01 Jan 2005

An HMM-Based Framework for Video Semantic Analysis

Gu Xu
Missouri University of Science and Technology

Yu-Fei Ma

Hong-Jiang Zhang

Shiqiang Yang

Recommended Citation

G. Xu et al., "An HMM-Based Framework for Video Semantic Analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), Jan 2005.
The definitive version is available at https://doi.org/10.1109/TCSVT.2005.856903

An HMM-Based Framework for Video Semantic Analysis

Gu Xu, Yu-Fei Ma, *Member, IEEE*, Hong-Jiang Zhang, *Fellow, IEEE*, and Shi-Qiang Yang, *Member, IEEE*

***Abstract*—Video semantic analysis is essential in video indexing and structuring. However, due to the lack of robust and generic algorithms, most of the existing works on semantic analysis are limited to specific domains. In this paper, we present a novel hidden Markove model (HMM)-based framework as a general solution to video semantic analysis. In the proposed framework, semantics in different granularities are mapped to a hierarchical model space, which is composed of detectors and connectors. In this manner, our model decomposes a complex analysis problem into simpler subproblems during the training process and automatically integrates those subproblems for recognition. The proposed framework is not only suitable for a broad range of applications, but also capable of modeling semantics in different semantic granularities. Additionally, we also present a new motion representation scheme, which is robust to different motion vector sources. The applications of the proposed framework in basketball event detection, soccer shot classification, and volleyball sequence analysis have demonstrated the effectiveness of the proposed framework on video semantic analysis.**

***Index Terms*—Event detection, hidden Markov models (HMMs), sports videos, video semantic analysis.**

I. Introduction

EFFICIENT video indexing and retrieval have emerged as challenging research problems in multimedia applications. Various features such as color, shape, texture, and motion have been used for video indexing and retrieval. However, the performance of the existing approaches to video indexing and retrieval using these features is far from satisfactory due to the gap between these low-level features and the high-level semantics presented in video data. Therefore, recent works on video indexing and retrieval have witness an interesting shift from low-level feature-based approaches to high-level semantic analysis.

In this paper, we argue that video semantic analysis should be decomposed into analysis at a number of granularities, instead of viewing it at a single level. To date, most existing approaches to video semantic analysis only focus on a particular semantic level. From the point view of human cognition, however, the process of understanding is an interaction of information and knowledge at different semantic levels or granularities. For instance, when we read a book, from the visual text to content understanding, the information is continually transformed and abstracted through different semantic granularities, as illustrated in Fig. 1. As a partial reproduction of human perception, automatic speech recognition also has achieved success by using constraints in several semantic granularities, e.g., phoneme, word, and language grammar. Therefore, integrating context constraints in different semantic granularities may be a key point to push on the research on video semantic analysis further.

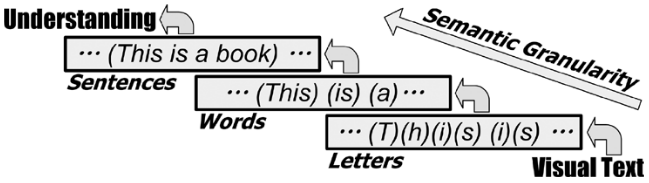


Fig. 1. Multilevel semantic understanding in book reading.

Video signals are not arbitrary but are ruled by the continuity of contents. Comparing with the book reading example, we also could explain the continuity of contents in the form of constraints in different granularities. For example, in broadcasted sports videos, the composition of events is controlled by shot category and the sequences of shots are partially determined by game genre. More specifically, we may make an analogy with the typical speech recognition for better understanding, viz. event versus phoneme, shot versus word, and sports genre versus grammar. The basic ides of our approach is to introduce some kinds of ”video grammar” to help video analysis and even general computer vision problems. However, because the semantic structure of videos, named “video grammar” also, is always under constraint and unstable in various applications, we will need a more general framework than that of the speech recognition to ensure that the approach could be applied.

In our approach, video semantic analysis may be viewed as a set of filters in different semantic granularities, which extract semantics hierarchically from visual or audio features. Consequently, we proposed a layered framework in which context constraints in different semantic granularities are represented at separate layers. With this framework, analysis tasks in different semantic granularities, e.g., event detection and shot classification, may be integrated into a single model and obtain more robust results by interacting and sharing internal information for the correlations of those tasks. There are two types of filters in this framework, namely, detectors and connectors. Detectors determine the existence and confidence of certain semantics according to low-level features, and connectors model the causalities between semantics. By assuming that the causalities between semantics follow a certain stochastic process, uniform connectors will be built. In this paper, we adopt hidden

Manuscript received May 12, 2003; revised October 10, 2003. This paper was recommended by Associate Editor H. Gharavi.

G. Xu and H.-J. Zhang are with Microsoft Research Asia, Beijing 100080, China (e-mail: t-gxu@microsoft.com; hjzhang@microsoft.com).

Y.-F. Ma is with the IBM China Research Laboratory, Beijing 100101, China (e-mail: mayufei009@yahoo.com).

S.-Q. Yang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: yangshq@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2005.856903

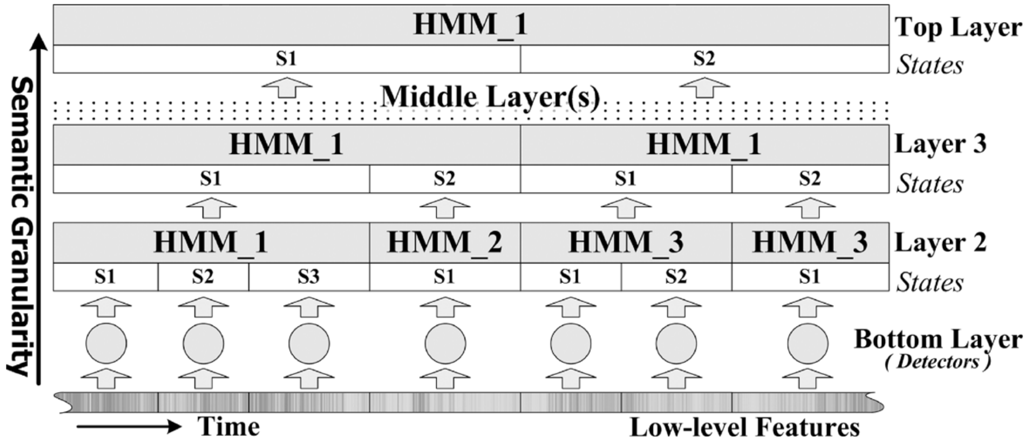


Fig. 2. Overview of the proposed framework.

Markov models for efficiency to constitute uniform connectors. As shown in Fig. 2, detectors produce hypotheses on the semantics at the bottom layer, and connectors optimize the hypotheses by maximizing the probability at the top layer.

The proposed framework is a generic and extendable model for video semantic analysis. The main differences and advantages of this framework compared with existing ones are as follows.

1) *A uniform solution to video semantic analysis.* The proposed methodology is not tied to a specific analysis task of videos in a specific domain, but is easy to adapt to all specific analysis tasks as it converts the analysis task into modeling detectors and connectors.
2) *An efficient and simple representation to context constraints.* The first-order hidden Markov models (HMMs) are simple yet powerful. Therefore, in terms of implementation, complex context constraints can be efficiently built by simple connectors in a uniform manner, that is, first-order HMMs.
3) *An integrated model for semantic extraction and segmentation in multiple semantic granularities.* In the proposed model, a number of semantic granularities are considered together. The analysis process is integrated by probabilities, and recognition and segmentation are solved synchronously.

