Sports Video Analysis: Semantics Extraction, Editorial Content Creation and Adaptation

Changsheng Xu, Jian Cheng, Yi Zhang, Yifan Zhang, Hanqing Lu
National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
Email: {csxu, jcheng, yizhang, yfzhang, luhq}@nlpr.ia.ac.cn

Abstract—**Advances in computing, networking, and multimedia technologies have led to a tremendous growth of sports video content and accelerated the need of analysis and understanding of sports video content. Sports video analysis has been a hot research area and a number of potential applications have been identified. In this paper, we summarize our research achievement on semantics extraction and automatic editorial content creation and adaptation in sports video analysis. We first propose a generic multi-layer and multi-modal framework for sports video analysis. Then we introduce several mid-level audio/visual features which are able to bridge the semantic gap between low-level features and high-level understanding. We also discuss emerging applications on editorial content creation and content enhancement/adaptation in sports video analysis, including event detection, sports MTV generation, automatic broadcast video generation, tactic analysis, player action recognition, virtual content insertion, and mobile sports video adaptation. Finally, we identify future directions in terms of research challenges remained and real applications expected.**

Index Terms—**Sports video, Semantic analysis, Audio keywords, Tactics analysis, Event detection**

I. Introduction

The proliferation of sports game broadcasting leads to the explosive growth of sports video content and the increasing need for the access of sports video content anytime, anywhere and over a variety of receiving devices. However, the rich sports video content has also resulted in much difficulty for the users to access and edit their favorite portions of sports games from huge amount of sports videos. It is clear that when accessing lengthy and voluminous sports video content, the ability to intelligently analyze sports video content to allow efficient browsing, indexing, enhancement, personalization and retrieval of sports video content is highly expected by content providers, service providers and end viewers.

In the past decades, extensive research efforts have been devoted to sports video analysis and applications due to its wide viewer-ship and high commercial potentials. Various approaches and prototypes have been proposed and developed to automatically or semi-automatically analyze sports video content, extract semantic events or highlights, intelligently adapt, enhance and personalize the content to meet users' preferences and network/device capabilities. Many applications, which may create commercial potentials, have also been identified in broadcast sports video editing and enhancement.

In this paper, we summarize our research work and highlight our technical achievements in sports video analysis and applications, particularly on semantics extraction and editorial content creation and adaptation. In section II, we present a generic framework from low-level feature extraction to high-level semantic understanding of sports video analysis. In section III, we describe feature extraction methods in sports video with emphasis on novel mid-level/semantic features we have developed for sports video analysis. In section IV, we introduce how to apply sports video analysis techniques into various applications, including event detection, tactic analysis, personalized sports MTV generation, automatic broadcast video generation, player action recognition, virtual content insertion, and interactive and mobile applications. In section V, we address research challenges and identify future promising directions of sports video analysis according to the current technologies used in sports video analysis and the demands from real-world applications.

II. A Sports Video Analysis Framework

The objective of the sports video analysis is to extract semantics from the source video and intelligently edit, enhance and adapt the sports video content to various applications. Humans tend to use high-level semantics for querying and browsing sports video, while generic low-level features acquired by automated processing of the source video cannot represent semantics directly. The framework of sports video analysis in Fig. 1 describes how to bridge the semantic gap between low-level features and high-level semantics. The framework starts with the low-level feature extraction from the source video. Three kinds of modals of low-level features including visual, audio and text can be directly obtained. Due to the difficulty of directly mapping low-level features to high-level semantics, a mid-level representation is built to bridge the semantic gap. The mid-level representation also contains visual, audio and text models which will be discussed in the following section. With the employment of domain knowledge and machine learning approaches, the mid-level models are selected and fused to describe the corresponding high-level semantics for different applications including event and highlight detection, broadcast sports video and MTV generation, tactic analysis, virtual content insertion and mobile application.

