Using Webcast Text for Semantic Event Detection in Broadcast Sports Video

Changsheng Xu, *Senior Member, IEEE*, Yi-Fan Zhang, Guangyu Zhu, Yong Rui, *Senior Member, IEEE*, Hanqing Lu, *Senior Member, IEEE*, and Qingming Huang, *Senior Member, IEEE*

***Abstract*—Sports video semantic event detection is essential for sports video summarization and retrieval. Extensive research efforts have been devoted to this area in recent years. However, the existing sports video event detection approaches heavily rely on either video content itself, which face the difficulty of high-level semantic information extraction from video content using computer vision and image processing techniques, or manually generated video ontology, which is domain specific and difficult to be automatically aligned with the video content. In this paper, we present a novel approach for sports video semantic event detection based on analysis and alignment of webcast text and broadcast video. Webcast text is a text broadcast channel for sports game which is co-produced with the broadcast video and is easily obtained from the web. We first analyze webcast text to cluster and detect text events in an unsupervised way using probabilistic latent semantic analysis (pLSA). Based on the detected text event and video structure analysis, we employ a conditional random field model (CRFM) to align text event and video event by detecting event moment and event boundary in the video. Incorporation of webcast text into sports video analysis significantly facilitates sports video semantic event detection. We conducted experiments on 33 hours of soccer and basketball games for webcast analysis, broadcast video analysis and text/video semantic alignment. The results are encouraging and compared with the manually labeled ground truth.**

***Index Terms*—Broadcast video, semantic event detection, Webcast text.**

I. Introduction

THE explosive proliferation of multimedia content made available on broadcast and Internet leads to the increasing need for its ubiquitous access at anytime and anywhere, and over a variety of receiving devices. It is clear that when accessing lengthy, voluminous video programs, the ability to access highlights and to skip the less interesting parts of the videos will save not only the viewer's time, but also data downloading/airtime costs if the viewer receives videos wirelessly from remote servers. Moreover, it would be very attractive if users can access and view the content based on their own preferences. To realize above needs, the source video has to be tagged with semantic labels. These labels must not only be broad to cover the general events in the video, e.g., goals scored, near-misses, fouls in soccer, it has to be also sufficiently deep to cover the semantics of the events, e.g., names of the players involved. This is a very challenging task and would require to exploit multimodal and multicontext approaches.

The ever increasing amount of video archives makes manual annotation extremely time-consuming and expensive. Since humans tend to use high-level semantic concepts when querying and browsing video database, there is an increasing need for automatic video semantic annotation. Video annotation is also able to facilitate video semantic analysis such as video summarization and personalized video retrieval. Therefore, automatic video annotation has become a hot research area in recent years and is an important task in TRECVID benchmark [1]. Various approaches have been proposed for concept detection and semantic annotation of news video [2] and home video [3].

In this paper, we focus on semantic annotation and event detection of sports video which has wide view-ship and huge commercial potential. We present a novel approach for sports video semantic annotation and event detection using webcast text under a probabilistic framework. We use soccer video and basketball video as our test-bed. The proposed approach is generic and can be extended to other sports domains.

A. Related Work

Extensive research efforts have been devoted to sports video event detection in recent years. The existing approaches can be classified into event detection based on video content only and event detection based on external knowledge (ontology).

1) Event detection based on video content only: Most of the previous work of event detection in sports video is based on audio/visual/textual features directly extracted from video content itself. The basic idea of these approaches is to use low-level or mid-level audio/visual/textual features and rule-based or statistical learning algorithms to detect events in sports video. These approaches can be further classified into single-modality based approach and multimodality based approach. Single-modality based approaches only use single stream in sports video for event detection. For example, audio features were used for baseball event detection [4] and soccer event detection [5]; visual features were used for soccer event detection [6], [7]; and textual features (caption text overlaid on

Manuscript received July 16, 2007; revised May 24, 2008. Current version published November 17, 2008. The work of Y.-F. Zhang and H. Lu was supported by National Sciences Foundation of China under Grant 60475010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kiyoharu Aizawa.

C. Xu, Y.-F. Zhang, and H. Lu are with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China and also with the China-Singapore Institute of Digital Media, Singapore (e-mail: csxu@nlpr.ia.ac.cn; yfzhang@nlpr.ia.ac.cn; luhq@nlpr.ia.ac.cn).

G. Zhu is with the School of Computer Science, Harbin Institute of Technology, Harbin, 150001, China (e-mail: gyzhu@jdl.ac.cn).

Y. Rui is with the Microsoft China R&D Group, Beijing 100080, China (e-mail: yongrui@microsoft.com).

Q. Huang is with the Graduate University, Chinese Academy of Sciences, Beijing, China, 100039 (e-mail: qmhuang@jdl.ac.cn).

Digital Object Identifier 10.1109/TMM.2008.2004912

the video) were utilized for event detection in baseball [8] and soccer [9] videos. The single-modality based approaches have low computational load, but the accuracy of event detection is low as broadcast sports video is the integration of different multimodal information and only using single modality is not able to fully characterize the events in sports video. In order to improve the robustness of event detection, multimodality based approaches were utilized for sports video analysis. For example, audio/visual features were utilized for event detection in tennis [10], soccer [11] and basketball [12]; and audio/visual/textual features were utilized for event detection in baseball [13], basketball [14], and soccer [15]. Compared with the single-modality based approaches, multimodality based approaches are able to obtain better event detection accuracy but have high computational cost.

Both single-modality based approaches and multimodality based approaches are heavily replying on audio/visual/textual features directly extracted from the video itself. Due to the semantic gap between low-level features and high-level events as well as dynamic structures of different sports games, it is difficult to use these approaches to address following challenges: 1) ideal event detection accuracy $(\sim 100\%)$; 2) extraction of the event semantics, e.g., who scores the goal and how the goal is scored for a "goal" event in soccer video; 3) detection of the exact event boundaries; 4) generation of personalized summary based on certain event, player or team; 5) a generic event detection framework for different sports games; and 6) robust performance with the increase of the test dataset. In order to address these challenges, we have to seek available external knowledge for help.

2) Event detection based on external knowledge: Event detection based on video content only faces the difficulty of high-level semantic information extraction from broadcast video content using computer vision and image processing techniques. Therefore, in order to deal with this problem, ontology-based approaches are proposed [16]–[21] for video annotation and event detection. Ontology is a formal and explicit specification of knowledge domain, which consists of concepts, concept properties, and relationship between concepts [21]. Domain specific linguistic ontology with multilingual lexicons and possibility of cross document merging for multimedia retrieval was presented in [16]. A hierarchy of ontology [17] was defined to represent the results of video segmentation. Concepts were represented in keywords and mapped into an object ontology, a shot ontology and a semantic ontology. The possibility of extending linguistic ontology to multimedia ontology was proposed [18] to support video understanding. Multimedia ontology was constructed by extracting text information and visual features in the videos and manually assigning the extracted information to concepts, properties, or relationships in the ontology [19]. A Visual Descriptors Ontology and a Multimedia Structure Ontology based on MPEG-7 Visual Descriptors were used together with domain ontology to support content annotation [20]. An improvement of the approach in [20] was proposed and applied to soccer video annotation [21]. Visual features were extracted and used in the domain ontology and a clustering algorithm was employed to extend the domain knowledge through visual feature analysis.

42' Kaka (Milan) wins a free kick on the left wing.
42' Foul by Nicolas Burdisso (Inter Milan).
40' Goal! Inter Milan 0, Milan 1. Ronaldo (Milan) left footed shot from outside the box to the bottom left corner. Assisted by Ivan Gennaro Gattuso following a fast break.
39' Attempt missed. Luis Figo (Inter Milan) right footed shot from the left side of the box is close, but misses to the right. Assisted by Maxwell.
37' Offside, Milan. Kaka tries a through ball, but Ronaldo is caught offside.

(a)

11:41 Shaquille O'Neal makes dunk (Dwyane Wade assists) 2-0
11:26 Chauncey Billups misses 17-foot jumper 2-0
11:23 Antoine Walker defensive rebound 2-0
11:17 Jason Williams misses 16-foot jumper 2-0
11:15 Richard Hamilton defensive rebound 2-0
11:14 Antoine Walker shooting foul (Ben Wallace draws the foul) 2-0

(b)

Fig. 1. Webcast text.

B. Problem Formulation

The approaches of event detection based on video content only face a challenge called "semantic gap." To bridge this gap from low-level features to semantics, ontology based approaches are explored to utilize the external knowledge which has rich semantic information. However, most of the external knowledge used is built independently of the video content, thus it lacks of an automatic mechanism to connect external knowledge to video content at semantic level. In this paper, we propose to employ the external sources which are coproduced together with the video content during the sports game broadcasting and develop an automatic mechanism to semantically combine the video content and external sources for event detection. Currently, closed caption and webcast text are two external sources used for sports video semantic analysis.