The remainder of this paper is organized as follows. In Section II, we briefly review the previous works on video semantic analysis and compare the proposed framework with existing HMM-based approaches. In Section III, the HMM based computational framework is described in detail. In Section IV, example applications of this novel framework are presented and discussed. All experimental results are given in Section V. Finally, Section VI concludes this paper. Moreover, possible future work is also presented in this section.

II. Related Work

There have been many attempts at video semantic analysis, and some limited successes have been reported in general application domains, such as periodical motion finding work [1], [2] and temporal matching work [3]. Although those unsupervised approaches have more expansibility when dealing with general video data, the semantics extracted are limited to simple ones. Other recent work mostly focused on extracting high-level semantics in confined application domains. In this manner, video content may be modeled by supervised approaches. Such supervised semantic analysis mainly falls in three categories, namely, video genre classification, concept learning, and event detection/recognition.

The video genre is the broad class to which a video may belong, such as sports, news, and feature movies. Roach *et al.* [4] have classified video sequences into three predefined broad classes of genres by motions. A reliable approach was presented in [5], in which more genres have been classified. At a finer semantic level, Messer *et al.* [6] proposed semantic cues to classify sports videos into different types, such as tennis, swimming, and yachting. However, these classification approaches are all limited to the coarsest level of semantics.

The objective of video concept learning is trying to automatically index video content in a database by concepts, usually represented by keywords. Liu and Bhanu [7] proposed a color-based approach to learn visual concepts in videos. Bayesian networks were adopted in [8] to classify video contents according to feature sensors. In order to exploit the interrelationships of concepts, language-like model [9] and graphical model [10], [11] have been used to model the dependences between concepts. Instead of classifying key frames into several categories, shot classification may be deemed as a special case of concept learning in the clip level. Although there is a lack of general learning models for arbitrary videos, it has been widely used in specific domains, for example, news [12] and sports [13].

Events are the most important part of video semantic contents. Events represent the temporal interactions and variation that compose the story line in a video program. Because sports activities have clear semantic structure, many research efforts have been attracted on sports video analysis. Tan *et al.* [14] proposed a system to detect basketball events using motion vectors from MPEG streams. In the work presented in [15] and [16], events in tennis and soccer games were recognized by semantic object information, such as court-line and players locations. Zhong and Chang [17] combined the statistical model and manual verification rules for recurrent event detection, such as pitching and serving views in baseball games. The rules in these works were defined explicatively and manually. A different approach was presented in [18], in which rules were trained by an entropy-based inductive tree-learning algorithm.

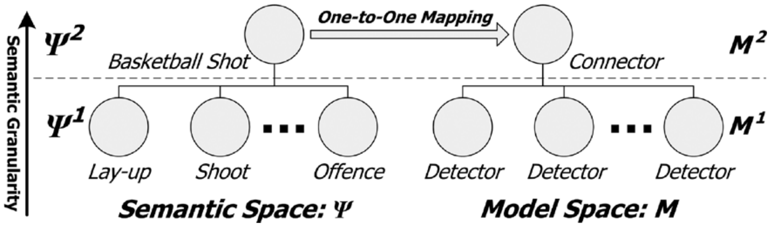


Fig. 3. Semantic space and model space in basketball event detection.

As another important aspect of video content analysis, context constraints have been addressed in many literatures. Due to the temporal interactions and interrelations between frames in video, HMMs are a natural and simple way to model temporal relationships. Liu and Kender [19] used HMMs to explore the structure of documentaries. Meanwhile, they also have conjectured that the parameters of trained HMM's may represent certain semantics, named "*stylistic*." As an extension of this idea, Sánchez *et al.* [20] proposed an unsupervised scheme for video content clustering. The approach is based on the parameters of Markov chains instead of low-level features. Rather than the relationships between video frames, Markov assumptions are also made, separately, at different levels depending on the applications. In the event level, Petkovic *et al.* [21] have recognized strokes in tennis by HMMs. Xie *et al.* [22] proposed an HMM-based approach for soccer play/break event detection. Video contents in one shot are temporally continuous, indicating that shots with similar contents may show similar temporal patterns. For example, Huang *et al.* [23] employed HMMs for video shot classification. The attempts at exploiting the relationships between shots by using HMMs are also reported. Wolf [24] has used HMMs to parse video programs for semantic recognition. A similar approach is adopted in [25] for slow-motion highlight detection. In order to utilize the video syntax more efficiently, multilayer HMMs have been used in [22] and [26]. Meanwhile, they also were employed for other purposes, such as modality fusing [10] or gesture abstraction [27] in video content analysis. In these approaches, isolated HMMs in different layers are connected directly or indirectly, and the final results are given by maximizing the probability of the entire model instead of certain HMM. However, few of these aforementioned approaches are able to extend to more generic analysis tasks.

The proposed framework may be similar to the aforementioned works based on HMM. However, the idea is quite different. Connectors in our proposed framework not only provide the video syntax constraints, but also represent video semantics. The video semantics in different granularities are entirely mapped to a model space composed of detectors and connectors. From the point view of applications, the framework will allow correlative tasks to be solved more efficiently by bringing them together rather than dealing with them separately.

III. HMM-Based Framework for Video Semantic Analysis

A video program is not merely a sequence of images. The temporal context information in videos is one of significant cues for content understanding. The purpose of video semantic analysis may be of discovering the hidden states or semantics behind video signals. In this view, video signals could be looked as the observations of semantics. If we denote video signals as O_t and semantics as S_t, therefore in terms of probabilities, the basic approach in semantic analysis may be formulized as follows:

$$S_t = \arg\max_s \Pr(s\,|O_t)\Pr(s\,|S_{t-1}, S_{t-2}, \ldots). \tag{1}$$

Obviously, any analysis approach could be divided into two parts, that is, likelihood models $\Pr(s|O_t)$ and temporal context constraints $\Pr(s|S_{t-1}, S_{t-2}, \ldots)$, represented as detectors and connectors, respectively, in this paper. Though the form of detectors depends on features and applications, the connectors may be universal if we assume that context constraints follow a certain stochastic process. HMMs have been proven to be effective in sequential pattern analysis and have been successfully applied in speech recognition [28]. In this section, therefore, we present how to build a general framework using uniform connectors, namely HMMs.

A. Architecture of Framework

As aforementioned, to be a generic solution to video semantic analysis, the proposed framework also should be composed of two components, that is, likelihood models and context constraints. In this paper, we use the concepts of detectors and connectors to describe them precisely.

Before the detailed descriptions are given, we first introduce two terms, that is, semantic space and model space. Actually, from the point view of supervised approaches, semantics involved in a specified application are finite, which constitute the semantic space Ψ. An alternative denotation is Ψ^k, which is the subset of semantics in the kth semantic granularity. In addition, we virtually create a model space M equivalent to the semantic space, as shown in Fig. 3. Each semantic in Ψ is mapped to a submodel (detector or connectors) in M. For convenience, we also use M^k to denote the subset of the submodels corresponding to semantics in Ψ^k. Obviously, there is a one-to-one relation between semantic space Ψ and model space M. In this manner, video semantic analysis may be translated into a problem of seeking an optimal sequence of submodels in the model space.

In this mapping, we divide the semantic space into two parts corresponding to detectors and connectors, respectively, in the model space. Actually, semantics in each granularity are evidently correlated if we take into account the abstracted semantics in a higher granularity. Therefore, connectors are used to model the context constraints provided by the abstracted semantics and rule the composition of semantics at the lower granularity. However, in the lowest granularity, the relationships become much more complicated because low-level features are

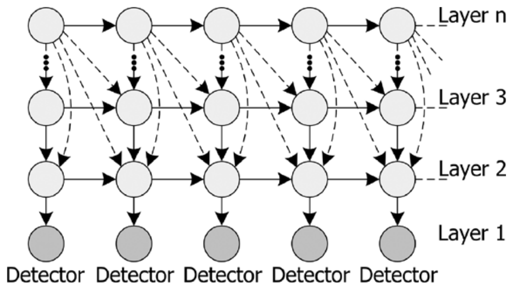


Fig. 4. State dependences unrolled in time.

involved. So, in order to make the definition of the framework more general, another kind of submodels, that is detectors, is proposed to draw a confidence inference of certain semantics from the low-level features. In conclusion, M^1 is composed of detectors and the rest are connectors.