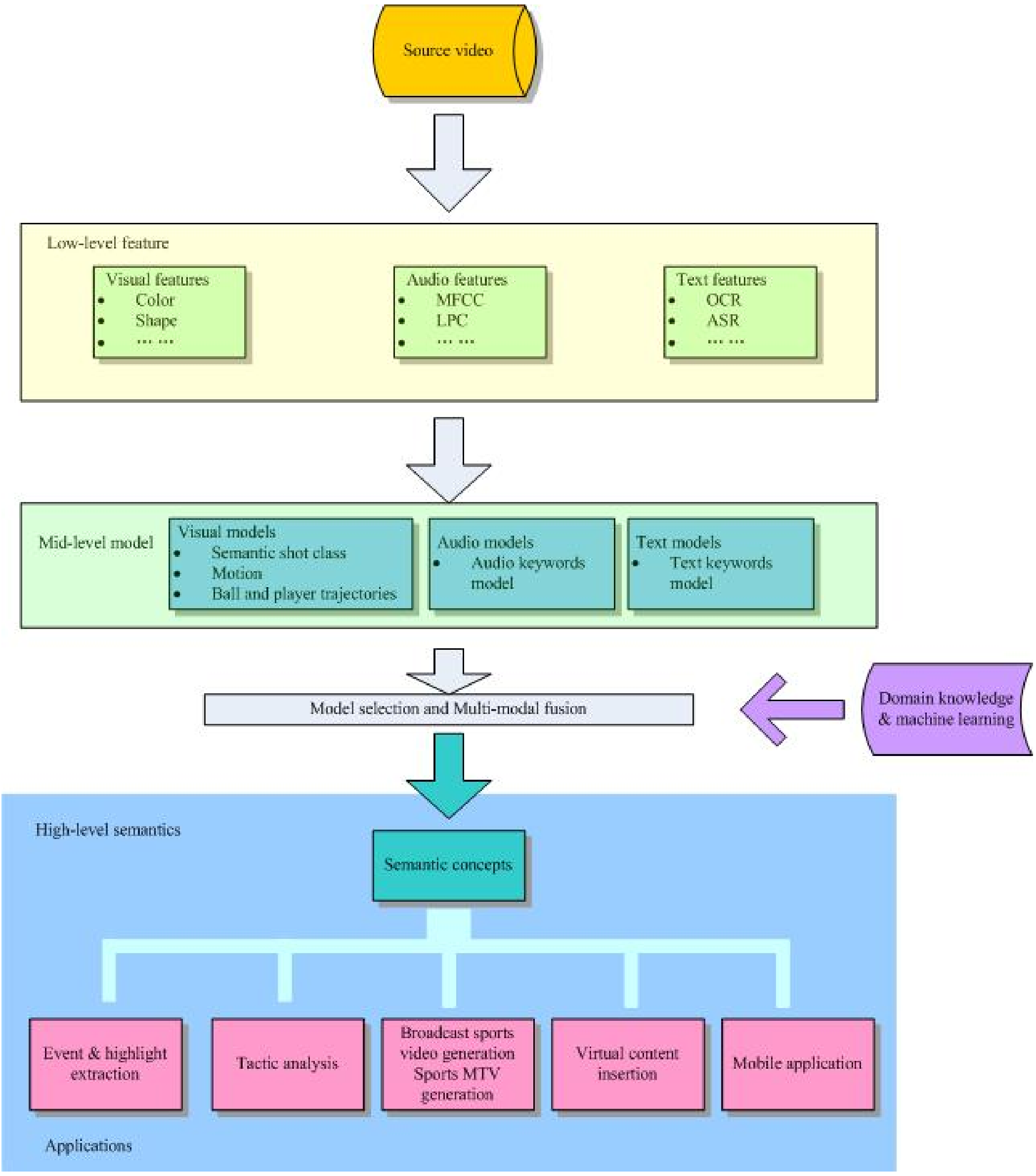

Figure 1. Sports video analysis framework.

III. MID-LEVEL FEATURE EXTRACTION

Humans tend to use high-level semantic concepts, which are precise and abstract to describe meaning, when querying and browsing sports videos. Unfortunately, the information we can obtain directly from the source video is the low-level features which are only simply and raw description of visual or audio content. There exists a large semantic gap between the high-level semantic concept and the low-level features extracted from video sources. In order to bridge the gap, it is necessary to build a mid-level representation between the low-level features and the high-level semantic concepts. The mid-level representation is designed to establish the links from the feature descriptors and the syntactic elements to the domain semantics. The integration of domain knowledge and low-level features strengthens the semantic description ability. It is effective to combine multimodal mid-level representations for high-level semantic content analysis. In this section, we will introduce a set of effective mid-level features developed by us and how they are derived from low-level features using machine learning techniques.

A. Semantic Shot

Sports videos are composed of a series of semantic shots, which appear repeatedly with similar actions and events in the same shot class. The structure of the sports video makes it feasible to classify these shots according to the semantic meanings. As basic unit a shot describes part of or an entire semantic concept. Therefore, the representation and classification of shot are crucial to semantic understanding and provide the basic video structure for further video analysis.

The challenge of the automatic shot representation and classification lies in the frequent variation of the shots in visual appearance. Because of the projection from the three-dimension world to the two-dimension image, the geometric features of objects are ambiguous. In order to build a robust and flexible semantic shot classification system, a set of effective representations with domain and computation constrains are required.

Most existing approaches focused on shot clustering by aggregating shots or key-frames using similar low-level features [1][2]. These shot classification approaches usually resorted to domain knowledge for low-level feature selection. As a result, the selected low-level features were unable to deal with videos of other games and the application range was limited. To tackle these problems, we proposed a novel scheme for semantic shot representation and classification with the emphasis on knowledge representation and acquisition [3]. Our scheme improves the process with employment of supervised learning to perform a top-down video shot classification. Moreover, the supervised learning procedure was constructed on the basis of effective mid-level representations [4] instead of exhaustive low-level features. It extracted low-level features from video data and maps them to the mid-level representation with the nonparametric feature space analysis. The mid-level features including Motion Vector Field Model (MVFM)[5], Color Tracking Model (CTM) and Shot Pace Model (SPM) enables the uniform representation of the shots. MVFM is a nonparametric characterization of the Motion Vector Field (MVF) with feature space analysis. Kernel density based clustering methods are chosen to learn the multidimensional motion patterns from the dataset. Unlike traditional histogram features, the color tracking was performed with the distinguishing color tracker to capture semantic concepts. Temporal and spatial structure constraints of sports video were also added to the color tracking. SPM measured the length of the shots. It was motivated by the fact that there is a distinguishable and consistent shot length difference between major shot classes in team sports video. The rate of each shot's length to the maximum shot length within a symmetric sliding window was regarded as the shot pace measurement. The mid-level features mentioned above described the attributes of the shots with numerical vectors, which provided the basis of the shot classification.

The key issue of the media processing design was the usage and storage of the domain knowledge. We did not attempt to exhaustively extract large amount of low-level features and match them to the stored models. Only the mid-level representations with the important semantic were selected, which significantly accelerated the classification. The supervised learning method (support vector machine) was used to classify the shots into predefined shot categories. Our method combined the domain knowledge of predefined shot categories and human directed feature selection with the automatic machine learning scheme. It took advantages of both the human constructed knowledge and machine learning. The semantic shot classification can be extended to other sports video analysis work such as event/highlight detection and video summarization.

Our method assumed that it was feasible to predefine a set of shot classes with a large coverage for a specific sports video. This assumption is valid for most field-ball type sports video with prominent structure constraints. However, there do exist large amounts of sports videos with a loose structure such as golf, racing. An alternative is to apply a shot clustering approach to the mid-level representations.