Closed caption is a manually tagged text stream encoded into video signals. Since it is the content of the live commentary of the sports game, it is a useful information source for sports video semantic analysis and event detection [22]–[24]. However, using closed caption has to face several difficulties: 1) it lacks of a decent structure; 2) it contains redundant information of the games; and 3) currently it is only available for certain sports videos and in certain countries.

Webcast text is another external source which is coproduced together with broadcast sports video during the sports game broadcasting. It was first used together with match report to assist event detection in soccer video [25], [26], but the proposed event detection model is hard to extend in other sports domains. We employed webcast text from common sports websites to detect and identify semantic events in broadcast sports video and proposed a live soccer event detection system [27], [28]. However, the event detection from webcast text was based on predefined keywords which are domain specific and also hard to extend to other sports domains.

In this paper, we present a novel approach for sports video semantic event detection based on analysis and alignment of webcast text and broadcast video. Webcast text is text broadcasting

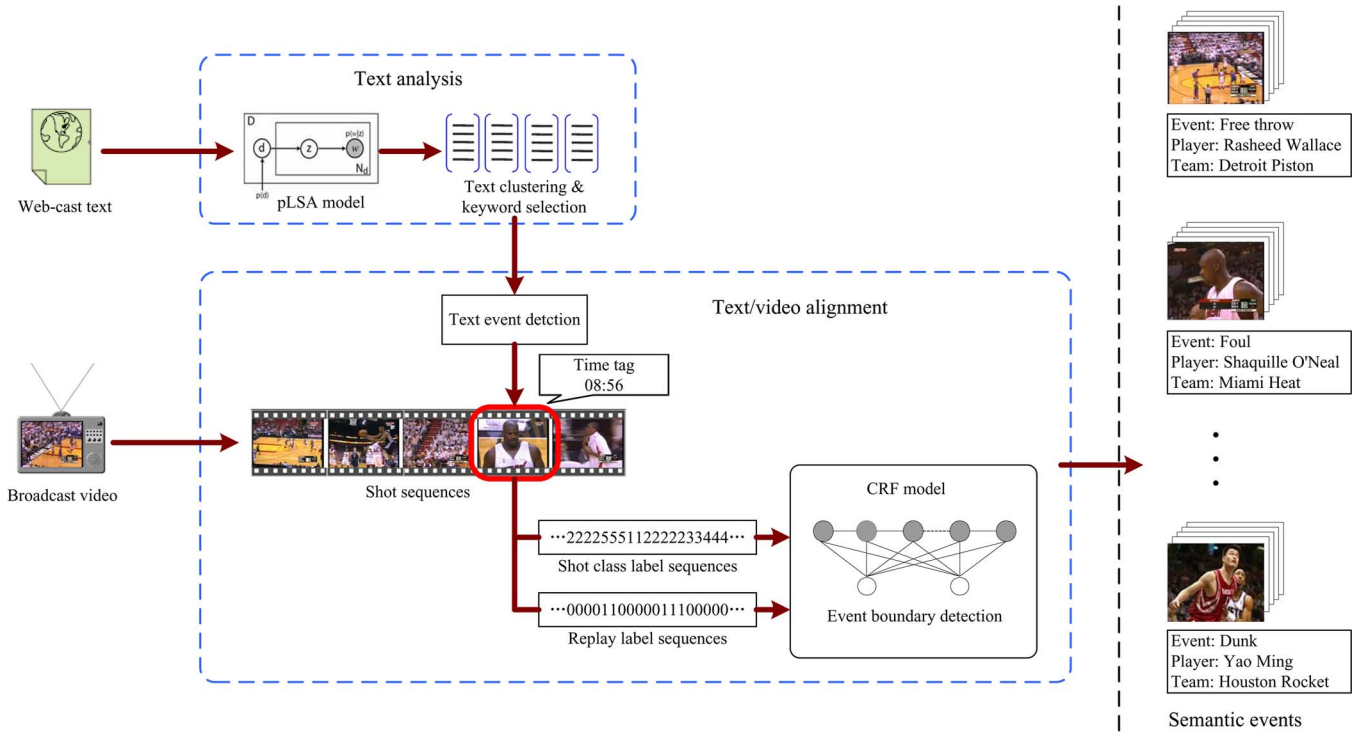

Fig. 2. Framework of proposed approach.

of live sports games and can be considered as a text counterpart of broadcast sports video. Webcast text contains rich semantics related to the event (see Fig. 1) and can be freely accessed from many sports websites [29], [30], [34]. We formulate the problem to be addressed as follows: given a webcast text and a broadcast video of the same sports game, we develop an automatic mechanism to align webcast text and broadcast video at both temporal and semantic levels to achieve semantic event detection in broadcast video.

Two research challenges need to be addressed for sports video semantic event detection: 1) automatic event detection from webcast text; and 2) automatic synchronization of detected text event to the broadcast video, namely, detection of event moment and event boundary in broadcast video corresponding to the text event. From machining learning point of view, the first challenge can be treated as unsupervised learning, while the second can be considered as supervised learning. Due to the different presentation styles of webcast text and different sports domains, it is difficult to use prior knowledge to detect event from webcast text in order to achieve a generic event detection framework. Therefore, we have to investigate unsupervised learning to address the first challenge. We will employ probabilistic latent semantic analysis (pLSA) to automatically cluster text event from webcast text and extract keywords (semantics) from the events in each cluster without using any prior knowledge (Section II). For different broadcast sports videos, there are some common features (e.g., shots, shot types, replay, etc.) and common production rules (especially for the events/highlights). These common features and production rules can be used as prior knowledge to model event structure in the video and help for event boundary detection in the video. To address the second challenge, we will apply conditional random field model (CRFM) to detect event with exact start/end boundary in broadcast video (Section III). The framework of proposed approach is illustrated in Fig. 2.

The rest of the paper is organized as follows. Event detection from webcast text using pLSA is described in Section II. The semantic alignment of webcast text and broadcast video using CRFM is presented in Section III. Experimental results are reported in Section IV. We conclude the paper with future work in Section V.

II. Web-Cast Text Analysis

The advantage of using text to help sports video analysis is to bridge the semantic gap between low-level features and high-level events. Researchers thus attempt to use text information, such as caption text overlaid on the video and closed caption, to assist sports video analysis. However, the result of recognizing captions overlaid on the video by OCR may be affected by video quality. Closed caption is a direct transcript from speech to text thus contains redundant information irrelevant to the games and lacks of well-defined text structure.

Webcast text (Fig. 1) is another text source. It is coproduced together with broadcast sports video during the sports game broadcasting and can be easily accessed in the web [29], [30] during or after the game. The content of webcast text is more focused on events of sports games and contains detailed information of events in sports games such as time of the event moment, the development of the event, players and team involved in the event, etc., which are very difficult to be automatically extracted from the broadcast video content directly. These advantages of webcast text motivate us to utilize webcast text to facilitate event semantics extraction and sports video annotation.

A. Text Event Detection

In webcast text, usually each sentence records an event. A simple way to detect event from webcast text is to use predefined keywords to match related words in the webcast text to detect event [27], [28]. However, there are several limitations for such an approach: 1) Since the event types are different for different sports games, we have to pre-define keywords for different sports games; 2) Even for the same sports game (e.g., soccer), due to different presentation style and language, it is difficult to use one predefined keyword to represent one event type; and 3) Since the description of an event is on a sentence level, single keyword is not enough for robust event detection and sometimes may cause incorrect detection. These limitations make the predefined keywords based approach less generic to different sports domains. In order to come up with a generic approach for various sports games, we need to explore an unsupervised approach without using any prior knowledge for webcast text analysis. Based on our observation, the descriptions of the same events in the webcast text have similar sentence structure and word usage. We can therefore employ an unsupervised approach to first cluster the descriptions into different groups corresponding to certain events and then to extract keywords from the descriptions in each group for event detection. Here we apply probabilistic latent semantic analysis (pLSA) for text event clustering and detection.

pLSA is a generative model to discover topics in a document using the bag-of-words document representation [31]. It is derived and improved from standard Latent semantic analysis (LSA) [32]. LSA is a method widely used in natural language processing. LSA builds a term-document matrix to describe the occurrences of terms in different documents. The matrix is decomposed by singular value decomposition (SVD). The cosine distance between vectors in the matrix is computed to measure the document similarity in the reduced dimensional space. However, LSA stems from linear algebra and is based on a L_2-optimal approximation which corresponds to an implicit additive Gaussian noise assumption on counts. In addition, the representation of texts obtained by LSA is unable to well handle polysemy (i.e., a word may have multiple senses and multiple types of usage in different context). The coordinates of a word in the latent space is written as a linear superposition of the coordinates of the documents that contain the word, but the superposition principle is unable to explicitly capture multiple senses of a word. Compared with LSA, pLSA is based on a mixture decomposition derived from a latent class model, which results in a solid statistic foundation. It is able to apply statistical techniques for model fitting, model selection and complexity control. More specifically, pLSA defines a proper generative data model and associates a latent context variable with each word occurrence, which explicitly accounts for polysemy. Therefore, pLSA is appropriate here for clustering webcast text from different presentation styles.