Because the form of detectors depends on low-level features and applications, detectors could be considered as black boxes in our formulation, whose inputs are low-level features and outputs are probabilities of the presence of semantics. For connectors, a uniform definition may be achieved by inheriting from that of HMMs, that is,

$$M_i^k = (\pi_{M_i^k}, A_{M_i^k}, M^{k-1}), \qquad M_i^k \in M^k, k > 1 \tag{2}$$

where π is the initial probability vector and A is the transition probability matrix. They are only trainable parameters in connectors. The specification of states is completely determined by M^{k-1}. The state number equals to $|M^{k-1}|$ and the state likelihoods are the outputs of M^{k-1}.

More generally, we suppose that the semantics in Ψ span n granularities. Therefore, the model space is also composed of n layers. Fortunately, the n-layer model space could be considered one HMM with $\prod_{k=1}^{n} \left|M^k\right|$ hidden states. If we align the whole model with video timeline, the state dependences unrolled in time are shown in Fig. 4.

As mentioned earlier, the n-layer model space could be denoted as one HMM $\{S, \pi, A, B\}$, where S is the set of states, π is a vector of state initial probabilities, A is a matrix of state transition probabilities, and B is the collection of observation probability distribution in each state. More specifically, $\pi(s)$ is the initial probability of state s, $A(s_1, s_2)$ is the transition probability from state s_1 to s_2, and $B_s(O)$ is the probability of state s given observation O. Each element s in S is a combination of submodels in each layer, written as $s = (s^1, s^2, \ldots, s^n)$, where s^k is any element of M^k. The submodel $s^k (k > 1)$ is a connector, which is denoted in the form of (2). All of the parameters in the n-layer model may be computed according to all the submodels as follows:

$$|S| = \prod_{k=1}^{n} \left|M^k\right| \tag{3}$$

$$\pi(s) = \pi\left(s^1, s^2, \ldots s^n\right) = \prod_{k=1}^{n-1} \pi_{s^{k+1}}\left(s^k\right) \tag{4}$$

$$\begin{aligned} A(s_1, s_2) &= A\left(\left(s_1^1, s_1^2, \ldots s_1^n\right), \left(s_2^1, s_2^2, \ldots s_2^n\right)\right) \\ &= \begin{cases} A_{s_1^{p+1}}\left(s_1^p, s_2^p\right) \prod_{k=1}^{p-1} \pi_{s_2^{k+1}}\left(s_2^k\right), & n > p \geq 2 \\ A_{s_1^2}\left(s_1^1, s_2^1\right), & p = 1 \end{cases} \\ p &= \max\left\{k \left| s_1^k \neq s_2^k, n \geq k \geq 1\right.\right\} \end{aligned} \tag{5}$$

$$B_s(O) = B_{(s^1, s^1, \ldots s^n)}(O) = s^1(O) \tag{6}$$

where $\pi_{s^k}(x_0)$ and $A_{s^k}(x_1, x_2)$ are the parameters of the submodel s^k, which are defined as the initial probability of x_0 and the transition probability from x_1 to x_2 (x_0, x_1, and x_2 are the elements of M^{k-1}), and $s^1(O)$ is the probability output of detector s^1 when the observation O is given. Thus, we have integrated all submodels into one HMM, and the optimization may be achieved by *Viterbi* algorithm. When the optimal sequence of sub-models is determined, the most likely semantics in each granularity are also obtained, including temporal segmentation boundaries.

B. Submodel Training

As basic components of the proposed framework, detectors and connectors are highly coupled in recognition. From the perspective of training, however, each submodel is completely independent. In this manner, a complex model is decomposed into several simple parts, which greatly reduces the complexity of model training. The training of detectors should be trying to maximize the probability outputs with respect to training samples. However, for the various applications and available features, the implementation of detectors may quite different. Therefore, in this section, we mainly discuss the training of connectors.

As aforementioned, connectors introduce constraints on the composition of submodels at lower layer. So the objective of connectors' training is to model the dependences between semantics at lower granularity along temporal axis. Given training samples and semantic space, the samples can be hierarchically segment into semantic clips at each granularity from the top down, as shown in Fig. 5. For a connector $M_i^k (M_i^k \in M^k, k > 1)$, equivalent to Ψ_i^k, all the clips labeled as Ψ_i^k are picked out and their label sequences in the $k - 1$th granularity are taken as training data. Denote the set of training samples of M_i^k as $D^{M_i^k} = \{D_1^{M i^k}, D_2^{M i^k}, \ldots, D_C^{M i^k}\}$. For each training sequence $D_j^{M i^k}$, the ith element is denoted as $D_j^{M i^k}(i)$, where $D_j^{M i^k}(i)$ an element of M_i^{k-1}. In the maximum-likelihood formulation, the state initial probability vector π and the transition probability matrix A are calculated as follows:

$$\pi_{M_i^k}(m) = \frac{1}{C}\left|\left\{j \left| D_j^{M_i^k}(1) = m, C \geq j \geq 1\right.\right\}\right| \\ \left(m \in M^{k-1}\right) \tag{7}$$

$$A_{M_i^k}(m_1, m_2) = \frac{N(m_1, m_2)}{\sum\limits_{m \in M^{k-1}} N(m_1, m)}\left(m_1, m_2 \in M^{k-1}\right)$$

$$N_{(a,b)} = \left|\left\{j \left| D_j^{M_i^k}(i) = a, D_j^{M_i^k}(i+1) = b, C \geq j \geq 1\right.\right\}\right| \tag{8}$$

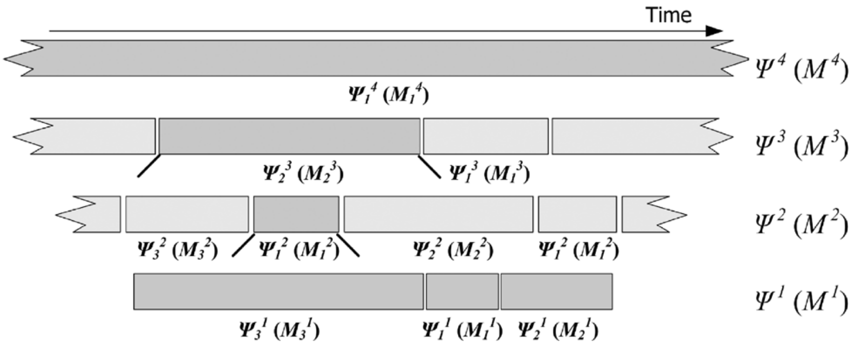


Fig. 5. Hierarchical segmentation.

where $N(a,b)$ calculates the number of pairs in sequences in which a is followed by b.

Although the specific implementations of detectors are not presented in this section, how to balance the payload between detectors and connectors is still critical for the overall model. The detectors should not be too strong so that the connectors have chances to correct errors caused by noises. Meanwhile, weak detectors are easy to be trained and reused in other application domains.