B. Audio Keywords

Most of the event detection methods in sports video are based on visual features. In sport video, the transition on events often goes with the change of sounds. Therefore, audio features also play an important role in the semantic understanding of sport video. Nowadays the importance of effective management for audio databases relying on audio content analysis has begun to be realized. There are plenty of sound categories in sports video. Utilization of the audio information to discover interesting content has drawn more attention.

As one of the most used techniques, speech recognition is commonly adopted to acquire the semantics from audio streams of sport video. However, it is still unsatisfactory due to the low precision of the automatic speech recognition. In fact, most of the existing approaches are direct to use the low-level audio features to assist video analysis in sports video. The widely used low-level audio features can improve visual analysis to some extent, but they only provide limited hints for events. A mid-level feature representation is strongly required to audio analysis for diverse sports types.

We proposed a mid-level representation called audio keywords to assist sports video analysis [6]. "Keyword", which was borrowed from the text mining domain, contained strong inference of the high-level semantic events. Audio keywords were created from the low-level features with the feature selection and machine learning approach. Therefore, the audio keywords were employed to connect the low-level features and high-level semantic content. There were two categories of the audio keywords: generic and specific. Generic audio keywords were defined as the common sounds of sports games such as commentator speech and audience sound.

Domain-specific audio keywords described the representative sound of different games with their own rules and playing fields. Creation of these keywords was conducted with the classification of low-level features extracted for the corresponding keywords. There were a large amount of audio features as the basis of the classification including Zero-Crossing Rate (ZCR), Linear Prediction Coefficients (LPC), Linear Prediction Cepstrum Coefficient (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), Short Time Energy (STE), Spectral Power (SP), etc. These low-level features were selected to find the most effective features for each audio keyword classification. The feature selection was automatically implemented with a feature selection vector and Support Vector Machines (SVM). The SVM was designed as the hierarchical structure according to the domain knowledge of different sports games for hierarchical classification. These audio keywords show a strong correlation with interesting events in sports video than other existing audio techniques. The applications to event detection verified the effectiveness of the proposed audio keywords.

C. Camera Motion

Motion information is another important mid-level representation for visual perception. In many situations, high-level semantic content is closely related to the object and camera motions. Most approaches for motion estimation and pattern recognition were based on predefined parametric models [7][8]. The disadvantage of the parametric model based approaches lies in the difficulty of determining the appropriate class model.

In order to overcome the disadvantage of the parametric model, we proposed a nonparametric based motion analysis approach [9]. The nonparametric approach introduces the feature of motion vector represented by the cone-shaped Motion Vector Space (MVS). Because of the low quality of the broadcast sports video, the directly extracted MVS vectors were unable to fulfill the requirement for motion analysis. Thus, it was necessary to reprocess the vectors to reduce the affection of noise and take account of the spatial-range constraints. Mean shift [10] was employed as the filter of the feature vectors for its discontinuity preservation and mode-seeking performance. Motion Vector Fields (MVF) were smoothed and transformed into a mosaic comprising a set of colored pieces. Histograms of the pixels' horizontal and vertical projection consisted of the feature vector for motion pattern analysis. The camera motion classification was achieved by the C-support vector classifier with an RBF kernel function.

The contribution of the nonparametric approach is the development of a novel effective mid-level motion representation without predefined models for shot description. Diverse camera shots and frequent occurrences of bad optical flow estimation make it impractical to keep a proper parametric assumption valid in a wide range of video scenarios. Existing parametric approaches are unable to solve the model selection problem. Our approach provides the uniform feature for motion pattern analysis with the solution of the novel nonparametric feature.

D. Object Trajectory

Player and ball are the main moving objects in sports video. Their trajectories contain abundant semantic information which is useful for video analysis, such as event detection, tactics analysis, etc. In order to obtain trajectory, player or ball detection and tracking are the necessary process. However, object detection and tracking encounter additional difficulties in broadcast sports video. Most of object detection and tracking approaches are based on the videos in certain conditions such as camera motions, object motions, background and video qualities. However, the circumstance of the ball game and the specialty of the broadcast video make the tracking results unsatisfying with the direct application of existing object tracking algorithms. The low resolution, motion blur and noise in the broadcast video significantly affect the result of detection and tracking. The low video resolution and the widely distributed noise in the video reduce the signal-to-noise ratio. The ball and players in long shots of the broadcast video have a very small size. The motion blur and the occlusion make the shape of the ball and player variable. In the multi-player ball games, another difficulty lies in the occlusion between players and the ball. These difficulties make it unreliable to directly apply the existing object detection and tracking approaches to the trajectory extraction of ball and player in broadcast sports video.

Traditional ball detection approaches detected the ball with the heuristic rules of the size, shape and color. However, the traditional approaches often fail to detect ball due to wide variation in broadcast sports video, i.e. occlusion, high speed motion blur, the merge with other objects. We proposed a trajectory based ball detection and tracking method to overcome the challenges [11]. The detection and tracking process was partitioned into two steps. A set of ball candidates were generated firstly. The anti-model approach was applied to the candidate generation. A group of sieves were designed with the properties of the ball including the size, color and shape to remove non-ball objects from the candidates. The rest ball candidates were selected and processed to compute the set of ball trajectories utilizing the temporal and spatial motion information of the ball. By plotting the candidate locations of the ball over the time, the ball and non-ball objects can be differentiated with the length of their consistent trajectories. The ball has a longer trajectory than the non-ball objects. A heuristic measure of candidate ball trajectory's index confidence was proposed to select the most likely ball trajectory in the candidate set.

