Human Behavior Analysis for Highlight Ranking in Broadcast Racket Sports Video

Guangyu Zhu, Qingming Huang, *Member, IEEE*, Changsheng Xu, *Senior Member, IEEE*, Liyuan Xing, Wen Gao, *Senior Member, IEEE*, and Hongxun Yao, *Member, IEEE*

***Abstract*—The majority of existing work on sports video analysis concentrates on highlight extraction. Little work focuses on the important issue as how the extracted highlights should be organized. In this paper, we present a multimodal approach to organize the highlights extracted from racket sports video grounded on human behavior analysis using a nonlinear affective ranking model. Two research challenges of highlight ranking are addressed, namely affective feature extraction and ranking model construction. The basic principle of affective feature extraction in our work is to extract sensitive features which can stimulate user's emotion. Since the users pay most attention to player behavior and audience response in racket sport highlights, we extract affective features from player behavior including action and trajectory, and game-specific audio keywords. We propose a novel motion analysis method to recognize the player actions. We employ support vector regression to construct the nonlinear highlight ranking model from affective features. A new subjective evaluation criterion is proposed to guide the model construction. To evaluate the performance of the proposed approaches, we have tested them on more than ten-hour broadcast tennis and badminton videos. The experimental results demonstrate that our action recognition approach significantly outperforms the existing appearance-based method. Moreover, our user study shows that the affective highlight ranking approach is effective.**

***Index Terms*—Action recognition, affective analysis, highlight ranking, semantic analysis, sports video analysis.**

I. Introduction

WITH the explosive growth in the amount of video data, entirely manual annotation of huge video databases is neither feasible nor appropriate. This trend has resulted in growing research interest in automatic content-based video analysis. From a general user's point of view, efficient video archiving facilitates the convenient content browsing to easily access the highlight segments in the media stream. Nevertheless, automatic highlight analysis for video summarization is still far from satisfactory in terms of semantic gap between the richness of user semantics and the simplicity of perceptual features of video data.

Two major issues are included in highlight analysis for video summarization [10], which are 1) how to extract highlight content and 2) how to organize extracted content into limited display duration or space. One of the possible solutions to the first issue is highlight extraction which has been extensively studied [2], [4]–[7], [37], [38]. With the extracted highlight content, the second issue is emphasized on organizing the structure of video summary to adapt for the user preference. In terms of the general experience for digital multimedia broadcasting, most users prefer to first browse the more interesting scenes rather than insipid content, and sometimes prefer to view the most interesting highlights due to the device capacity and time limitation. There exists a compelling case to organize the highlight content in a ranking manner preferably according to highly personalized requirement and affective criteria. Endowing an automated system with such affective ranking capability for video summarization will lead to exciting applications that enhance existing video retrieval systems. However, immediately related work in affective highlight ranking is little. This is mainly due to the difficulty of bridging the affective gap, especially in the case where high-level semantics are labeled by low-level perceptual cues of video content [11].

In this paper, we propose an affective highlight analysis approach for racket sports [12]. As an important video document, sports video has attracted increasing attention in automatic video analysis [1]–[9] due to its wide viewership and tremendous commercial potential. Inside the category of sports games, a distinction can be generally made between time-constrained games and score-constrained games. Sports like soccer, basketball, and football are time-constrained in the sense that there is relatively loose structure in the progress of the game and have distinct highlights e.g., goal in soccer. Contrastively, the structure of score-constrained games, such as racket sports like tennis and badminton, is highly formulaic with the result that such kind of sports video is composed of a restricted number of typical scenes producing a repetitive pattern. In racket sports video, there is one dominant camera view (e.g., in-play shot in tennis) that contains almost all the semantically meaningful content (scoring events). All the scoring events can be aggregated as the highlight content. Consequently, it is more crucial for racket sports to effectively rank and structure

Manuscript received August 12, 2006; revised February 8, 2007. This work was supported in part by National Hi-Tech Development Program (863 Program) of China under Grant 2006AA01Z117, in part by "Science100 Program" of the Chinese Academy of Sciences under Grant 99T3002T03, in part by the Beijing Natural Science Foundation under Grant 4063041, and sponsored by NEC Research China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jie Yang.

G. Zhu, W. Gao, and H. Yao are with the School of Computer Science, Harbin Institute of Technology, Harbin, China, 150001 (e-mail: gyzhu@jdl.ac.cn; wgao@jdl.ac.cn; yhx@vilab.hit.edu.cn).

Q. Huang and L. Xing are with the Graduate School of Chinese Academy of Sciences, Beijing, China, 100039 (e-mail: qmhuang@jdl.ac.cn; lyxing@jdl.ac.cn).

C. Xu is with the Institute for Infocomm Research, Singapore, 119613 (e-mail: xucs@i2r.a-star.edu.sg).

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

Digital Object Identifier 10.1109/TMM.2007.902847

the highlight segments to match the user's personalized query and summarize them to fit within an allocated browsing time. Unfortunately, few efforts have been devoted to this problem.

Humans perceive the impressive degree of highlight content via the nature of emotions. Yet it is a difficult task to create automatic methods for highlight evaluation. The main challenge lies in the extraction of audio-visual cues from video in the affective context, namely affective features which can effectively reflect the highlight impression from human perception. Though low-level feature has the advantages of easily computing and wide range of applications, it cannot discover the deep insight of highlight at the affective level. The mid-level representation, e.g., player's action or trajectory which is generated based on low-level features resorting to learning and recognition approaches, characterizes the semantic concepts of video content in a certain extent. As an analogy to bridge the semantic gap using mid-level cues, we can extract the affective features from mid-level video representation to overcome the existence of affective gap. In addition, ranking model is the most important element in the highlight organization scenario. From machine learning point of view, ranking process can be treated as the simulation of human's evaluation on the highlight impression from subjective perception. Thus, we can construct the ranking model using supervised learning methods.