IV. Applications of the Proposed Framework

In this section, we introduce the usage of the proposed generic framework in detail. As sports activities are governed by rules, the semantics in sports videos are usually well defined. Therefore, we use sports videos as sample applications to evaluate the proposed framework. Recent work on semantic analysis in sports videos mainly focuses on three semantic granularities, i.e., events, shot categories and genres. In this work, the proposed framework has been applied to three kinds of video, one at each of three semantic granularities: basketball event detection, soccer shot classification, and volleyball sequence analysis. Since motion is the most important feature for capturing the semantic contents in video, especially for sports videos, we have proposed a novel representation of dominant motions for these sample applications. Also, instead of precisely detecting and tracking object as many other works, a statistical learning approach is adopted to build semantic models based on a set of motion filters.

A. Statistical Motion Feature

The proposed motion representations are generic and may be applied to common motion analysis for broad video content analysis. There are two key ideas in the proposed motion representation converting a video to a temporal feature sequence. First, motions between two video frames are viewed as an energy redistribution process, in which each video frame is represented by an energy distribution function. Second, a set of motion filters are designed, and each is most responsive to a certain type of dominant motion. These filters are applied to the sequence of energy distribution functions, respectively, and each response sequence presents the characteristics of certain motion pattern on time axis.

1) Energy Redistribution: This motion representation is derived from motion vector fields (MVFs), which are estimated by block-based motion estimation algorithms. Although the real motion in videos often cannot be accurately described in the way of MVFs, the loss is trivial compared to its efficiency. Particularly, if videos are in MPEG format, the motion vector fields are readily available. The proposed representation views motion vectors as the forces to alter the distribution of "energy" associated with each block from which one motion vector is extracted, and measures the distribution change between two frames by an energy redistribution function.

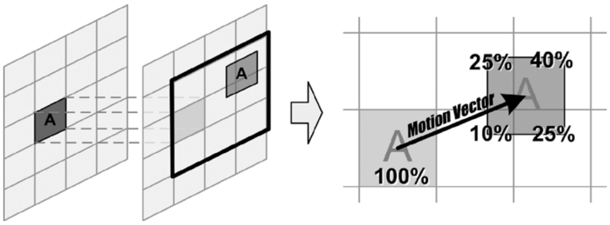


Fig. 6. Hierarchical segmentation.

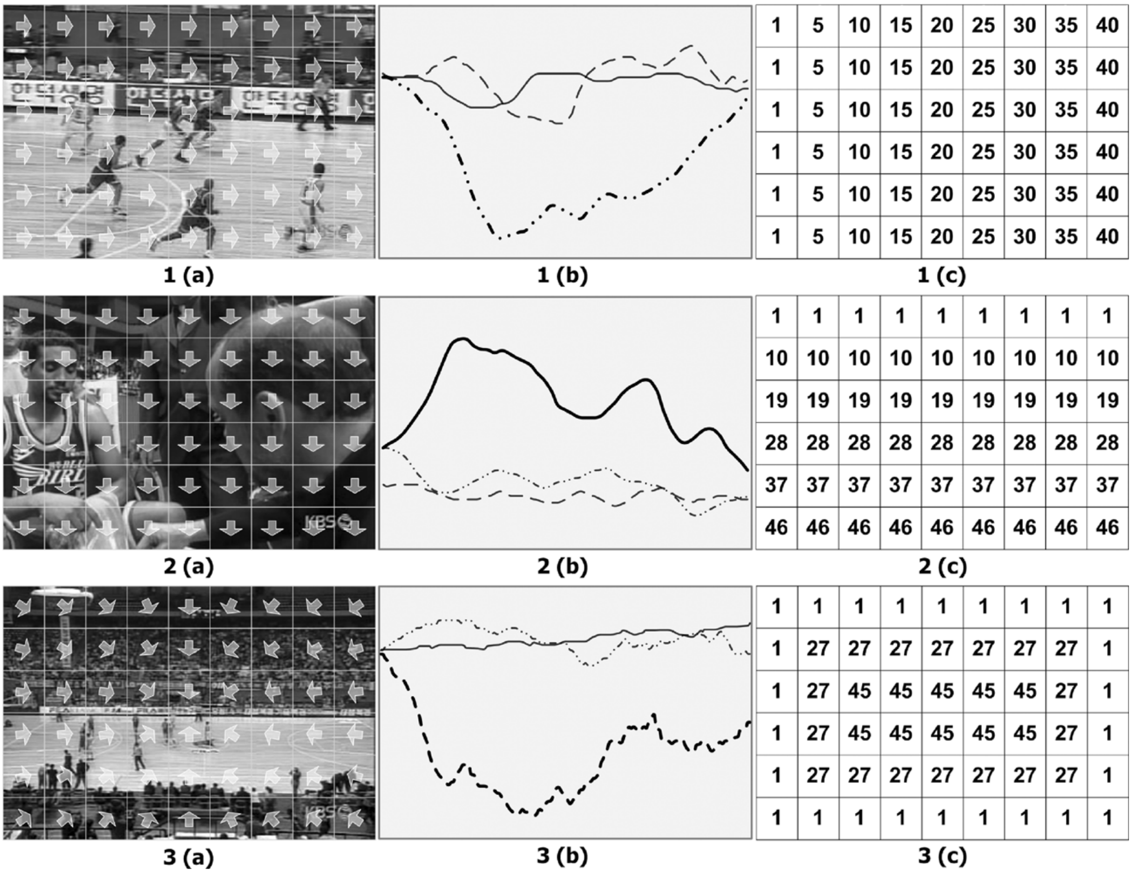
More specifically, each block in MVFs is viewed as a basic energy container. We assume that all of blocks in the initial frame have the same amount of energy. Motion vectors are figured as the outside forces that cause energy exchange between blocks, as show in Fig. 6. Therefore, the change of energy distribution may reflect motion characteristics.

The redistribution of energy depends on the corresponding position in the next frame. The energy at $block(x,y)$ is denoted by $E_{x,y}$. The energy redistribution function is represented as

$$E'_{x,y} = \frac{\sum_{i,j} (\text{overlap}S_{i,j,x,y} \times E_{i,j})}{W_b^2}, \qquad i,j \in [1, W_b] \tag{9}$$

where $overlapS_{i,j,x,y}$ denotes the overlap portion of the rectangular regions corresponding to $block(i,j)$ in the previous frame and $block(x,y)$ in the current frame, and W_b is the size of blocks. If a block moves out of the frame boundary, in order to keep the same amount of energy, we decrease the magnitude of vector to ensure the block is just in frame.

2) Motion Filters: In order to discover temporal motion patterns, we need to convert such measure of energy distribution function into temporal motion features. For this purpose, we have designed a set of motion filters, each of which is a weight matrix with the same size as blocks divided in video frames. Elements in the weight matrix are denoted by $w_{i,j}$. When we



1 (c)

1	5	10	15	20	25	30	35	40
1	5	10	15	20	25	30	35	40
1	5	10	15	20	25	30	35	40
1	5	10	15	20	25	30	35	40
1	5	10	15	20	25	30	35	40
1	5	10	15	20	25	30	35	40

2 (c)

1	1	1	1	1	1	1	1	1
10	10	10	10	10	10	10	10	10
19	19	19	19	19	19	19	19	19
28	28	28	28	28	28	28	28	28
37	37	37	37	37	37	37	37	37
46	46	46	46	46	46	46	46	46

3 (c)

1	1	1	1	1	1	1	1	1
1	27	27	27	27	27	27	27	1
1	27	45	45	45	45	45	27	1
1	27	45	45	45	45	45	27	1
1	27	27	27	27	27	27	27	1
1	1	1	1	1	1	1	1	1

Fig. 7. Examples of feature curves. (a) Key frame of clips. (b) Feature curve. (c) Example for weight templates (1: horizontal motion; 2: vertical motion; 3: radial motion).

apply each of these filters to the energy distribution of a video frame, the response in a frame is defined as

$$E_R = \sum_{i,j} E_{i,j} \times w_{i,j}, \qquad i,j \in [1, W_b]. \qquad (10)$$

By arranging elements in a weight matrix with different values, we may design filters with different sensitivities to different motion patterns. As shown in Fig. 7, each plot (b) contains three curves, i.e., time series of the responses to three filters. The key-frame of the corresponding video clips are shown in (a), and the filters used and result in strong responses (indicated by bold curves) are listed in (c). The crests on curves indicate the presence of salient motions, and the type and shape of crests show the direction and characters of the motions.