Different from most ball detection algorithms which perceive the ball as a single object in the frames, our ball detection and tracking scheme considered the consistence of the whole ball trajectory. It was relatively easy to achieve very high accuracy in locating the ball among a set of ball-like candidates and much better to study the trajectory information of the ball because the ball was the "most active" object in the broadcast sports video.

Unlike the ball, player tracking is often considered as a multi-object problem because there are usually multiple players in a shot of broadcast sports video. The challenges of multi-player tracking in broadcast sports video can be summarized as follows: (1) The cameras used to capture sports game are not fixed and they are always moving to follow the players; (2) The broadcast video is an edited video where the broadcast sequence feed is selected from frequent switches among multiple cameras according to broadcast director's instruction; (3) There are numerous players moving in various directions in broadcast video; and (4) The background in broadcast sports video changes sharply. Therefore, detection and tracking of players in broadcast sports video remains difficult.

We proposed a unified framework for multi-player trajectory extraction [12]. Support vector classification combined with playfield segmentation was employed to detect the players for the initialization of the tracker. The particle filter was improved by integrating the support vector regression (SVR) into sequential Monte Carlo method. The proposed SVR particle filter was applied to the tracking of multiple players in broadcast sports video. As a nonparametric estimator of the posterior density, SVR was more suitable than other existing methods to tackle the nonlinear and non-Gaussian player tracking problem in broadcast sports video and more robust to the noise. The SVR particle filter enhanced the performance of the classical particle filter with small sample set and increased the efficiency of tracking system for multi-player tracking.

E. Attention Features

Audiences are more likely to be attracted by exciting segment than the other portion in sports video, such as goal, foul, etc. Therefore, the attention detection and representation would be very useful for sports video editing and adaption. Once we know what is attractive in sports video, some tasks, such as summarization, personalized customization, would be easy to be implemented. The traditional attention detection is realized by gaze detection, saliency detection, or motion analysis. These methods are often inconvenient or not much effective due to additional device required or single modality adopted.

We proposed an effective attention detection and representation by behavior analysis which utilizes both visual features and audio features [24]. Since the users pay most attention to player behavior and audience response in racket sport highlight, we extracted attention features from player behavior including action and trajectory, and game-specific audio keywords. To recognize player action, we started by player tracking and human-centric figure computation. Then optical flow was computed on the difference images of adjacent human-centric figures as low level feature. We proposed a novel motion descriptor based on grid partition with the relationship between human body parts and optical flow field regions. Finally, the action recognition was carried out in the framework of support vector machine. Voting strategy was utilized for action clip recognition by aggregating the frame classification over the temporal domain.

Besides visual feature, audio also contains much affection information. In racket game, audience is used to giving applause after a player score. The longer duration and higher average energy of the applause are, the more exciting the score event is. There the audio keywords--Applause is exploited to extract attention features as the response of audience form audio component of sports video. The features include duration of Applause, which is the time duration of audience applause, and Applause Average Energy (AAE) which is the energy measurement for applause signal.

IV. APPLICATIONS

In this section, we introduce several emerging applications of sports video analysis on editorial content creation and sports video enhancement/adaptation, including event detection, personalized sports MTV generation, automatic broadcast sports video generation, tactic analysis, player action recognition, virtual content insertion and mobile sports video adaptation. By employing the mid-level representations discussed in the previous section, these applications aim to build the semantic bridge between the source video and semantic concept. According to the characteristics of different applications, domain knowledge and machine learning approaches are utilized to select the mid-level features and learn the semantic concepts from the multi-modal representation.

A. Event Detection

Event detection aims at automatically detecting interest ing or significant events from sports video. It is essential for sports video summarization, indexing and retrieval. Extensive research efforts have been devoted to this area. The approaches we proposed for event detection can be classified into two categories: using video content only and using video content together with external sources.

The basic idea of approaches using video content only is to use low-level or mid-level features and rule-based or statistical learning methods to detect events in sports videos. Since single modality is not able to fully characterize the events in sports video, we propose an effective fusion scheme of visual and auditory modalities to detect events in tennis video [13]. We utilized our proposed unified framework for semantic shot classification in tennis video. In tennis video, shots are classified into Close-up, Court view, Player Medium-View, Audience, Bird-View and Replay six classes. Within these shots, game specific sound, commentator speech, and environmental sounds are detected and recognized by a hierarchical Support Vector Machine (SVM) classifier. Subsequently, we heuristically associated potential semantics of camera shots and sounds to identify five interesting events (i.e. serve, reserve, ace, return, score) in tennis videos.

We also combined audio/visual features to detect goal scoring moments as the highlights in soccer video [14]. We used Composite Fourier Transform (CFT) to detect dominant speech portions in the noisy commentary that corresponds to excited vocal. Cumulative CFT density values were obtained by summing all density values within a sliding a 5-sec window over successive audio frames. These dominant speech segments are the candidates of highlights and need to take a visual verification. First, the current shot was examined for the appearance of goal-mouth. Then the subsequent shots were examined for the rate of shot changes. This cascaded process filters spurious candidates from the noisy audio and enhances the highlight detection precision.

Due to the semantic gap between low-level features and high-level events as well as dynamic structures of different sports games, the method of using video content only is difficult to address following challenges: (1) ideal event detection accuracy; (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 and live videos. In order to address these challenges, we attempted to seek available external sources to help.

