65] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. ACM SIGKDD*, 2004.
- [66] R. L. Gruen, S. Knox, H. Britt, and R. S. Bailie, "The surgical nosology in primary-care setting (SNIPS): A simple bridging classification for the interface between primary and specialist care," *BMC Health Services Res.*, vol. 4, no. 8, 2004.
- [67] H. Luo, J. Fan, J. Yang, I. Satoh, and W. Ribarsky, "Large-scale news video visualization," in *Proc. IEEE VAST*, 2006.
- [68] K. Crammer and Y. Singer, "On the algorithmic implementation of multi-class kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2001.

- [69] T. Dietterich and G. Bakiri, "Solving multiclass learning problem via error-correcting output codes," *J. Artif. Intell.*, vol. 2, pp. 263–286, 1995.
- [70] D. E. Knuth, *The Art of Computer Programming, Sorting and Searching*. Reading, MA: Addison-Wesley, 1978, vol. 3.
- [71] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, to be published.
- [72] M. Koskela, A. F. Smeaton, and J. Laaksonen, "Measuring concept similarities in multimedia ontologies: Analysis and evaluations," *IEEE Trans. Multimedia*, to be published.
- [73] G. P. Nguyen and M. Worring, "Similarity based visualization of image collections," in *Proc. AVIVDiLib*, 2005.
- [74] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Proc. IEEE ICCV*, 1998, pp. 59–66.
- [75] D. Stan and I. Sethi, "eID: A system for exploration of image databases," *Inform. Process. Manage.*, vol. 39, pp. 335–361, 2003.
- [76] B. Moghaddam, Q. Tian, N. Lesh, C. Shen, and T. S. Huang, "Visualization and user-modeling for browsing personal photo libraries," *Int. J. Comput. Vis.*, vol. 56, pp. 109–130, 2004.
- [77] J. A. Walter and H. Ritter, "On interactive visualization of high-dimensional data using the hyperbolic plane," in *Proc. ACM SIGKDD*, 2002.
- [78] J. Lamping and R. Rao, "The hyperbolic browser: A focus+content technique for visualizing large hierarchies," *J. Vis. Lang. Comput.*, vol. 7, pp. 33–55, 1996.



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 science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997.

He was a Researcher at Fudan University, Shanghai, during 1998. From 1998 to 1999, he was a Researcher with Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From September 1999 to 2001, he was a Researcher in the Department of Computer Science, Purdue University, West Lafayette, IN. In 2001, he joined the Department of Computer Science, University of North Carolina at Charlotte as an Assistant Professor and then became Associate Professor. His research interests include content-based image/video analysis, classification and retrieval, surveillance videos, and statistical machine learning.

Hangzai Luo received the B.S. degree in computer science from Fudan University, Shanghai, China, in 1998. He received the Ph.D. degree in information technology in 2006 from the University of North Carolina at Charlotte.

His research interests include computer vision, video retrieval, and statistical machine learning.

Dr. Luo received the second place award from the Department of Homeland Security at 2007 for his excellent work on video analysis and visualization for homeland security applications.

Yuli Gao received the B.S. degree in computer science from Fudan University, Shanghai, China, in 2002. He received the Ph.D. degree in information technology from the University of North Carolina at Charlotte in 2007.

His research interests include computer vision, image classification and retrieval, and statistical machine learning.

Dr. Gao received an award from IBM as Emerging Leader in Multimedia in 2006.

Ramesh Jain is the Bren Professor of Information and Computer Science, Department of Computer Science, University of California, Irvine. He has been an active researcher in multimedia information systems, image databases, machine vision, and intelligent systems. While he was at the University of Michigan, Ann Arbor, and the University of California, San Diego, he founded and directed artificial intelligence and visual computing labs. He was the founding Editor-in-Chief of *IEEE Multimedia Magazine* and *Machine Vision and Applications* and serves on the editorial boards of several magazines in multimedia, business, and image and vision processing. He has co-authored more than 250 research papers. His current research is in experiential systems and their applications.

Dr. Jain is a Fellow of ACM, IAPR, AAAI, and SPIE.