3) Sequential Feature Curves: Like in audio signal processing, a sliding window is used in calculating motion features of a video sequence. The first frame in the window is the initial frame with even energy distribution. Then, the energy redistribution function (9) is applied to frames in the window one by one, until the last frame is reached. This process produces a sequence of energy distribution functions. Each function is filtered by three motion filters, respectively, to produce three sequences of responses, as defined by (10). We calculate the means of three response sequences within sliding window separately and use them to present the motions simply but effectively. The three motion filters are designed to detect three kinds of dominant motions, say, horizontal, vertical, and radial, respectively, as shown in Fig. 7. The width of the sliding window and the sampling frequency (defined by the number of skipped frames when the window slides) determine the accuracy of results. Therefore, it is easy to balance computational complexity and performance by adjusting the two parameters.

B. Applications to Sports Videos

In this section, the applications to sports videos will be discussed in detail. According to the proposed motion features, a direct approach to detector design is a single Gaussian model. Although simple detectors have more tolerance on noises, a single Gaussian model may be too simple to model meaningful semantics, which leads to an inconvenience of manual annotation. Therefore, we have to find those "hidden" detectors through the semantics in a higher granularity. Fortunately, if we consider the detector layer and the first connector layer together, we obtain the common HMMs [28], named compound connectors. In this manner, we only need to determine the number of detectors for each abstracted semantic, that is, the number of states of each connector in the second layer. So, the detectors at the bottom layer and the connectors at the second layer are trained together. In our implementation, the same architecture will be used in all sample applications. It must be mentioned that the form of detectors is not limited to a single Gaussian model what is used in this paper.

Although sample applications have different focuses in research, the proposed framework, a general solution to video semantic analysis, is implemented in a similar manner. In the

TABLE I
MODEL AND SEMANTIC SPACE IN THE BASKETBALL APPLICATION

Layer	Sub-models	Semantics
3rd	Connectors, (HMM)	Basketball Shot
2nd	Connectors, (HMM)	16 Pre-defined Basketball Events
1st	Detectors, (Single GM)	Hidden Semantics

TABLE II
MODEL AND SEMANTIC SPACE IN THE SOCCER APPLICATION

Layer	Sub-models	Semantics
4th	Connectors, (HMM)	Soccer Video Sequence
3rd	Connectors, (HMM)	7 Pre-defined Soccer Shot Categories
2nd	Connectors, (HMM)	12 Pre-defined Soccer Sub-semantics
1st	Detectors, (Single GM)	Hidden Semantics

feature part, only the proposed motion representation is used, except the soccer application. Three filters presented in Section III-A, which are sensitive to horizontal, vertical, and radial motions, respectively, are employed to generate three motion curves. In theory, they are complete for any dominant motion that may be regarded as a linear combination of the three-directional motions. Because some temporal patterns are only distinct in the differential curves, we combine three original and three differential curves, and one six-dimensional vector is used as observation vector and input into compound connectors. In training process, video sequences are manually annotated and segmented at each granularity. The video clips at the lowest granularity are the training samples of compound connectors and the sequences of semantics in each granularity are used to build regular connectors described in Section III-B. After all submodels are trained independently, the integrated model constituted by submodels is able to segment the test sequences into separate clips and annotate those clips in different granularities. In this manner, video sequence may be automatically converted into a number of semantic sequences at different granularities. By using the information of semantic sequences, such as in sports videos, users are allowed to quickly locate the clips of interests from a long video sequence based on semantics and further customize the highlight generated by computers.

1) Basketball Event Detection: Most of existing works on event detection assume that video sequences are presegmented into clips and each clip contains only one event. Our approach is different. In the proposed framework, our goal is to find an optimal sequence of semantics to explain video content. Therefore, the tasks of our framework are not only recognition but also segmentation. We have applied the proposed framework to basketball videos to decompose each shot in basketball videos into a sequence of predefined events. Semantic space and corresponding model space in this application are shown in Table I.

According to the game rules, editing manners, and viewers' interests, we defined 16 basketball events: 1) offence at left court; 2) offence at right court; 3) fast break to left court; 4) fast break to right court; 5) lay-up at left court; 6) lay-up at right court; 7) shot at left court; 8) shot at right court; 9) track player to left; 10) track player to right; 11) lay-up in close-up view; 12) shot in close-up view; 13) foul shot in close-up view; 14) general close-up; 15) wipe; and 16) stillness. Obviously, some of these definitions are not considered as events in existing work. Our approach is to explain video content with semantics, which requires that the predefined semantics are complete for a specific domain. Therefore, we may need to define some "nonevent" semantics, such as "stillness" and "general close-up," to tolerate noises on the timeline and avoid model breakdown in the HMM recognition process.

The detailed implementations of the framework have been presented at the beginning of this section. However, how to define a state topology for each compound connector still remains a problem. The direct approach is to define a specific topology for each compound connector by manually studying training samples. However, it is too time-consuming, especially when the event variation is large. Considering the fact that the connections between states can be broken in training process, we use a complete connected six-state compound connector as a general prototype for all basketball events. According to the experimental results, the six-state compound connector is reasonable for all events' modeling. In addition, while definitely requiring more training data, increasing the number of states has no distinct improvement on recognition performance.

In this application, the optimization is achieved at shot level. That is, the model segments each shot into event clips independently. The results of basketball event detection given by our system are the event sequences of shots, including event boundaries.

2) Soccer Shot Classification: Soccer is one of the most popular worldwide sports and attracts a great deal of research attentions [13], [22]. Play/break status in soccer videos is important information for soccer video analysis, in which shot classification is a key technology [22]. Therefore, we also applied this framework to soccer shot classification. Although the granularity of shot categories is higher than that of events, we still define coarse semantics at event level to increase the overall toleration of shot variations. The semantic space and corresponding model space in soccer shot classification are shown in Table II.

The seven predefined shot categories are: 1) wide-angle; 2) zoom-in; 3) close-up (after wide-angle); 4) close-up (after replay); 5) wipe (before replay); 6) wipe (after replay); and 7) replay. Also, we decompose those categories into several subsemantics at event level, including: 1) play in the field; 2) shoot or attempt to shoot in the field; 3) track player to left in zoom-in view; 4) track player to right in zoom-in view; 5) general close-up in zoom-in view; 6) track player to left in close-up view; 7) track player to right in close-up view; 8) general close-up in close-up view; 9) wipe; 10) track player or ball to left in replay view; 11) track player or ball to right in replay view; and 12) general close-up in replay view. Such definitions of semantics at event level are coarse and do not even contain most of highlighted events, such as "corner kick" or "free kick."

In soccer games, the dominant color is quite stable in wide-angle shots due to the large range of grass field. In order to utilize this significant cue, we use multistream HMMs [29] in com-

TABLE III
MODEL AND SEMANTIC SPACE IN THE VOLLEYBALL APPLICATION

Layer	Sub-models	Semantics
4th	Connectors, (HMM)	Volleyball Video Sequence
3rd	Connectors, (HMM)	8 Pre-defined Volleyball Shot Categories
2nd	Connectors, (HMM)	14 Pre-defined Volleyball Events
1st	Detectors, (Single GM)	Hidden Semantics

pound connectors. A multistream HMM is obtained by combining the multiple single-stream HMMs and introducing the weights for each stream. The first stream is still the motion feature vector same as in the basketball application, and the second stream is the vector of the mean RGB values of each frame. The color feature is only the secondary information and unreliable in other kind of shots, thus we use a lower weight on the color feature stream. The topology of each single-stream HMM follows the definition of basketball application.