Web-casting text is a text broadcast source for sports game and can be live captured from the web. Incorporating web-casting text into sports video analysis, we developed a live soccer event detection system and conduct a live trial on World Cup 2006 games [15]. In this system, given a web-casting text and a broadcast video of corresponding soccer game, it can automatically synchronize text and video at both temporal and semantic level to achieve event detection and semantic annotation. The system contained four live modules: live text/video capturing, live text analysis, live video analysis, and live text/video alignment. The live text/video capturing module captures the web-casting text from the web and the broadcast video from TV. In the captured text, for each type of the events in soccer game, we defined a keyword. By searching these keywords, the relevant event can be detected. Then we recorded the type of event and players/team involved in the event for personalized summarization. The time stamp of each detected event was also logged for text/video alignment. In video analysis module, we detected the start point of the game in the video and used it as a reference point to infer the game time. Then we detected the digital clock overlaid on the video and recognized the time from the video clock. These two steps were combined together to detect the game time in the video. For text/video alignment, the event moment in the video was detected by linking the time stamp in the text event to the game time in the video. The event moment is the reference point to determine the start and end boundary of the event. A Finite State Machine (FSM) was employed to model the shot type transition pattern in the video and detect the event boundary. The whole system was tested on both live games and recorded games. The results were encouraging and comparable to the manually detected events.

Based on the live event detection system, we proposed a framework for semantic annotation and personalized retrieval of sports video [16]. We extended the applications from soccer to basketball. The detected text events and semantics were used to annotate the related events in the video and generated a text summary for the sports game. A platform was provided to support personalized retrieval based on the queries of the users. It enables users to generate video summaries related to certain event, player or team, which is different from the one-to-many mode of the current sports video summary generation.

To make the event detection approach more generic, we further improved our work [17]. In web-casting text analysis module, we used pre-defined keywords for text event detection [15][16], which is less generic to different sports domain. Due to the different presentation styles of web-cast text and different sports domains, it is difficult to use prior knowledge to detect event from web-cast text in order to achieve a generic event detection framework. We proposed an unsupervised approach for text analysis [17]. We employed probabilistic Latent Semantic Analysis (pLSA) to automatically cluster text event from web-casting text and extracted keywords from the events in each cluster. For text/video alignment, we used a Conditional Random Field model (CRF) instead of the Finite State Machine (FSM) method in [15] and Hidden Markov Model (HMM) in [16] to model the temporal event structure and detect the event boundary. CRF is a probabilistic framework which brings together the merits of discriminative and generative models. It relaxes the independence assumptions and avoids the label bias problem, which makes it outperform HMM on real-word sequence labeling problem. We used it to label the visual feature sequence extracted from the video in order to determine the exact boundaries of events.

B. Personalized Music Sports Video Generation

Automatic generation of personalized music sports video is another video editing application [20]. Current production of such music video is very labor-intensive and inflexible. The selected clips must be found manually from huge volumes of sports video documents by professional producers, who then cut and edit each selected segment to match a given music. By introducing tools which can automate the process, production efficiency for professionals can be improved. We proposed a personalized music sports video generation approach [20]. In our approach we addressed two research challenges: semantic sports video content selection and automatic video composition. Multimodal features (audio, video and web-casting text) were utilized to detect events and extract high-level semantics from the events. Based on the semantics and detect events, sports video contents can be selected for music video composition. We introduced two schemes to generate music sports video: the video-centric scheme and music-centric scheme. The first one was conducted in a video summarization manner by sorting and selecting video contents according to user preference and matching them with music clips. For music-centric scheme, it is more difficult as it is driven by both the video and the music. Both content and tempo matching between the video and music are required. We parsed the music structure with three types of boundary information: beat boundary, lyric boundary and semantic music structure boundary. A rule based method was defined to select suitable content to match the music structure including the portion of Intro, Verse, Chorus, Bridge and Outro. To match the music tempo, a scoring scheme was introduced to find an event whose shot durations could perfectly match the length of respective music boundaries. In addition, the motion of each shot was also computed and matched with that of the music such that the speed of video and music are similar.

C. *Automatic Broadcast Video Generation*

Automatic broadcast sports video generation is able to not only improve broadcast video generation efficiency, but also provide the possibility to customize sports video broadcasting. Research in this area is rare. We proposed an automatic composition system of broadcast soccer video [21]. There are two major issues required in the implementation of the system: the automatic replay generation and the camera view selection/switching from the raw unedited main/sub-camera video seeds.

Automatic replay generation [22] is to detect replay-worthy events from the raw sports video and find suitable time slot to insert the replays. It aims to reduce the labor cost during the live sports video production process. Compared with the broadcast sports video, automatic event detection in unedited video is more challenging and difficult as there are few cues that can be utilized. There is also no direct relationship between the low-level feature patterns and high-level events. To bridge the gap between the features and events, we proposed a mid-level representation which consists of five mid-level features: active play position keyword, ball trajectory, goalmouth location, motion activity and audio keyword. Using this multimodal mid-level representation, replay-worthy events were detected via SVM classifiers which classified the five synchronized mid-level feature sequences into event and no-event. For replay insertion, it contains two situations: instant replay and delayed replay. Most replays are instant replays that are inserted almost immediately following the event if subsequent segments are uninteresting. Delayed replay occurs for several reasons: a) the event is missed by the main camera , b) the event to be replayed is followed by an interesting segment, hence the broadcaster has to delay the replay, and c) the event is important and worth being replayed many times. Our system examined whether an instant replay can be inserted at the following no-event segment, and reacted accordingly. In addition, the system examined whether the same event met the delayed replay condition. If so, the system buffered the event and inserted the replay in a suitable subsequent time slot. The automatic replay generation results are promising and found to be comparable to those generated by broadcast professionals.

For camera view selection/switching, our system mimicked the professional director's choice of using the main camera capture and at proper instances launching the sub-camera segment which has the clearest game view among all the other sub-cameras. A HMM-based classifier was employed to segment the sub-camera capture into alternative partitions of suitable and unsuitable for composition selection. Then during the no-event segment in the raw video, we selected a proper duration of the main-camera view and switched to a sub-camera view, and repeated this alternative switching until the end of the no-event segment.