pLSA applies a latent variable model called aspect model for co-occurrence data which associates an unobserved class variable $z \in Z = z_1, z_2, \ldots, z_K$ with each observation. In our case, the different types of semantic events (such as goal in soccer and dunk in basketball) in the webcast text are considered as the latent topic corresponding to the variable z. In webcast text, each sentence description of an event is considered as a document and all the documents can be represented as $D = d_1, d_2, \ldots d_N$ with words from a vocabulary $W = w_1, w_2, \ldots w_M$. Based on this representation, webcast text can be summarized in a $N \times M$ co-occurrence table of counts $N_{ij} = n(d_i, w_j)$, where $n(d_i, w_j)$ denotes how often the word w_j occurred in a document d_i. A joint probability model $p(w, d)$ over $N \times M$ is defined by the mixture:

$$p(w, d) = \sum_{z \in Z} p(z)p(d|z)p(w|z) \tag{1}$$

where $p(w|z)$ and $p(d|z)$ are the class-conditional probabilities of a specific word/document conditioned on the unobserved class variable z, respectively.

The Expectation Maximization (EM) algorithm is applied to estimate the values of $p(w|z)$ and $p(d|z)$ in latent variable models. The EM algorithm consists of two steps:

1) Expectation step to compute the following posterior probability:

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{i=1}^{K} P(z_i)P(d|z_i)P(w|z_i)}. \tag{2}$$

2) Maximization step to estimate following terms:

$$P(w|z) \propto \sum_{i=1}^{N} n(d_i, w)P(z|d_i, w) \tag{3}$$

$$P(d|z) \propto \sum_{i=1}^{M} n(d, w_i)P(z|d, w_i) \tag{4}$$

$$P(z) \propto \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j)P(z|d_i, w_j). \tag{5}$$

Before applying pLSA, the webcast text corpus is preprocessed by a standard stop-word list and a name entity recognition algorithm [33] to filter out the names of the players and teams due to the consideration of their affecting to the analysis result. For example, "Rasheed Wallace shooting foul (Shaquille O'Neal draws the foul)" is processed into "shooting foul draws foul". By applying pLSA, a $N \times K$ matrix of the class-conditional probability of each document is obtained:

$$\begin{bmatrix} P(d_1|z_1) & \bullet\bullet\bullet & P(d_1|z_K) \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ P(d_N|z_1) & \bullet\bullet\bullet & P(d_N|z_K) \end{bmatrix} \tag{6}$$

where N is the total number of documents contained in webcast text and K is the total number of topics corresponding to the event categories in webcast text. The row of the matrix represents the probability of each document given the latent event classes.

We determine the optimal number of event categories K and cluster the documents into these categories $(C_1, C_2, \ldots, C_K)$ in an unsupervised manner. Each document can be clustered

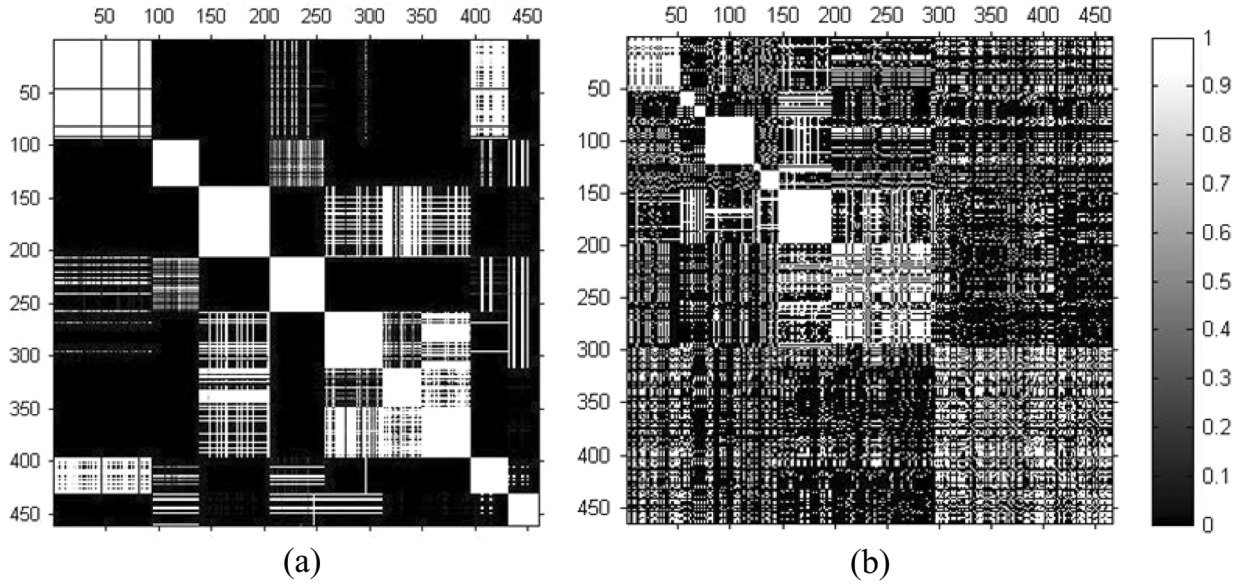

Fig. 3. Similarity matrix of webcast text by (a) pLSA. (b) LSA: The higher the intensity, the more similar the two documents.

into one category based on its maximum class-conditional probability:

$$C_k(d_i) = \arg\max_k \{P(d_i|z_k)\}_{k=1}^K \quad i = 1, \ldots, N \tag{7}$$

where $C_k(d_i)$ denotes that document d_i is clustered into the category C_k.

Each document can be represented by its class-conditional probability distribution on the K categories as a K dimensional feature vector $\vec{x}$. To measure the similarity between two documents in a K dimensional latent semantic space, a similarity function is defined as follows:

$$S(\vec{x}_i, \vec{x}_j) = \frac{1}{K} \sum_{k=1}^{K} \left\| x_i^k - x_j^k \right\|^2 \tag{8}$$

where x_i^k and x_j^k are the kth component of $\vec{x}_i$ and $\vec{x}_j$ respectively.

A clustering algorithm should achieve the compactness of intra-cluster and isolation of inter-cluster and consider the comprehensive effects of several factors including the defined squared error, the geometric or statistical properties of the data, number of the input data, the dissimilarity (or similarity), and the number of clusters [42]. With the context of document clustering in our work, the documents within the same cluster should have maximum similarity while the documents in the different clusters should have minimum similarity. Based on this criterion and (8), we define a cost function J_K to determine the optimal number of event categories.

$$J_K = \frac{\sum_{i=1}^{K} \sum_{\vec{x} \in C_i} S(\vec{x}, \vec{m}_i)}{\frac{1}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} S(\vec{m}_i, \vec{m}_j)} \tag{9}$$

$$\vec{m}_i = \frac{1}{n_i} \sum_{\vec{x} \in C_i} \vec{x} \tag{10}$$

where n_i is the number of documents in category C_i.

In practice, we define the candidate set of the number of clusters as $C = \{2, \ldots, K\}$ where $K < N$ and N is the number of all the documents used for clustering. Based on the selection criterion for the parameter of the number of clusters [42], the optimal number of event categories K^* can be determined by minimizing the value of J_K.

$$K^* = \arg\min_k \{J_k\}_{k=1}^K. \tag{11}$$

In sports games, an event means an occurrence which is significant and essential to the course of the game progress, hence the number of event categories recorded in the webcast text is normally less than 15. In the experiment the candidate set of number of clusters was initialized as $C = \{2, \ldots, 15\}$, that is, the range of K is between 2 and 15.

Fig. 3 illustrates the similarity matrix of the clustered soccer webcast text collected from Yahoo MatchCast Central [34], which contains 460 documents. The documents are sorted by their category number as follows:

$$\underbrace{\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_i}_{\mathrm{cat}1}, \underbrace{\mathrm{d}_{i+1}, \ldots, \mathrm{d}_j}_{\mathrm{cat}2}, \ldots, \underbrace{\mathrm{d}_n, \ldots, \mathrm{d}_N}_{\mathrm{cat}K}.$$

The x and y axis indicate the document number arranged in above order. The similarity between every two documents is measured by the similarity function (8) between the feature vectors in the latent semantic space. To make a comparison with LSA, the similarity matrix of the clustered result by LSA using same data is also shown. By using pLSA, we can see that the documents within the same clusters have the higher similarity while the documents in the different clusters have the lower similarity. In contrast, the result by LSA does not exhibit the high similarity of the documents in the same clusters.