In soccer shot classification, we perform the optimization at sequence level. Therefore, adjacent shots are no longer independent in recognition. In effect, the recognition results given by this model are composed of two parts, viz. the sequence of shot categories and the sequences of coarse events for each shot. Since what are concerned with is only shot classification, the sequences of coarse events are discarded even though they may provide more detailed information than shot categories. In our implementation, the shot boundaries are labeled manually, in order to suppress negative effects of inaccurate shot boundaries. Therefore, the recognition process is divided into two steps, say, shot level recognition and sequence level recognition. These two steps are interactively and tightly connected by probabilities.

3) Volleyball Sequence Analysis: Volleyball analysis is another application example of the proposed framework. Volleyball videos are more predictable in structure, which motivates the idea of sequence analysis. That is, an integrated approach is designed for both event detection and shot classification in a uniform manner. The semantic space of this application and corresponding model space are given in Table III. From the implementation point of view, this application is similar to the soccer shot classification presented in Section IV-B2 excepting the features used. Therefore, the detailed implementations of those parts are omitted.

Basically, we divide volleyball shots into eight categories, that is: 1) serve; 2) wide-angle (after serve); 3) wide-angle (after zoom-in); 4) zoom-in; 5) close-up; 6) wipe (before replay); 7) wipe (after replay); and 8) replay. However, in order to distinguish more detailed events, we attach the information of serving team to each shot category, so the number of shot categories is doubled. In this manner, more precise events may be recognized, which is impossible for one individual shot analysis. So, the predefined events in volleyball are divided into 14 categories. Some events are further extended to a number of events by attaching high level semantic attributes, viz. error and scoring. The 14 event categories include: 1) serve at right or left court; 2) offence at right or left court; 3) attack to right or left court; 4) stroke at right or left court; 5) block at right or left court; 6) zoom-in when playing; 7) jump serve/serve in close-up view by right or left team; 8) cheer in right or left court; 9) track player to right or left; 10) general close-up; 11) stillness; 12) wipe; 13) slow motion of stroking in right or left court; and 14) slow motion in close-up view. The semantics predefined in this application is extraordinary comprehensive and describe the content of volleyball videos precisely. Without regarding to computational complexity, we can build an analysis system in a broad domain by integrating a number of models like this one.

V. EXPERIMENTS

Before experimental results and discussions are given, we first present the evaluation of the proposed motion features. All three applications are based on this motion features, hence the feature performance is critical to any further applications.

A. Feature Effectiveness Study

In Section III-A, we proposed a novel representation of dominant motions which are derived from MVFs. However, different block matching algorithms produce different MVFs. Although the full-search strategy is employed to obtain MVFs in basketball application, it is extremely time-consuming. Therefore, our goal is to find a balance point between computational complexity and recognition performance in the following experiments. We use basketball application to characterize the changes of recognition performance when different sources of MVFs are employed.

Three MVFs obtained by three different ways are studied in this experiment. The first one is full-search strategy, which emphasizes accuracy and disregards speed. The second is the diamond-search strategy [30], which pays attention to both accuracy and speed, and the last one is the MVFs directly extracted from MPEG-I streams, specifically, P-frame motion vectors. Though a variety of block matching algorithms may be used in the MPEG-I coding systems, all of them are highly accelerated for real-time compression with a compromise on accuracy. Usually, there are three types of frames in MPEG-I streams, namely, I-frames, P-frames, and B-frames. As P-frames are not continuous and the interval is variable, we fill the "blanks" between P-frames to make the P-frame MVFs usable for our experiment. In normal MPEG-I streams, frame sequences always begin with an I-frame and end with a P-frame, which means each "blank" sequence is followed by a P-frame. Supposing there are n continuous "blank" frames preceded by a P-frame, the MVFs of the $n+1$ frames, including n "blank" frames and a posterior P-frame, are equal to $\text{MVF}_P/(n+1)$, where MVF_P denotes the MVF of the posterior P-frame.

Total test video sequences have been composed of six sessions of basketball games, about 15 min each. Two sessions were used for training and the remains for testing. All sequences were segmented into shots and further labeled with predefined events. Event transcriptions of test sequences were used as ground truth. According to the three aforementioned MVFs, three different recognition results were obtained based on the same training and test data. As event transcriptions are given as final recognition results, it is not reasonable to compare them with ground-truth by orderly one by one matching. We need to deal with the cases of the event insertion, deletion and replacement. A dynamic-programming (DP) approach, similar

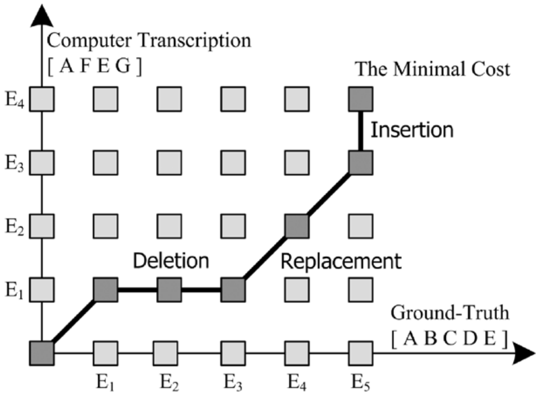


Fig. 8. DP approach for event transcription comparison.

TABLE IV
RECOGNITION PERFORMANCE USING DIFFERENT MVFS

Clip	Full-search		Diamond-Search		MPEG-I Stream	
	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
I	70.65%	19.74	73.17%	20.59	58.41%	23.51
II	76.59%	19.63	73.36%	19.68	56.62%	26.08
III	71.21%	19.55	74.24%	20.10	60.20%	25.77
IV	77.82%	19.28	75.16%	19.67	56.11%	24.69

to the measure of speech recognition [31], is employed to assess recognition results obtained from three MVFs, respectively, as shown in Fig. 8. Because the edges of some predefined events may be unclear sometime, the ground truth is not deemed as the exclusive explanation for the given video sequences. Considering this fact, some event pairs, including "lay-up" and "shot," "close-up" and "tracking," have been combined by setting the replacement costs to zero. In other cases, the costs of insertion, deletion, and replacement are all equal to one.

Based on the DP matching scheme, we have defined two measures for result evaluation, that is, the correctness and the accuracy, which score the performance of event detection and segmentation respectively. They are defined as follows:

$$R_{\text{cor}} = \frac{C_{\max} - C_{\min}}{C_{\max}} \times 100\% \tag{11}$$

$$R_{\text{acc}} = \frac{1}{2|\Omega|} \sum_{x \in \Omega} \left| cB^x_{\text{begin}} - gB^x_{\text{begin}} \right| + \left| cB^x_{\text{end}} - gB^x_{\text{end}} \right| \tag{12}$$

where $C_{\min}$ and $C_{\max}$ denote the minimal and maximal cost of DP matching, respectively, Ω is the set of matched events, cB_{begin^x} and cB^x_{end} are the two boundaries of event x in recognition results, and gB^x_{begin} and gB^x_{end} are in ground truth. The experimental results are listed in Table IV.

Because the motion vectors in MPEG-I streams are not only inaccurate but also incomplete, the scores of correctness and accuracy are the lowest when MPEG MVFs are employed. However, even in this case, the highest score of correctness approaches to 60% and the accuracy score also shows that the average error of segmentation is no more than one second. The full-search- and diamond-search-based approaches all obtain higher scores compared to that of the MPEG MVF-based approach. The correctness is evidently increased, although the improvement on accuracy is not distinct. This experiment has verified that the proposed features are robust to various MVFs in terms of segmentation. The correctness of diamond-search is not higher than that of full search, but more stable. Therefore, the diamond-search algorithm, designed based on the matching probability distribution, is likely to produce true motion vectors under noises. In conclusion, the diamond-search algorithm is the best solution to the proposed motion features at aspects of both performance and computational complexity.