D. *Tactic Analysis*

Unlike the traditional semantic event extraction presenting the most interesting facts to the audience without any further analysis, tactic analysis aims to help the coaches and professionals to understand the game process, tactic patterns in the attack and defense and tactics in certain events. The statistics obtained by the tactic analysis is able to assist to adapt the training plan and improve the performance of the team.

The challenge of tactic analysis lies in the representation of the tactics and the classification of the tactic patterns based on the representation. Tactics is a high-level semantic concept. It is required to develop an effective mid-level representation to bridge the gap between low-level features and tactic concepts. Description of different tactics in games for classification and clustering is also challenge for tactic analysis in sports video.

Most existing work are unable to tackle team games such as soccer for their complex tactics executed by group of players. We proposed a tactic analysis framework based on the ball and player trajectories in certain events [18][19]. With the web-cast text retrieved on the Internet, events in the sports videos were detected and classified. The tactics used in the game were characterized by the behavior of individual player and the interactions among the players and ball. The ball and player trajectories reflected such characterization, where ball and players were located and analyzed with their mutual relationship. We developed an aggregate trajectory for tactic representation, which was constructed based on multiple trajectories using a novel analysis of temporal-spatial interaction among the players and the ball. Geometric features of the aggregated trajectories containing location, speed and curvature were employed to represent the substantial tactics included in the trajectories. The tactics patterns of route and interaction were classified with heuristic rules according to the features extracted from the aggregate ball and player trajectories.

This was the first solution for soccer game tactic analysis based on broadcast video. Object trajectory was developed as one kind of the generic features for the team sports. The tactic representation and information extracted from the trajectory was consequently general for the tactics analysis of team sports. In the future, the proposed tactic representation and temporal-spatial interaction analysis can be extended to other team sports video such as hockey and American football.

E. Player Action Recognition

Player action can be perceived as a kind of mid-level representation. Player action contains important semantic information which indicates the occurrence of certain events, tactics and highlights. The actions performed by players in sports games reveal the process of the game and the tactics of the players. The movement of players in sports video provides useful information for analysis and understanding of the game. Thus, recognition of player actions is essential for the analysis of sports games and is desired by sports professionals and audiences for technical and tactic assistance.

Action recognition was widely studied in surveillance videos. However, there still existed challenges for action recognition in broadcast sports video. Most of the sports videos are composed of long shots where the whole field is focused on. The ratio of player's body to the frame size is much lower in sports video than in surveillance video. The difficulty increases greatly when the scale of the player to be recognized in the video reduces. Moreover, in broadcast video, the player action recognition is also affected by the random motion of the camera, noise and motion blur.

We proposed a novel motion analysis approach to solve the player actions recognition problem in far-view frames of the broadcast sports video [23][24]. Our player action recognition method was based on motion analysis, which was different from appearance-based approaches. The motion features for small scale players were more robust than appearance-based features. In order to overcome the affect of the noise in the video, we employed the optical flow and treated it as spatial patterns of noisy measurements instead of precise pixel displacements. We proposed a new motion descriptor, slice based optical flow histograms, as the motion representation. The new descriptor emphasized the motion relationship among different parts of the player's body. Several other features including audio keywords consisted of the multimodal feature vector for the supervised learning classifier. According to the recognized player actions we ranked the highlights for the summarization of the sports video.

The proposed player action framework can be applied to other applications such as match content enrichment and immersive environment reconstruction.

F. Virtual Content Insertion

Virtual Content Insertion (VCI) is an emerging application of video analysis and has been used in video augmentation and advertisement insertion. There are three critical issues for a VCI system: when (time), where (place) and how (method) to insert the Virtual Content (VC) into the video. We proposed a generic VCI system [25] based on visual attention analysis. This system selected the insertion time and place by performing temporal and spatial attention analysis, which predicted the attention change along time and the attended region over space. In order to enable the inserted VC to be noticed by audience while not to interrupt the audience's viewing experience to the original content, the VC was inserted at the time when the video content attracts much audience attention and at the place where attracts less. Dynamic insertion in general videos was performed by using affine rectification and Global Motion Estimation (GME). The method needs only two pair of parallel lines, which are relatively easy to be obtained in most videos. The VCI system is able to obtain an optimal balance between the notice of the VC by audience and disruption of viewing experience to the original content.

G. Mobile Sports Video Adaptation

Nowadays, the handheld Internet-capable devices have been undergoing a booming prosperity in our daily lives. With these portable devices, it is convenient for the mobile users to access the Internet anywhere and anytime. Hence providing on-demand sports video to mobile device is another necessary application. To enable the mobile users to really enjoy the ease of video resources, the limited accessible bandwidth is one key bottleneck to be handled. We proposed a system [26] which focused on reducing bandwidth consumption with transmitting important video segments to the users instead of the whole sports game video. The system architecture comprised two core components: the highlight extraction component at the video server side, and the client UI component at the mobile client side. The mobile client began a request to an online sports video. The front end of sports video server obtained the client request and forwarded the request to the video fetching component. The video fetching component examined the client request and determined whether to call the highlight extraction component. Video service component transferred the original or highlight video to the frond end. The client interacted with the video via the local browsing UI. On the server side, the highlights of the sports video were extracted based on replay detection. On the client side, an advanced UI was designed to effectively browse these highlights on the mobile devices. We designed a browsing interface on a Dopod 575 smart phone with a 176x224 pixel resolution and Microsoft Smartphone 2003 as the operation system.

To improve our system, we combined quality-domain video compression with content-based sports highlights extraction to reduce the bandwidth consumption more effectively [27]. We customized normal general video coding schemes for mobile devices from two aspects. The one is to select the Intra-frame adaptively according to the shot length in the video. The other is to pay more attention on interesting regions coding. Taking an example of soccer video, we detected the attention region based on field and object segmentation, then decreased the visual quality of the other non-attention regions. Experimental results showed that our video coding scheme can reduce more than 77.5% of the bandwidth consumption.