After all the documents have been clustered into related categories, the words of the documents in each category are ranked

Fig. 4. Examples of shot class.

by their class-conditional probability $P(w_j|z_k)$. The top ranked word in each category is selected as the keyword to represent the event type. After the keywords for all the categories are identified, the text events can be detected by finding the sentences which contains the keywords and analyzing context information in the description. Here the reason we need to apply context information analysis is because using keywords alone for event detection sometimes cannot guarantee the occurrence of the related events, which may cause false alarms. For example, "kick" is the top ranked word for "free kick" event in soccer game. If its previous word is "free," the event is correctly detected; while if its previous word is "goal," a false alarm is caused (goal kick is a different event from free kick). Without using context information, the precision of "free kick" event detection will be reduced from 98.5% to 89.7% in our test dataset. In basketball game, the word "jumper" corresponds to the jumper event if its previous word is "makes" (e.g., Kobe Bryant makes 20-foot jumper); while it will cause a false alarm if its previous word is "misses" (e.g., Marcus Camby misses 15-foot jumper). Without using context information, the precision of "jumper" event detection will be reduced from 89.3% to 51.7% in our test dataset. Therefore, we need to search both keyword and its adjacent words to ensure the event detection accuracy.

The keyword is used to identify event category. For all the events in each event category, we save the time tags of the events from the beginning of the event sentence, then extract players and teams' names using a name entity recognition algorithm [32], which can achieve 95.3% precision and 98.8% recall for the detected events in our test dataset. The extracted semantic information (event category, player, team, time-tag) is used to build a data structure to annotate the video event and facilitate the searching using metadata. The original descriptions are attached to the video events as log files.

III. Text/Video Alignment

In order to semantically annotate broadcast video using webcast text, we need to synchronize the webcast text event to the event occurred in the video, that is, we need to detect a start boundary and an end boundary in the video that cover the event clip corresponding to the webcast text event. From a webcast text event, we can obtain a time tag which indicates when the event occurs in the game. According to the time tag, we can detect event moment in the video. Then based on the event moment, we detect event boundary using CRFM. The features used for CRFM are related to broadcasting production rules which are generic to different sports videos. The details of feature extraction, event moment detection and event boundary detection are described in the following subsections.

A. Feature Extraction

Broadcast sports video is a post-edited video containing video seeds from multiple cameras and editing effects such as replays. Hence, features related to broadcast production rules can be employed for video analysis. In our approach, we use shot class and replay to model the event structure of the broadcast sports video as these features are domain independent and robust to different sports videos. Different broadcast sports videos usually exhibit similar structure for an event due to the common production rule in sports video broadcasting. For example, when an event occurs, the broadcast director generates the related broadcast video segment by smoothly concatenating the video feeds from different cameras and inserting the replays.

1) Shot classification: Shot is a basic unit in broadcast videos which is generated by camera changing and switching. In broadcast sports video, shot transition patterns may reflect the occurrence of events, hence understanding shot class is important for sports video analysis. Before shot classification, we first detect shot boundary using our previous approach [27]. Based on the video production rules [35], the shot can be classified into three classes: far view shot, medium view shot, and close-up shot. These three basic shot classes relate to three different degrees of concentration and narrow an audience's attention from the overall scene, through a group of people, to the single individual. This rule also applies to broadcast sports video. In our application, we further classified medium view/close-up shot into in field medium view/close-up shot and out field medium view/close-up shot. Thus we have totally five shot classes. Fig. 4 illustrates examples of frames in far view shot, in/out field medium view shot and in/out field close-up shot in soccer video and basketball video respectively.

For shot classification, we first classify each frame within a shot into the five classes defined above. We extract the court region of each frame using color and filed lines. If the court region is lower than a threshold, the frame class is determined by the edge pixel number. Otherwise, the frame class is determined by the proportion of the maximal foreground object region against the court field region. After frame classification, a

simple weighted majority voting of frames within an identified single shot boundary is conducted to perform shot classification. More details can be referred to our previous work [36].

2) Replay detection: Replays are post-edited effects to highlight the events in broadcast sports video, hence they indicate the occurrences of the events in the video. Usually at the beginning and end of a replay, some special digital video effects (SDVE) are inserted, which can be used as cues to identify the replays. To obtain robust and distinguishable representation of the SDVE, we utilize a spatial-temporal mode seeking algorithm [37] to describe the features of SDVEs. Then we use a sliding window to perform similarity matching within a window over the whole video data to detect SDVEs. Compared with the previous work [38] which also tried to detect the special shot transitions and SDVEs, our replay detection approach employed the spatial-temporal feature modes which are obtained by a mean shift procedure and expected to capture the prominent color, motion and texture features which achieve the invariance properties from large amounts of frames. Hence, the comparison of two groups of frames is transformed into the distance measure between two sets of spatial-temporal feature modes. The Earth Mover's Distance (EMD) is applied to measure the video clip similarity in order to detect the SDVEs. Since our replay detection approach does not rely on any robust shot boundary detection of gradual transition, key frame extraction or complex logo tracking, it exhibits more flexibility and generic properties.

B. Event Moment Detection

Event moment is defined as a time point when the event exactly occurs. Since the text event contains the time tag (event moment) indicating when the event occurs, the intuitive way to detect the event moment in the video is to find the same time in the video. However, in broadcast sports video, the game time and video time are not synchronized due to nongame scenes such as player introduction, half-time break before the game starting and time-out during the game. Therefore, we need to recognize the game time in the video to detect event moment.

In many sports games such as soccer and basketball, a video clock is used to indicate the game lapsed time. We proposed an approach [27] to first detect a clock overlaid on the video and then recognize the digits from the clock to detect the game time. This approach worked well for soccer video due to the nonstopping clock time in soccer video. Once the game time is recognized at certain video frame, the time corresponding to other frames can be inferred based on recognized game time and frame rate. However, the clock in some broadcast sports video such as basketball video may stop or disappear at any time of the game, which makes the game time recognition more challenging than the video with nonstopping clock time. Due to unpredictable clock stopping, game time recognition in each frame is necessary in the video. As the clock region in the video may disappear during the clock stopping time, we employ a detection–verification–redetection mechanism, which is an improvement to our previous approach [27], to recognize game time in broadcast sports video.

We first segment the static overlaid region using a static region detection approach. Then the temporal neighboring pattern similarity measure is utilized to locate the clock digit position because the pixels of clock digit region are changing periodically. The technique detail of digit region location and digit number template capturing can be referred to [27]. Once we get the template and clock digit position, we recognize the digits of the TEN-MINUTE, MINUTE, TEN-SECOND, and SECOND of the clock. As the clock may disappear during the clock stopping time in some sports games (e.g., basketball), we need to verify the matching result of the SECOND digit using following formula:

$$M(i) = \max_i \left\{ \sum_{x,y \in R} D(x,y) \Theta T_i(x,y) \right\}, \quad i = 0, 1, \ldots, 9, 10 \tag{12}$$

where $T_i(x,y)$ is the image pixel value in position (x,y) for the template of digit number i, $D(x,y)$ is the image pixel value in position (x,y) for the digit number to be recognized, R is the region of digit number, Θ is EQV operator. $M(i)$ is calculated in the SECOND digit of the clock in each frame. If $M(i)$ is smaller than a matching threshold lasting for a certain time (2 seconds in our work), a failure of verification will occur and lead to the re-detection of clock digit. After the success of verification, the game time will be recorded as a time sequence and is linked to the frame number, which is called time-frame index. If the verification failure occurs during the clock stopping time, the related frames are also denoted as the label of "- - : - -" in the time-frame index.

C. Event Boundary Detection

Event moment can be detected based on the recognized game time. However, an event should be a video segment which exhibits the whole process of the event (e.g., how the event is developed, players involved in the event, reaction of the players to the event, etc.) rather than just a moment. Hence, we have to detect the start and end boundaries of an event in the video. Previous approaches on event detection only gave a coarse boundary for the event, while here we aim to detect event boundaries which can be comparable to the event manually generated by professionals. In this section, we present a novel approach to model the temporal transition patterns of a video event clip to recognize the boundaries. Certain temporal patterns recur during the event portion of the sports video due to the existence of common broadcast video production rules. By modeling the transition patterns, the duration of event can be identified from the video.

We use a conditional random field model (CRFM) to model the temporal event structure and detect the event boundary. We first train the CRFM using labeled video data to obtain the parameters for CRFM. The features described in Section III-A are used to train CRFM. After training, the CRFM can be used to detect event boundaries.