TABLE V
EVALUATION OF BASKETBALL EVENT DETECTION

Shot Num	Assessment				Score
	Good	Mod.	Bad	Refu.	
85	576	156	112	33	77.5%
441	3385	918	685	102	77.1%
532	3822	1093	769	97	76.9%
699	3201	885	834	50	74.1%
470	2990	758	919	30	72.2%
359	1500	321	457	9	72.9%
2586	**15474**	**4131**	**3776**	**321**	**75.0%**

B. *Evaluation of Basketball Event Detection*

The total duration of our experimental basketball videos is more than 6 h and over 2500 shots, including MPEG-7 test videos. About 20 sample clips have been used to train each compound connector and 380 event transcriptions to build the top-layer connector.

Considering the ambiguity of semantics, we have carried out evaluation experiments by user study. A web-based evaluation system has been designed for user assessment from the website. The subjects were invited and required to score computer-generated results online. In our evaluation system, the event information was extracted from the database and presented to users by speeches and texts simultaneously when an event was emerging. At each end of shots, the subjects were required to select an assessment: *good*, *neutral*, or *bad*. If a subject thinks the current shot cannot be described by our predefined events, he/she is allowed to refuse the assignment. We define the rate of users' satisfaction as follows:

$$SatRate = \frac{N_{\text{Good}} \cdot 100 + N_{\text{Neutral}} \cdot 50}{N_{\text{Good}} + N_{\text{Neutral}} + N_{\text{Bad}}} \cdot 100\% \tag{13}$$

where N_{Good} is the number of the "*good*" assessments, N_{Neutral} is the number of the "*neutral*" assessments, and N_{Bad} is the number of the "*bad*" assessments. In total, 15 subjects were involved in this experiment. The results of user study are listed in Table V, which indicates that the predefined events are reasonably complete, as only 1.35% assignments were refused by subjects. The average user satisfaction rate approaches 75%, which indicates that our framework is effective for basketball event detection.

C. *Evaluation of Soccer Shot Classification*

Three soccer matches have been used in this evaluation, in which one match (about 76 min) was for training and the other two matches (about 3 h) for testing. Based on the conclusion of the feature study, we adopted the diamond-search approach

TABLE VI
EVALUATION OF SOCCER SHOT CLASSIFICATION

Shot Category	Precision Rate	Recall Rate
Wide-angle	90.2%	88.6%
Zoom-in	85.2%	89.2%
Close- up	86.3%	90.8%
Wipe	99.0%	89.2%
Replay	94.8%	79.8%

to extract motion vectors. In this manner, the speed of feature extraction is tripled approximately. As the shots in soccer videos change frequently, the training data for subsemantics and shot categories are adequate. We have obtained more than 20 sample clips to train each subsemantic model and over 700 subsemantic transcriptions of shots to build connectors at upper layers.

In this application, subsemantics are only defined as mid-level steps between low-level features analysis and high-level shot categorization. Though subsemantic transcriptions are also given in recognition results, we do not take them as a part of the result evaluation. Because shot boundaries are determined before recognition, we need not concern the accuracy of segmentation. Meanwhile, the semantic granularity of shot categories is relative high, so the divergences on shot classification may not be distinct. We manually labeled each shot in test sequences as ground truth and used precision rate and recall rate to evaluate recognition results. The results are given in Table VI.

From this table, we can see that both the precision rate and recall rate are satisfactory. As motion patterns in wipes are most distinguishable, the highest score has been obtained in this category. Wide-angle, zoom-in, and close-up, which are helpful for the determination of play/break status in games, are also well classified. In conclusion, the experimental results have demonstrated the good capability of the proposed framework in shot classification.

D. *Evaluation of Volleyball Sequence Analysis*

In volleyball application, two sessions (about 40 min) are used for training, and seven sessions (about 110 min) for testing. Because the occurrences of different events are extremely uneven, we cannot obtain adequate samples for some events though 40-min videos are employed. However, most of the compound connectors are well trained by at least 10 samples. More than 400 event transcriptions of shots are utilized to build connectors at upper layers.

The recognition results are composed of two parts, e.g., subsemantic transcriptions for event detection/segmentation and shot category transcriptions for shot classification. We evaluate the two parts of recognition results, respectively, with the similar schemes in basketball and soccer applications. In the evaluation experiment for event transcriptions, 10 subjects were invited for assessment. The experimental results of shot classification and event detection are listed in Tables VII and VIII, respectively.

Comparing Table VII with Table VI, it is evident that better performance on shot classification is obtained in this application. However, the recall rate of zoom-in is only 58.8%, which is much lower than that of the other categories. Unlike in soccer videos, zoom-in shots are rare in volleyball videos, which may result in insufficient training of the related submodels. As aforementioned, each shot category also preserves which is the serving team. We use this information to predict the scores of match teams. The predict error is no more than two points for each team in our experiment, which shows that the serving team information can also be effectively extracted.

TABLE VII
EVALUATION OF VOLLEYBALL SHOT CLASSIFICATION

Shot Category	Precision Rate	Recall Rate
Wide-angle	99.3%	97.6%
Zoom-in	100%	58.8%
Close-up	100%	85.7%
Serve	87.6%	99.4%
Wipe	100%	94.1%
Replay	100%	94.1%

TABLE VIII
EVALUATION OF VOLLEYBALL EVENT DETECTION

Shot Num	Assessment				Score
	Good	Mod.	Bad	Refu.	
203	1074	464	275	30	72.0%
218	1312	484	276	4	75.0%
107	727	209	125	2	78.4%
157	1138	286	138	17	82.0%
260	1682	494	409	6	74.6%
217	1487	403	273	4	78.1%
135	889	241	187	1	76.7%
1297	**8309**	**2581**	**1683**	**64**	**76.4%**

In Table VIII, only 0.5% of assignments are refused by subjects, which is much fewer than that in basketball analysis. It is because more comprehensive events can be detected by the integrated model. The average user satisfaction exceeds 76%, which is also higher than that in basketball. Usually, the finer event category results in lower satisfaction rate due to more recognition errors. However, a higher score has been obtained though the predefined events in volleyball are finer than those in basketball. In general, we can conclude that semantics at different granularities may be more efficiently modeled by the proposed framework than those dealing with them separately. The reason is that the recognition results given by the integrated model are optimized in a broader model space.

VI. CONCLUSION

In this paper, we have presented a novel HMM-based framework as a generic solution to video semantic analysis. There are three advantages in the proposed framework compared to existing work: 1) it is a uniform solution to video semantic analysis; 2) an efficient and simple representation is proposed to sufficiently utilize context constraints in video sequences; and 3) it is an integrated model for semantic recognition and segmentation in multiply semantic granularity. Specifically, in the proposed framework, the semantics at different granularities are mapped to a hierarchical model space which is composed of detectors at bottom layer and connectors at upper layers. Detectors are application-dependent models which convert low-level features to weak hypotheses of semantics. On the other hand, connectors, a kind of HMM, are universal models, which optimize those hypotheses from detectors or connectors at lower layer according to context constraints. In this manner, the proposed model decomposes a complex issue into simple subissues represented by detectors or connectors when training and automatically integrates those submodels for recognition. The applications to basketball event detection, soccer shot classification and volleyball sequence analysis have demonstrated that the proposed framework is not only suitable for a broad range of applications, but also capable of handling semantics at different granularities. Another contribution of this paper is the robust temporal motion representation scheme. The evaluation experiments have validated the effectiveness of this motion representation. Moreover, the robustness of this representation is also testified by the performance study.