V. FUTURE DIREDTIONS

Sports video is one of the important video genres and has become a key driver content for various services and applications. In the past decades, significant achievements have been made in certain research areas and applications in sports video analysis. However, research challenges still remain due to the difficulty of high-level semantic information extraction from video content using computer vision and image processing techniques and semantic alignment of video content with the manually generated ontology. On the other hand, current user expectations for sports video technologies and applications still far exceed the achievement of today's state of the art technologies. These challenges and expectations will motivate researchers to explore new research approaches. Based on current technologies achieved in sports video analysis and the expectations from real-world applications, we identify following future directions in sports video analysis.

A. Cross-media Semantic Annotation and Retrieval

Semantics extraction is a hard problem in sports video analysis. Mining single modality alone has reached its limit in semantics extraction. Nowadays rich external sources are available and able to assist sports video semantic analysis. Therefore, analysis of multiple modalities to achieve synergy among the semantic cues from different information sources is a trend for sports video semantics extraction and annotation.

From users' point of view, a Google similar sports video search engine is desired. Different from text-based Google search, sports video search and retrieval is challenging due to the difficulty to fuse and synchronize various complex, heterogeneous and dynamic information sources. Therefore, advanced search and retrieval frameworks based on cross-media analysis and fusion is a research trend for sports video retrieval. This may need exploration to build a semantic space to facilitate similarity measure at semantic level among different heterogeneous modalities (e.g. text, audio, video, etc.).

B. Generic Solution and Knowledge Integration

Although we have seen encouraging results achieved in different specific sports games, it is difficult or almost impossible to extend the approach developed for certain sports domain to a new sports domain or even the same domain with different conditions. This is because those approaches heavily rely on the features appropriate for the specific games and game specific rules for sports video analysis, which lack of generic representation and structure for different sports domains. This will motivate researchers to investigate the advanced approaches to integrate the cues of different information sources and the knowledge of various sports games in structure analysis and semantics extraction.

C. Robust Performance on Large Scale Test Data

Most of the approaches proposed for sports video analysis are research oriented, which were tested on small dataset created by individual researchers. Although it is possible to achieve promising results on limited dataset by adopting advanced learning approaches or strong domain rules, the performance may dramatically decrease if the test data get larger. In sports video research community, a large scale common video database is desired to provide a benchmark for testing different research approaches. This has already been applied to other domains, e.g. TRECVID for news video analysis, ImageCLEF for medical image retrieval.

D. User Study and Personalization

One limitation of current sports video analysis research and sports video services is seldom to consider users' real needs. For example, current sports video events and highlights are generated through a one-to-many mode and users have to passively watch them despite different users may have different preferences for events and highlights. As rich media will take center-stage and play a bigger role in our daily lives, media consumption is no longer just passively watching, but actively engaging. To realize this, we need to know users' preferences and provide an effective tool to facilitate users to interact with the systems. Therefore, automatic preference acquisition is a research trend to achieve personalization of video content. Automatic personal preference acquisition is a challenging task due to the difficulty of capturing and analyzing human behaviors as well as semantic connection of human behaviors to video content. Since different persons may have different preferences even for the same content, how to create a generic model to analyze human behaviors corresponding to video content is extremely important for automatic personal preference acquisition. Developing techniques capable of learning user profiles from their behaviors or habits in media experience is also important to create personalized video content and less intrusive recommendation.

REFERENCES

[1] A. Ekin, A. M. Tekalp, "Shot type classification by dominant color for sports video segmentation and summarization," *ICASSP*, pp.173-176, 2003.

[2] S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of goal segments in basketball videos," in *Proc. ACM Multimedia*, Ottawa, ON, Canada, Sep. 30, 2001, pp. 261–269.

[3] L. Duan, M. Xu, Q. Tian, C. Xu, J. S. Jin, "A Unified Framework for Semantic Shot Classification in Sports Video," *IEEE Transaction on Multimedia,* 2005, vol. 7(6), pp. 1066-1083.

[4] L. Duan, M. Xu, T. Chua, Q. Tian, C. Xu, "A mid-level representation framework for semantic sports video analysis," in *Proc. ACM Multimedia,* Berkeley, CA, Nov. 1, 2003, pp. 33–44.

[5] L. Duan, M. Xu, Q. Tian, and C. Xu, "Mean shift based nonparametric motion characterization," in *Proc. Int. Conf. Image Processing,* Singapore, Oct. 24, 2004, pp. 1597–1600.

[6] M. Xu, C. Xu, L. Duan, J. S. Jin, S. Luo, "Audio Keywords Generation for Sports Video Analysis," *ACM Transaction on Multimedia Computing, Communication and Application,* 2008, vol. 4(2), pp. 11.1-11.23.

[7] P. Bouthemy, M. Gelgon, F. Ganansia, "A Unified Approach to Shot Change Detection," *IEEE Tran. CSVT*, vol. 9(7), pp. 1030-1044, 1999.

[8] Y. Tan, D. D. Saur, S. R. Kulkami, P. J. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation," *IEEE Tran. CSVT*, vol. 10(1), pp. 133-146, 2000.

[9] L. Duan, M. Xu, Q. Tian, and C. Xu, "Nonparametric Motion Characterization Using Mean Shift for Robust Camera Motion Pattern Classification," *IEEE Transaction on Multimedia*, 2006, vol. 8(2), pp. 323-340.

[10] D. Comaniciu, P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," *IEEE Tran. PAMI*, vol. 24(5), pp. 1-18, 2002.