Conditional random field (CRF) is a probabilistic framework and is a discriminatively trained undirected graphical model of which hidden nodes have Markov property [39] conditioned on the observation. CRF brings together the merits of discriminative and generative models. The features are combined using exponential functions, which are trained discriminatively to maximize the conditional probability of the label given observation. But like generative models, the probabilities are propagated through neighboring nodes to obtain globally optimal labeling. The primary advantage of CRF over hidden Markov model (HMM) is its conditional nature, resulting in the relaxation of the independence assumptions required by HMM in order to ensure tractable inference. Additionally, CRF avoids

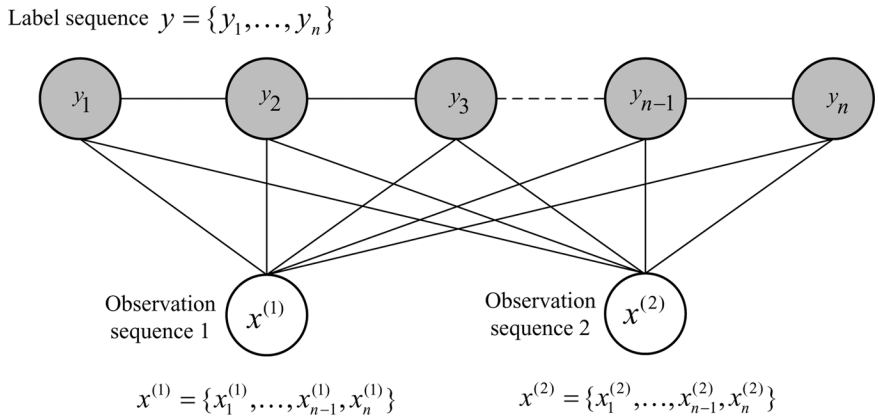

Fig. 5. CRF model for event boundary detection.

the label bias problem [39], a weakness exhibited by maximum entropy Markov model (MEMM) and other conditional Markov models based on directed graphical models. CRF outperforms both HMM and MEMM on a number of real-world sequence labeling tasks [39].

CRF model is shown in Fig. 5. In this model, the upper chain is a label shot sequence indicating whether the shot belongs to an event or not. We use two observation sequences. The left white node represents the replay shot sequence indicating whether the shot belongs to a replay or not and the right white node represents the shot view type sequence indicating the one of five shot classes. The reason we use two observation sequences is based on the consideration that the replay and shot class play an important role for event editing in the video broadcasting.

Based on the CRFM in Fig. 5, the conditional probability of label sequence $\mathbf{y} = \{y_1, \ldots, y_n\}$ given observation sequences $\mathbf{x}^{(1)} = \{x_1^{(1)}, \ldots, x_n^{(1)}\}$ and $\mathbf{x}^{(2)} = \{x_1^{(2)}, \ldots, x_n^{(2)}\}$ is defined as

$$p(\mathbf{y}|\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \Lambda) = \frac{1}{Z(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})} \times \exp\left(\sum_{j=1}^{J} \lambda_j F_j(\mathbf{y}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)})\right) \quad (13)$$

where $\Lambda = \{\lambda_1, \ldots, \lambda_J\}$ is the set of parameters of CRFM and J is equal to the number of feature functions, $Z(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ is a normalization factor which is defined as

$$Z\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = \sum_{\mathbf{y}} \exp\left(\sum_{j=1}^{J} \lambda_j F_j(\mathbf{y}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)})\right) \quad (14)$$

and $F_j(\mathbf{y}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ is the feature function which is defined as

$$F_j\left(\mathbf{y}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = \sum_{i=1}^{n} f_j\left(y_{i-1}, y_i, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, i\right) \quad (15)$$

where $f_j(y_{i-1}, y_i, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, i)$ is the feature used in CRFM, n is the length of label and observation sequences.

Given above CRFM, we need to provide training data to train CRFM to obtain the parameters of the model. The training process can be formulated as follows: given the training data set with N data $(X^{(1)}, X^{(2)}, Y) = \{(\mathbf{x}_1^{(1)}, \mathbf{x}_1^{(2)}, \mathbf{y}_1), (\mathbf{x}_2^{(1)}, \mathbf{x}_2^{(2)}, \mathbf{y}_2), \ldots, (\mathbf{x}_N^{(1)}, \mathbf{x}_N^{(2)}, \mathbf{y}_N)\}$, find the best model parameters such that the log likelihood $\log p(Y|X^{(1)}, X^{(2)}, \Lambda)$ is maximized. The partial derivative of the log likelihood function is computed as shown in (16):

$$\begin{aligned}
&\frac{\partial \log p(Y|X^{(1)}, X^{(2)}, \Lambda)}{\partial \lambda_j} \\
&= \sum_{k=1}^{N} F_j\left(\mathbf{y}_k, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) - \sum_{k=1}^{N} \frac{\partial \log Z\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right)}{\partial \lambda_j} \\
&= \sum_{k=1}^{N} F_j\left(\mathbf{y}_k, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) - \sum_{k=1}^{N} \frac{1}{Z\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right)} \\
&\quad \times \sum_{\mathbf{y}} \left(F_j\left(\mathbf{y}, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) \times \exp\left(\sum_{j=1}^{J} \lambda_j F_j\left(\mathbf{y}, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right)\right)\right) \\
&= \sum_{k=1}^{N} F_j\left(\mathbf{y}_k, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) - \sum_{k=1}^{N} \sum_{\mathbf{y}} \left(F_j\left(\mathbf{y}, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) \frac{1}{Z\left(\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right)} \right. \\
&\quad \left. \times \exp\left(\sum_{j=1}^{J} \lambda_j F_j\left(\mathbf{y}, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right)\right)\right) \\
&= \sum_{k=1}^{N} F_j\left(\mathbf{y}_k, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) - \sum_{k=1}^{N} \sum_{\mathbf{y}} \left(F_j\left(\mathbf{y}, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}\right) \right. \\
&\quad \left. \times p\left(\mathbf{y}|\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \Lambda\right)\right). \quad (16)
\end{aligned}$$

The first term in the final result of (16) is the expected value of $F_j(\mathbf{y}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ under the empirical distribution $\tilde{p}(\mathbf{y}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)})$, while the second item is the expectation of $F_j(\mathbf{y}, \mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)})$ under the model distribution $p(\mathbf{y}|\mathbf{x}_k^{(1)}, \mathbf{x}_k^{(2)}, \Lambda)$. We use L-BFGS quasi-Newton method [40] to train CRFM because it converges much faster than traditional iterative scaling algorithms [40].

Two sets of models are trained for soccer game and basketball game respectively. Each event category is trained using a CRFM. After the training, the CRFMs can be used for event boundary detection by finding the best label Y such that the log likelihood $\arg\max_Y p(Y|X^{(1)}, X^{(2)}, \Lambda)$ is maximized. The marginal probabilities of CRF can be inferred using dynamic programming in linear time.

During the detection process, the shot containing the detected event moment is used as a reference shot to obtain a search range for event boundary detection. The search range is empirically set to start from the first far view shot before the reference shot and end at the first far view shot after the reference shot. According to broadcast sports video production rule, the temporal transition patterns of an event would occur in this range. We then use the trained CRFM to calculate the probability scores of all the shots within the search range and label the shot with event or non-event shot based on the highest probability. The labeled shot sequence may contain some isolated points. For example, a given sequence $\mathrm{X} = \{x_1, x_2, \ldots, x_{20}\}$ is labeled as shown at the bottom of the page, where the label $+1$ denotes that the shot belongs to an event while -1 for nonevent. There are isolated points in this sequence which can be considered as noise and may affect event boundary detection. We apply a neighboring voting scheme to sequentially correct these points from the beginning of the sequence. Given the label sequence $L = \{l_1, \ldots, l_n\}$ where $l_i \in \{-1, +1\}, 1 \le i \le n, n$ is the length of the label sequence. Two steps are applied to identify and correct the noise points, respectively. For each l_i, we first use the following criteria to identify whether it is a noise point

$$l_{i-1} + l_i = 0 \text{ and } \left| \sum_{j=i}^{m} l_j \right| < m \tag{17}$$

where $m (m < n)$ is the length of the identified interval predefined on the sequence L. If l_i satisfies (17), it is identified as the noise point. Then a window based neighboring voting (NV) scheme is used to correct the value of the noise point l_i. For l_i, the NV value computed on the neighbors fallen in the window centered at l_i is defined as following:

$$\mathrm{NV}(l_i) = \sum_{j=-k}^{+k} l_{i+j} \tag{18}$$

where $[-k, k]$ is the width of the neighboring voting window. Therefore, the corrected value of l_i is determined as

$$\hat{l}_i = \begin{cases} +1, & \text{if } \mathrm{NV}(l_i) \ge 0 \\ -1, & \text{if } \mathrm{NV}(l_i) < 0. \end{cases} \tag{19}$$

If the calculated $\mathrm{NV}(l_i)$ is zero, which means the number of event label $(+1)$ and non-event label (-1) is equal in the neighboring voting window of l_i, to avoid the occurrence of marginal classification the event label $(+1)$ is assigned to l_i. This is reasonable due to the assumption that the probability of the event is larger than the probability of non-event in the defined sequence (search range). In the experiment, the values of m and k are both set to be 5 empirically.