The proposed framework is open and extendable. Rather than the single Gaussian model, various generative models may be used as detectors, such as mixture Gaussian Model and Bayesian networks, for better performance in specific application domain. However, weak detectors are still expected in terms of the performance of whole model.

References

[1] F. Liu and R. W. Picard, "Finding periodicity in space and time," in *Proc. IEEE Int. Conf. Computer Vision*, 1998, pp. 376–383.

[2] R. Polana and R. Nelson, "Detecting activities," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 1993, pp. 2–7.

[3] L. Zelnik-Manor and M. Irani, "Vent-based analysis of video," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2001, pp. 123–130.

[4] M. J. Roach, J. D. Mason, and M. Pawlewski, "Video genre classification using dynamics," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2001, pp. 1557–1560.

[5] B. T. Truong and C. Dorai, "Automatic genre identification for content-based video categorization," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2000, pp. 230–233.

[6] K. Messer, W. Christmas, and J. Kittler, "Automatic sports classification," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2002, pp. 1005–1008.

[7] J. Liu and B. Bhanu, "Learning semantic visual concepts from video," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2002, pp. 1061–1064.

[8] N. Vasconcelos and A. Lippman, "A Bayesian framework for semantic content characterization," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, 1998, pp. 566–571.

[9] M. R. Naphade, M. R. Naphade, S. Basu, J. R. Smith, C.-Y. Lin, and B. Tseng, "A statistical modeling approach to content based video retrieval," in *Proc. IEEE. Int. Conf. Pattern Recognition*, 2002, pp. 953–956.

[10] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems," in *Proc. IEEE Int. Conf. Image Processing*, 1998, pp. 536–540.

[11] M. R. Naphade and T. S. Huang, "Semantic video indexing using a probabilistic framework," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2000, pp. 79–84.

[12] H. J. Zhang, Y. Gong, S. W. Smoliar, and S. Y. Tan, "Automatic parsing of news video," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1994, pp. 45–54.

[13] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796–807, Jul. 2003.

[14] Y.-P. Tan, D. D. Saur, S. R. Kulkarni, and P. J. Ramadge, "Rapid estimation of camera motion from compressed video with application to video annotation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 2, pp. 133–146, Feb. 2000.

[15] G. Sudhir, J. C. M. Lee, and A. K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," in *Proc. IEEE Int. Workshop Content-Based Access of Image and Video Database*, 1998, pp. 81–90.

[16] Y. Gong, T. S. Lim, and H. C. Chua, "Automatic parsing of tv soccer programs," in *Proc. IEEE Int. Conf. Multimedia Computing and Systems*, 1995, pp. 167–174.

[17] D. Zhong and S. F. Chang, "Structure analysis of sports video using domain models," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, 2001, pp. 713–716.

[18] T. Liu and J. R. Kender, "Rule-based video classification system for basketball video indexing," in *Proc. ACM Multimedia Workshops*, 2000, pp. 213–216.

[19] W. Zhou, A. Vellaikal, and C. C. J. Kuo, "A hidden Markov model approach to the structure of documentaries," in *Proc. IEEE Int. Workshop Content-Based Access of Image and Video Libraries*, 2000, pp. 111–115.

[20] J. M. Sanchez, X. Binefa, and J. R. Kender, "Coupled Markov chains for video contents characterization," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2002, pp. 461–464.

[21] M. Petkovic, W. Jonker, and Z. Zivkovic, "Recognizing strokes in tennis videos using hidden Markov models," in *Proc. IASTED Int. Conf. Visualization, Imaging and Image Processing*, 2001, pp. 512–516.

[22] L. Xie, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2002, pp. 4096–4099.

[23] C. L. Huang and C. Y. Chang, "Video summarization using hidden Markov model," in *Proc. IEEE Int. Conf. Information Technology: Coding and Computing*, 2001, pp. 473–477.

[24] W. Wolf, "Hidden Markov model parsing of video programs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1997, pp. 2609–2611.

[25] H. Pan, P. van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2001, pp. 1649–1652.

[26] J. C. Huang, Z. Liu, and Y. Wang, "Joint video scene segmentation and classification based on hidden Markov model," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, 2000, pp. 1551–1554.

[27] G. S. Chambers, S. Venkatesh, G. A. W. West, and H. H. Bui, "Hierarchical recognition of intentional human gestures for sports video annotation," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2002, pp. 1082–1085.

[28] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 256–286, Feb. 1989.

[29] H. Bourlard, S. Dupont, and C. Ris, "Multi-Stream Speech Recognition," IDIAP, IDIAP-RR 07, 1996.

[30] S. Zhu and K. K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. Image Process.*, vol. 9, no. 2, pp. 287–290, Feb. 2000.

[31] S. Young *et al.*. (2001) The HTK Book (for HTK Version 3.2), Online Manual. [Online]. Available: http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf

Gu Xu received the B.Sc. degree in biochemistry and molecular biology from Sichuan University, Chengdu, China, in 1999 and the M.Eng. degree in computer science from Tsinghua University, Beijing, China, in 2003.

He joined Microsoft Research Asia, Beijing, China, in 2003. Before 2004, he undertook research in video content analysis and computer vision. His recent research interests are in web-scale search and mining.

Yu-Fei Ma (M'00) received the B.S. degree from Harbin Engineering University, Harbin, China, in 1994 and the M.S. degree in computer science from Tsinghua University, Beijing, China, in 2000.

From 1994 to 1997, he was engaged in computer network system analysis. From 2000 to 2005, he was with Microsoft Research Asia, Beijing, China, where he worked in the areas of video content analysis, multimedia signal processing, and pattern recognition. In 2005, he joined the IBM China Research Laboratory, Beijing, as a Research Staff Member. His current interests are in the areas of mobile multimedia and pervasive computing. He has authored more than 30 publications in these areas.

Hong-Jiang Zhang (M'91–SM'97–F'03) received the B.S. degree from Zhengzhou University, Henan, China, in 1982 and the Ph.D. degree from the Technical University of Denmark, Lyngby, Denmark, in 1991, both in electrical engineering.

From 1992 to 1995, he was with the Institute of Systems Science, National University of Singapore, Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. From 1995 to 1999, he was a Research Manager with Hewlett-Packard Laboratories, Palo Alto, CA, where he was responsible for research and development in the areas of multimedia management and intelligent image processing. In 1999, he joined Microsoft Research Asia, Beijing, China, where he is currently the Managing Director of the Advanced Technology Center, Beijing. He has coauthored/coedited four books, over 300 referred papers and book chapters, eight special issues of international journals on image and video processing, content-based media retrieval, and computer vision, as well as over 50 patents or pending applications.

Dr. Zhang currently serves on the editorial boards of five IEEE/ACM journals and a dozen committees of international conferences and is the Editor-in-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA.

Shi-Qiang Yang (M'97) received the B.Eng. and M.Eng. degrees in computer science from Tsinghua University, Beijing, China, in 1977 and 1983, respectively.

He is currently a Professor, Ph.D. supervisor, and the Executive Head of the Department of Computer Science and Technology, Tsinghua University. He is the Co-Director of the Tsinghua-Microsoft Multimedia Joint Research Laboratory. His research interests include multimedia technology and systems, video compression and streaming, content-based retrieval for multimedia information, and pervasive computing. He has published approximately 80 papers.

Dr. Yang serves as the Program Co-Chair of the Workshop on ACM Multimedia 2005, Pacific-Rim Conference on Multimedia PCM 2006, and the Internation Multimedia Modeling Conference 2006.