[11] X. Yu, H. Leong, C. Xu, Q. Tian, "Trajectory-Based Ball Detection and Tracking in Broadcast Soccer Video," *IEEE Transaction on Multimedia*, 2006, vol. 8(6) pp. 1164-1178.

[12] G. Zhu, C. Xu, Q. Huang, W. Gao, "Automatic Multi-player Detection and Tracking in Broadcast Sports Video Using Support Vector Machine and Particle Filter," *IEEE International Conference on Multimedia and Expo*, Toronto, Canada, 9-12 July 2006, pp. 1629-1632.

[13] M Xu, L. Duan, C. Xu, Q Tian "A Fusion Scheme of Visual and Auditory Modalities for Event Detection in Sports Video," *IEEE International Conference on Acoustics, Speech, & Signal Processing*, Hong Kong, China, 2003, vol. 3, pp. 189-192.

[14] K. Wan, C, Xu, "Efficient Multimodal Features for Automatic Soccer Highlight Generation," *International Conference on Pattern Recognition*, Cambridge, UK, 23-26 Aug. 2004, vol. 3, pp. 973-976.

[15] C. Xu, J. Wang, K. Wan, Y. Li, L. Duan, "Live Sports Event Detection Based on Broadcast Video and Web-casting Text," *ACM International Conference on Multimedia*, Santa Barbara, CA, 23-27 Oct. 2006, pp. 221-230.

[16] C. Xu, J. Wang, H. Lu, Y. Zhang, "An Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video," *IEEE Transaction on Multimedia*, 2008, vol. 10(3), pp. 421-436.

[17] C. Xu, Y. Zhang, G. Zhu, Y. Rui, H. Lu, Q, Huang, "Using Web-cast Text for Semantic Event Detection in Broadcast Sports Video," *IEEE Transaction on Multimedia*. To be appeared.

[18] G. Zhu, C. Xu, Q. Huang, Y. Rui, S. Jiang, W. Gao, H. Yao, "Event Tactic Analysis Based on Broadcast Sports Video," *IEEE Transaction on Multimedia*, To be appeared.

[19] G. Zhu, C. Xu, Q. Huang, Y. Rui, S. Jiang, W. Gao, H. Yao, "Trajectory Based Event Tactics Analysis in Broadcast Sports Video," *ACM International Conference on Multimedia*, Augsburg, Germany, 24-29 Sep. 2007, pp. 58-67.

[20] J. Wang, C. Xu, E. Chng, K. Wah, Q. Tian, "Automatic Replay Generation for Soccer Video Broadcasting," *ACM International Conference on Multimedia*, New York City, USA, 10-16 Oct. 2004, pp. 32-39.

[21] J. Wang, C. Xu, E. Chng, H. Lu, Q. Tian, "Automatic Composition of Broadcast Sports Video," *ACM Journal of Multimedia Systems*, 2008, vol. 14(4) pp. 179-193.

[22] J. Wang, C. Xu, E. Chng, L. Duan, K. Wan, Q. Tian, "Generation of Personalized Music Sports Video Using Multimodal Cues," *IEEE Transaction on Multimedia*, 2007, vol. 9(3): pp. 576-588.

[23] G. Zhu, C. Xu, Q. Huang, W. Gao, L. Xing, "Player Action Recognition in Broadcast Tennis Video with Applications to Semantic Analysis of Sports Game", *ACM International Conference on Multimedia*, Santa Barbara, CA, 23-27 Oct. 2006, pp. 431-440.

[24] G. Zhu, Q. Huang, C. Xu, W. Gao, L. Xing, "Human Behavior Analysis for Highlight Ranking in Broadcast Racket Sports Video," *IEEE Transaction on Multimedia*, 2007, vol. 9(6), pp. 1167-1182.

[25] H. Liu, S. Jiang, Q. Huang, C. Xu, "A Generic Virtual Content Insertion System Based on Visual Attention Analysis," Accepted, *ACM International Conference on Multimedia*, Vancouver, Canada, 26-31 Oct. 2008.

[26] Q. Liu, Z. Hua, C. Zang, X. Tong and H. Lu, "Providing on-demand sports video to mobile devices," *ACM International Conference on Multimedia*, Singapore, pp. 347 – 350, 2005.

[27] C. Zang, Q. Liu, X. Tong and H.Lu, "A framework for providing adaptive sports video to mobile devices," *Proceedings of the 2nd international conference on Mobile multimedia communications*, Alghero, Italy, 2006, vol. 324.

Changsheng Xu received the Ph.D. degree from Tsinghua University, Beijing, China in 1996.

Currently he is a 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 Senior Member of IEEE and Member of ACM. He 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.

Jian Cheng received the B.S. degree and M.S. degree in mathematics from Wuhan University in 1998 and in 2001, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences in 2004.

He is currently an associate professor of Institute of Automation. His research interests include image and video retrieval, machine learning, etc.

Yi Zhang received the B.S. and M.S. degrees in control science and engineering from Zhejiang University, Hangzhou, China, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree in pattern recognition at Institute of Automation, Chinese Academy of Sciences, Beijing, China.

From 2008, he joined China-Singapore Institute of Digital Media as an intern. His research interests include video processing, multimedia content analysis and pattern recognition.

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, computer vision and pattern recognition.

Hanqing Lu received the B.S. degree in computer science and the M.S. degree in electrical engineering from Harbin Institute of Technology in 1982 and 1985, respectively, and the Ph.D. degree in electronic engineering from Huazhong University of Science and Technology in 1992.

He is currently a professor of Institute of Automation, Chinese Academy of Sciences. He directs the Image and Video Analysis Research Group at National Laboratory of Pattern Recognition. His research areas include image retrieval, mobile computing, face detection, and video analysis. He has published over 50 international journals and conference papers.

Prof. Lu is Senior Member of IEEE and member of ACM.