IV. Experimental Results

We conducted our experiment on 20 hours of broadcast soccer videos including five World-Cup 2002 games, five Euro-Cup 2004 games, and five World-Cup 2006 games and 13 hours of broadcast basketball videos including five NBA 2005–2006 games and three Olympic Games 2004. We hope to use these diverse data sets to validate the robustness of the proposed approach to different sports games. These videos are captured by the Hauppauge WinTV-USB card. The correspondent webcast texts can be obtained from Yahoo (freestyle) and ESPN (professional) website for soccer and ESPN website for basketball. In our experiment, we used Yahoo webcast text for soccer event detection and ESPN for basketball event detection.

A. Text Analysis

We conducted keyword generation and text event detection experiment on all the soccer and basketball games. To make a comparison, we also tested LSA for keyword generation on the same data. For LSA, we represent the documents in reduced dimensional feature vectors and cluster them into different categories. Words in each category are ranked by their TF-IDF weights. The top ranked words are selected as the keywords. We first show the experimental results to automatically determine the number of event categories for basketball and soccer using pLSA and LSA. For LSA, the K-means approach is used for clustering and the optimal number of event categories is also determined by the cost function defined in (9). For each number of categories, we calculate a cost function value by (9). The number which has the smallest cost function value is selected as the optimal number of event categories. The results of the number of event categories and their corresponding cost function value for basketball and soccer using pLSA and LSA are plotted in Fig. 6.

From the results, for pLSA, the optimal number of event categories for basketball is 9 and 8 for soccer; for LSA, the optimal number of event categories for basketball is 9 and 10 for soccer. The optimal number of event categories will be used for the following experiments. The experimental results of keywords generation for basketball games using pLSA and LSA are listed in Tables I and II respectively, where the last column is the ground truth of event categories. If the number of events for one event type in the category is dominant (in our case, over 80% of all the events clustered in the category), this event type will be set as ground truth for the category.

From the results, we can see that there is no difference between pLSA and LSA if we select top ranked word in each event category from both tables as the keyword for the related

$$L = \{-1, -1, +1, +1, +1, +1, +1, -1, +1, +1, +1, +1, +1, -1, -1, -1, -1, +1, -1, -1\}$$

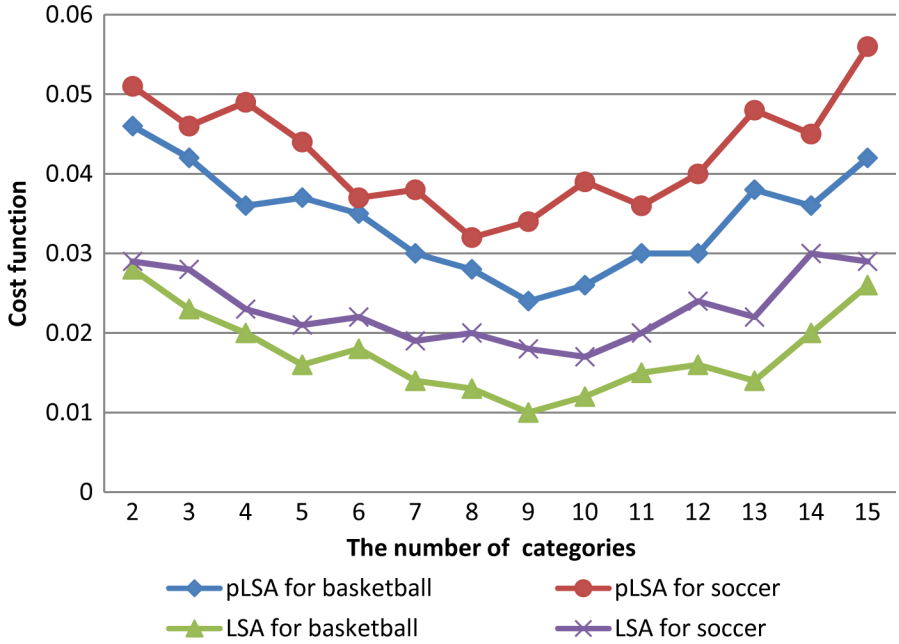

Fig. 6. Cost function of different number of event categories.

TABLE I
KEYWORDS GENERATED FROM BASKETBALL WEB-CAST TEXT BY PLSA

Category	Rank1	Rank2	Rank3	Event
1	shot	pass	bad	Shot
2	jumper	foot	misses	Jumper
3	layup	driving	blocks	Layup
4	dunk	makes	misses	Dunk
5	blocks	shot	assists	Block
6	rebound	defensive	offensive	Rebound
7	foul	draw	personal	Foul
8	throw	free	makes	Free throw
9	enters	game	timeout	Substitution

TABLE II
KEYWORDS GENERATED FROM BASKETBALL WEB-CAST TEXT BY LSA

Category	Rank1	Rank2	Rank3	Event
1	shot	point	foot	Shot
2	jumper	foot	makes	Jumper
3	layup	assists	makes	Layup
4	dunk	slam	assists	Dunk
5	blocks	jumper	shot	Block
6	rebound	assists	ball	Rebound
7	foul	draws	ball	Foul
8	throw	makes	misses	Free throw
9	enters	game	foul	Substitution

TABLE III
KEYWORDS GENERATED FROM SOCCER WEB-CAST TEXT BY PLSA

Category	Rank1	Rank2	Rank3	Event
1	corner	conceded	bottom	Corner
2	attempt	right	footed	Shot
3	foul	dangerous	for	Foul
4	yellow	shown	card	Card
5	kick	free	wins	Free kick
6	offside	ball	tries	Offside
7	substitution	replaces	lineups	Substitution
8	goal	shot	box	Goal

TABLE IV
KEYWORDS GENERATED FROM SOCCER WEB-CAST TEXT BY LSA

Category	Rank1	Rank2	Rank3	Event
1	conceded	corner	assisted	Corner
2	high	miss	bit	--
3	foul	wing	card	Foul
4	foul	bad	for	Foul
5	half	own	wins	Free kick
6	offside	caught	through	Offside
7	replace	substitution	defend	Substitution
8	left	box	foot	--
9	right	goal	center	Goal
10	announced	assisted	attacking	--

event. Since the webcast texts of basketball are collected from ESPN, which present a well-defined syntax structure. Therefore, we can conclude that both pLSA and LSA can work well for well-structured texts. Then we extended our test on freestyle webcast texts, which lack of decent structure for event description and exhibit dynamic structure and diverse presentation style. We choose the soccer webcast texts from Yahoo MatchCast Central which is freestyle texts as test texts. The results using pLSA and LSA are listed in Tables III and IV.

For freestyle webcast text, we can clearly see from Table IV that LSA failed to generate the event categories of "shot" and "card," generated the incorrect keywords for the event categories of "corner," "free kick," and "goal," and generated two categories for the same event "foul." In Table IV, "- -" in the last column indicates that there is more than one event types mixed in the generated event category and none of them is dominant, thus we cannot obtain the ground truth for that category. By analyzing the results, we found that LSA cannot well address the polysemy problem. For example, in Category 9 of Table IV, many descriptions of "goal kick" are misclassified into "goal" category because they also contain the word "goal." Goal kick is a type of restart of the game in which the ball is kicked from the goal area. Hence, the word "goal" has different meanings in "goal" and "goal kick." The same situation also occurs

TABLE V
TEXT EVENT DETECTION IN BASKETBALL

Event	Ground truth	Detect Num	Hit	Precision/Recall
Shot	162	181	159	87.8%/98.1%
Jumper	192	215	192	89.3%/100%
Layup	211	214	209	97.7%/99.0%
Dunk	146	151	146	96.7%/100%
Block	110	111	109	98.2%/99.1%
Rebound	515	519	511	98.5%/99.2%
Foul	626	618	618	100%/98.7%
Free throw	216	216	216	100%/100%
Substitution	307	323	307	95.0%/100%

TABLE VI
TEXT EVENT DETECTION IN SOCCER

Event	Ground truth	Detect Num	Hit	Precision/Recall
Corner	206	202	202	100%/98.1%
Shot	388	399	388	97.2%/100%
Foul	428	430	424	98.5%/99.1%
Card	147	129	126	97.7%/85.7%
Free kick	815	827	815	98.5%/100%
Offside	126	126	126	100%/100%
Goal	69	69	69	100%/100%
Substitution	87	88	86	97.7%/98.9%

in the sentences "Second *half* begins, Ascoli 1, Roma 0." and "Francesco Totti (Roma) wins a free kick in his own *half*." Here the word "half" is also the polysemy. LSA cannot avoid overestimating the true similarity between these sentences by counting common words that are used in different meanings. In contrast, we can see that pLSA is still able to achieve good performance on keyword generation for all the event categories in Table III, which validates that pLSA outperforms significantly than LSA for freestyle webcast text.

After keywords are generated, they can be used for events detection in the webcast text. The results of event detection in basketball and soccer are listed in Tables V and VI respectively.

From text event detection results, we can see that most of event categories can be correctly detected while some events have relatively lower precision or recall. For example, in Table V, "shot" and "jumper" events have relatively low precision. This is because some "block" events whose descriptions also have the term "shot" and "jumper" are misclassified into these two events. In Table VI, the card event has low recall because the word "yellow" is selected as the keyword which leads to missed detection of red card events. The errors in text event detection will propagate to text/video alignment for event boundary detection, which may lead to either missing the relevant event or detecting the irrelevant event with different category.

B. Text/Video Alignment

We conducted experiment for text/video alignment to detect event boundary in broadcast basketball and soccer videos. For basketball videos, two NBA games and one Olympic Game are used for training and the rest for testing. For soccer videos, three Euro-Cup 2004 and two World Cup 2002 games are used for training and the rest for testing.

To obtain the ground truth of shot classes and replays is a time-consuming task which has to be manually labeled in the videos. Since our purpose is to detect event boundary using shot classes and replays as visual features and the sports videos in our test set have similar quality in the same domain, we only labeled one NBA basketball game to demonstrate the performance of our shot classification and replay detection approaches. The shot classification task is generally more difficult in basketball video than in soccer video because the basketball playground is usually patterned and the camera converge is smaller, while the soccer pitch is normally dominated by green color. In addition, the basketball video is full of gray color (e.g., the audience, the roof), which might lead to incorrect dominant color detection. Therefore, the result of shot classification and replay detection obtained from a basketball video can provide a benchmark for the overall performance of our approaches.

TABLE VII
VISUAL FEATURE EXTRACTION RESULT FOR BASKETBALL

Feature	Total	Correct	Accuracy
Replay	37	37	100%
Far view shot	150	144	93.3%
In field medium view shot	31	21	67.7%
Out field medium view shot	12	9	75%
In field close-up shot	54	49	90.7%
Out field close-up shot	15	13	86.7%

The result of shot classification and replay detection for basketball video is listed in Table VII. For shot classification, the errors mainly come from the game noise such as unbalanced luminance, shadow, caption, etc. For replay detection, since all the replays contain SDEVs, our replay detection achieves very high accuracy.

We use boundary detection accuracy (BDA) to evaluate the detected event boundaries in the testing video set, which is defined as

$$\mathrm{BDA} = 1 - \frac{\alpha|t_{ms} - t_{ds}| + (1-\alpha)|t_{me} - t_{de}|}{\max\{(t_{de} - t_{ds}), (t_{me} - t_{ms})\}} \tag{20}$$

where t_{ds} and t_{de} are automatically detected start and end event boundaries respectively, t_{ms} and t_{me} are manually labeled start and end event boundaries respectively, α is a weight and set to 0.5 in our experiment.

Tables VIII and IX list the event boundary detection results using CRF and HMM for soccer and basketball videos, respectively. For CRF, we used FlexCRF [41] for the training and inference. For HMM, we first train the HMMs for different event categories in soccer and basketball games using labeled video data to obtain the parameters for each HMM. The features used to train HMMs are same as those used in CRFM. During the detection process, we use same defined search range as CRFM for event boundary detection. We then use the trained HMM to calculate the probability scores of all possible partitions (aligned with the shot boundary) within the search range. The partition which generates the highest probability score is selected as the detected event candidate and the event boundaries can be obtained from the boundaries of the first and last shot in this event.

From the test results, we can see that CRFM outperformed HMM for the event boundary detection for both soccer and basketball videos. This validates the advantages of CRFM over HMM. However, we found that BDA for some events is relatively low such as free kick and substitution in soccer and foul

TABLE VIII
SOCCER GAME EVENT BOUNDARY DETECTION

Event	BDA(CRF/HMM)	Event	BDA(CRF/HMM)
corner	95.3%/83.8%	Free-kick	75.8%/68.1%
shot	94.6%/88.5%	offside	93.2%/77.7%
foul	98.2%/79.2%	Goal	95.9%/80.1%
card	96.9%/79.2%	Substitution	68.2%/61.2%

TABLE IX
BASKETBALL EVENT BOUNDARY DETECTION

Event	BDA(CRF/HMM)	Event	BDA(CRF/HMM)
Shot	97.5%/88.3%	Rebound	90.4%/83.4%
Jumper	96.7%/90.6%	Foul	75.6%/69.8%
Layup	92.1%/82.9%	Free throw	85.1%/77.0%
Dunk	91.9%/79.8%	Substitution	65.0%/58.7%
Block	92.3%/84.5%		

and substitution in basketball. Based on our observation and validation, this can be explained as follows. 1) Such events do not have distinguishable temporal patterns for event boundary modeling. For example, in soccer video, the shot transition during a free kick event is very frequent and fast. In basketball, the foul event is always very short and quickly followed by free throw or throw in. The substitution event in both soccer and basketball video has loose structure and various temporal transition patterns. 2) The errors from shot classification, replay detection and inaccurate time tag of webcast text may lead to low detection result. 3) The event boundaries manually labeled are subjective. For some events such as "goal," the boundaries are labeled following the event temporal transition patterns. But for some events such as "foul," the labeled boundaries only cover the reference shot even there is a temporal transition pattern for such events. In our approach, we strictly follow the event temporal transition pattern to detect boundaries of all the events which may create bias as compared with the manually labeled boundaries for some events. However, we found that all detected event boundaries covered the correspondent event boundaries manually labeled though the length of some events detected using our approach is longer than the length of those generated manually.

V. CONCLUSION

Automatic sports video semantic event detection is a challenge task and is crucial for sports video semantic analysis such as summarization and personalized retrieval. We have presented a novel approach for semantic event detection in sports video by combining the analysis and alignment of webcast text and broadcast video. The experimental results for various soccer and basketball games are promising and validate the proposed approach.

The contributions of our approach include 1) We propose an unsupervised approach based on pLSA to automatically cluster text event and extract event keywords from webcast text in both professional and free styles; and 2) We propose a statistic approach based on CRFM to semantically synchronize the webcast text and broadcast sports video to detect the exact event boundary in the video.

The incorporation of webcast text into sports video analysis, which integrates the complementary strength of low-level features and high-level semantics, is believed to be able to open up a new possibility for sports video semantic analysis and create a promising business model for professional and consumer services.

Our future work will focus on extending the proposed approach for other sports domains, further improving text analysis and text/video semantic alignment approaches, and developing emerging applications (e.g., personalized sports video search engine) for the proposed approach.

REFERENCES

[1] *TRECVID: TREC Video Retrieval Evaluation*, [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid

[2] M. R. Naphade and J. R. Smith, "On the detection of semantic concepts at TRECVID," in *Proc. ACM Multimedia*, New York, 2004, pp. 660–667.

[3] M. Wang *et al.*, "Automatic video annotation by semi-supervised learning with kernel density estimation," in *Proc. of ACM Multimedia*, Santa Barbara, CA, 2006, pp. 967–976.

[4] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proc. ACM Multimedia*, Los Angeles, CA, 2000, pp. 105–115.

[5] M. Xu, N. C. Maddage, C. Xu, M. S. Kakanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, Baltimore, MD, 2003, vol. 2, pp. 281–284.

[6] Y. Gong *et al.*, "Automatic parsing of TV soccer programs," in *Proc. International Conference on Multimedia Computing and Systems*, 1995, pp. 167–174.

[7] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Trans. Image Processing*, vol. 12, no. 5, pp. 796–807, 2003.

[8] D. Zhang and S. F. Chang, "Event detection in baseball video using superimposed caption recognition," in *Proc. ACM Multimedia*, 2002, pp. 315–318.

[9] J. Assfalg *et al.*, "Semantic annotation of soccer videos: Automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 92, pp. 285–305, Nov. 2003.

[10] M. Xu, L. Duan, C. Xu, and Q. Tian, "A fusion scheme of visual and auditory modalities for event detection in sports video," in *Proc.IEEE Int. Conf. Acoustics, Speech, & Signal Processing*, Hong Kong, China, 2003, vol. 3, pp. 189–192.

[11] K. Wan and C. Xu, "Efficient multimodal features for automatic soccer highlight generation," in *Proc. Int. Conf. Pattern Recognition*, Cambridge, U.K., Aug. 23–26, 2004, vol. 3, pp. 973–976.

[12] M. Xu, L. Duan, C. Xu, M. S. Kankanhalli, and Q. Tian, "Event detection in basketball video using multi-modalities," in *Proc. IEEE Pacific Rim Conference on Multimedia*, Singapore, Dec. 15–18, 2003, vol. 3, pp. 1526–1530.

[13] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," in *Proc. ACM Multimedia*, 2002, pp. 347–350.

[14] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos," in *Proc. ACM Multimedia*, Ottawa, Canada, 2001, pp. 261–269.

[15] J. Wang, C. Xu, E. S. Chng, K. Wan, and Q. Tian, "Automatic generation of personalized music sports video," in *Proc. ACM Int. Conf. Multimedia*, Singapore, Nov. 6–11, 2005, pp. 735–744.

[16] D. Reidsma *et al.*, "Cross document ontology based information extraction for multimedia retrieval," in *Supplementary Proc. ICCS03*, Dresden, Germany, July 2003.

[17] V. Mezaris *et al.*, "Real-time compressed domain spatiotemporal segmentation and ontologies for video indexing and retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 606–621, 2004.

[18] A. Jaimes, B. Tseng, and J. R. Smith, "Model keywords, ontologies, and reasoning for video understanding," in *Proc. Int. Conf. Image and Video Retrieval (CIVR 2003)*, July 2003.

[19] A. Jaimes and J. R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2003.

[20] M. G. Strintzis *et al.*, "Knowledge representation for semantic multimedia content analysis and reasoning," in *Proc. Eur. Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, Nov. 2004.

[21] M. Bertini *et al.*, "Ontologies enriched with visual information for video annotation," in *Proc. 2nd Eur. Semantic Web Conf.*, Heraklion, Greece, 29 May–1 June 2005.

[22] N. Nitta and N. Babaguchi, "Automatic story segmentation of closed-caption text for semantic content analysis of broadcasted sports video," in *Proc. 8th Int. Workshop on Multimedia Information Systems'02*, 2002, pp. 110–116.
[23] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," *IEEE Trans. Multimedia*, vol. 4, pp. 68–75, Feb. 2002.
[24] N. Nitta, N. Babaguchi, and T. Kitahashi, "Generating semantic descriptions of broadcasted sports video based on structure of sports game," *Multimedia Tools Applicat.*, vol. 25, pp. 59–83, Jan. 2005.
[25] H. Xu and T. Chua, "Fusion of AV features and external information sources for event detection in team sports video," *ACM Trans. Multimedia Computing, Communications and Applications*, vol. 2, no. 1, pp. 44–67, Feb. 2006.
[26] H. Xu and T. Chua, "The fusion of audio-visual features and external knowledge for event detection in team sports video," in *Proc. Workshop on Multimedia Information Retrieval (MIR'04)*, Oct. 2004.
[27] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and webcasting text," in *Proc. ACM Int. Conf. Multimedia*, Santa Barbara, CA, Oct. 23–27, 2006, pp. 221–230.
[28] C. Xu, J. Wang, H. Lu, and Y. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 421–436, Apr. 2008.
[29] [Online]. Available: http://news.bbc.co.uk/sport2/hi/football/teams/
[30] [Online]. Available: http://sports.espn.go.com/
[31] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 41, pp. 177–196, 2001.
[32] S. Deerwester *et al.*, "Indexing by latent semantic analysis," *J. Amer. Soc. Information Science*, vol. 41, 1990.
[33] F. Erik and K. Tjong, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in *Proc. CoNLL-2003*, pp. 142–147.
[34] [Online]. Available: http://uk.sports.yahoo.com/fo/matchcast.html
[35] D. Cheshire*, The Complete Book of Video—Techniques, Subjects, Equipment.*. London, U.K.: Dorling Kindersley, 1990.
[36] J. Wang, E. S. Chng, and C. Xu, "Soccer replay detection using scene transition structure analysis," in *Proc. Int. Conf. Acoustics, Speech & Signal Processing*, 2005.
[37] L. Duan, M. Xu, Q. Tian, and C. Xu, "Mean shift based video segment representation and application to replay detection," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, QC, Canada, 2004, pp. 709–712.
[38] H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2002.
[39] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Machine Learning*, 2001, pp. 282–289.
[40] E. Chong and S. Zak*, An Introduction to Optimization*. New York: Wiley, 1996.
[41] *FlexCRFs: Flexible Conditional Random Fields*, [Online]. Available: http://www.jaist.ac.jp/~hieuxuan/flexcrfs/flexcrfs.html
[42] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, no. 3, 2005.

Changsheng Xu (M'97-SM'99) received the Ph.D. degree from Tsinghua University, Beijing, China in 1996.

Currently. he is Professor of Institute of Automation, Chinese Academy of Sciences and Executive Director of China-Singapore Institute of Digital Media. He was with Institute for Infocomm Research, Singapore. from 1998 to 2008. He was with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences from 1996 to 1998. His research interests include multimedia content analysis, indexing and retrieval, digital watermarking, computer vision and pattern recognition. He published over 150 papers in those areas.

Dr. Xu is an Associate Editor of ACM/Springer *Multimedia Systems Journal*. He served as Short Paper Co-Chair of ACM Multimedia 2008, General Co-Chair of 2008 Pacific-Rim Conference on Multimedia (PCM2008) and 2007 Asia-Pacific Workshop on Visual Information Processing (VIP2007), Program Co-Chair of VIP2006, Industry Track Chair and Area Chair of 2007 International Conference on Multimedia Modeling (MMM2007). He also served as Technical Program Committee Member of major international multimedia conferences, including ACM Multimedia Conference, International Conference on Multimedia & Expo, Pacific-Rim Conference on Multimedia, and International Conference on Multimedia Modeling.

Yifan Zhang received the B.E. degree from Southeast University, Nanjing, China, in 2004. He is currently pursuing the Ph.D. degree at Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His research interests include multimedia content analysis, machine learning and pattern recognition.

Guangyu Zhu received the B.S. and M.S. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively. He finished his Ph.D. study and submitted his Ph.D. thesis to Harbin Institute of Technology, Harbin, China in 2007.

Currently, he is a Postdoctoral Scientist in NEC Lab America. His research interests include image/video processing, multimedia content analysis, computer vision and pattern recognition, and machine learning.

Yong Rui (M'99-SM'04) received the Ph.D. degree from University of Illinois at Urbana-Champaign.

Dr. Rui serves as Director of Strategy of Microsoft China R&D (CRD) Group. Before this role, he managed the Multimedia Collaboration team at Microsoft Research, Redmond, WA. contributes significantly to the research communities in computer vision, signal processing, machine learning, and their applications in communication, collaboration, and multimedia systems. His contribution to relevance feedback in image search created a new research area in multimedia. He has published twelve books and book chapters, and over seventy referred journal and conference papers. Dr. Rui holds 30 issued and pending U.S. patents.

Dr. Rui is a senior member of both ACM and IEEE. He is an Associate Editor of ACM *Transactions on Multimedia Computing, Communication and Application*s (TOMCCAP), IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGIES. He was an editor of ACM/Springer *Multimedia Systems Journal* (2004–2006), *International Journal of Multimedia Tools and Applications* (2004–2006), and IEEE TRANSACTIONS ON MULTIMEDIA (2004–2008). He also serves on the advisory board of IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. He was on organizing committees and program committees of ACM Multimedia, IEEE CVPR, IEEE ECCV, IEEE ACCV, IEEE ICIP, IEEE ICASSP, IEEE ICME, SPIE ITCom, ICPR, CIVR, among others. He is a general chair of International Conference on Image and Video Retrieval (CIVR) 2006, a program chair of ACM Multimedia 2006, and a program chair of Pacific-Rim Conference on Multimedia (PCM) 2006.

Hanqinq Lu (M'05-SM'06) received the Ph.D. degree in Huazhong University of Sciences and Technology, Wuhan, China in 1992.

Currently, he is Professor of Institute of Automation, Chinese Academy of Sciences. His research interests include image similarity measure, video analysis, object recognition and tracking. He published more than 100 papers in those areas.

Qingming Huang (M'04-SM'08) received the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China in 1994.

He was a Postdoctoral Fellow with the National University of Singapore from 1995 to 1996, and worked in Institute for Infocomm Research, Singapore, as a Member of Research Staff from 1996 to 2002. Currently, he is a Professor with the Graduate University of Chinese Academy of Sciences. He has published over 100 scientific papers and granted/filed more than ten patents in the U.S., Singapore and China. His current research areas are multimedia processing, video analysis, pattern recognition and computer vision.

Dr. Huang was the Organization Co-Chair of 2006 Asia-PacificWorkshop on Visual Information Processing (VIP2006) and served as Technical Program Committee Member of various international conferences, including ACM Multimedia Conference, International Conference on Multimedia and Expo, International Conference on Computer Vision and International Conference on Multimedia Modeling.