In this paper, we develop a multimodal approach grounded on human behavior analysis to organize the highlight segments extracted from the broadcast racket sports video using nonlinear affective ranking model. Fig. 1 illustrates the flowchart of our approach. Low-level visual features are used to detect the in-play shots describing the scoring events, which are the highlight content of racket sports video. As sports highlights represent atomic entities at the semantic level, low-level features cannot capture the luxuriant insight of the highlight content. The moving object in sports video is one kind of effective mid-level representation for video content. In our approach, human behavior including the player's actions and trajectories performed in the match is analyzed and exploited to form the visual mid-level features for one in-play shot. Most techniques available in the literature for moving object based sports video analysis have focused on tracking the movement trajectory of the player and ball [13]–[15]. As one primary work, we propose a novel motion analysis method to recognize the player actions in the in-play shots of racket sports video. Since the content of video is intrinsically multimodal, a set of game-specific audio keywords are generated as the complementary of mid-level representation from auditory modality. For each in-play shot, affective features are extracted from action-trajectory-audio representation as the highlight attributes in the affective context. The basic principle of the affective feature extraction is to extract sensitive features which can easily stimulate user's emotion. Action and trajectory reflect the player behavior exhibited in the game, which are the most attractive focus paid attention by the user. On the other hand, the audio effects responded by audience in the video present the human perception about the highlight degree of video content. Finally, we conduct action-trajectory-audio based affective highlight ranking. The extracted affective features are exploited as the input of ranking model and the output of the model is the automatic estimation of impressive confidence for the in-play shot. Such highlight ranking organizes the highlight content with a hierarchical structure according to the ranking confidence and facilities two personalized content browsing fashions for broadcast racket sports video: scalable summarization and hierarchical browsing.

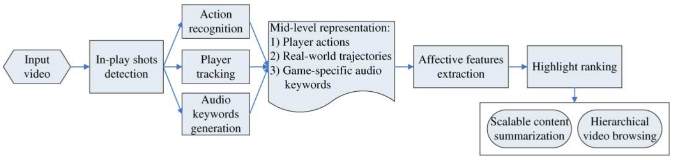
Input video
In-play shots detection
Action recognition
Player tracking
Audio keywords generation
Mid-level representation:
1) Player actions
2) Real-world trajectories
3) Game-specific audio keywords
Affective features extraction
Highlight ranking
Scalable content summarization
Hierarchical video browsing

Fig. 1. Flowchart of multimodal approach for affective highlight ranking in broadcast racket sports video.

The rest of the paper is organized as follows. Section II introduces the existing work devoted to sports video analysis. Taking tennis game as an example in Section III, we present the motion based player action recognition approach in the broadcast racket sports video. Section IV describes the multimodal approach of affective highlight ranking and depicts that this ranking approach facilities the scalable summarization and hierarchical browsing of racket sports video. In Section V, experimental results are reported and analyzed. Meanwhile, the experiments demonstrate that our action recognition approach can be easily extended to other racket sports domains. Finally, we conclude the paper in Section VI.

II. Related Work

In this section, we review the state-of-arts in sports video analysis according to three hierarchical layers for current research routine which are low-level, mid-level, and high-level analysis respectively. To distinguish our work from other related work, we also discuss human behavior recognition in the broadcast sports video and its role on semantic and affective video analysis.

A. Three Hierarchical Levels for Sports Video Analysis

1) Low-Level Processing Based Analysis: Low-level features have been widely used for sports video analysis and various approaches have been proposed on structure analysis, event detection and summarization in sports video. Low-level features can be extracted from different modalities (e.g., audio and visual) in the video. These features can be used separately or combinatorially. For example, visual features including dominant color ratio, shot boundary, and aspect ratio of referee region were employed to accomplish various events detection and summarization for soccer game [2]. Audio features including short time energy, Mel-frequency cepstral coefficients, and speech pitch were applied for semantic analysis of sports game [4]. The combination of audio and visual features was used for sports video analysis in [18]. However, it is difficult to robustly and accurately detect events using low-level features only due to the semantic gap between low-level features and high-level events.

2) Mid-Level Representation Based Analysis: Generic low-level features only deal with representing perceived content, but not with semantics. To bridge the semantic gap between low-level features and high-level events, one possible solution is to introduce a middle level representation. In [22], a mid-level feature layer was introduced in the framework of event detection for sports video. Various approaches have been proposed from the aspect of visual [22], [23], audio [19], [24] and text [6], [25], [53], respectively. The objects in sports video can be considered as an effective mid-level representation to facilitate semantic analysis. For example, the actions performed by players in tennis game reveal the process of the game and the tactics of the players [26]. The movement of players in the video provides useful information for analysis and understanding of the game [13], [14].

3) High-Level Analysis Based on Multimodality: The integrated use of various information sources is the trend in high-level analysis of sports video. With the enhancement of more information available, the result of sports video semantic analysis will improve when a multimodal approach is applied. Two categories can be differentiated for the approaches of high-level analysis: rule-based and model-based. For example, heuristic rules were exploited to combine the game-specific audio keywords, camera motion patterns, and homogeneous regions to detect events in tennis video [19]. Beyond heuristic rules, statistic models are frequently used for multimodal semantic analysis, e.g., hidden Markov model [18], support vector machine [22], and finite state machine [5]. Different from semantic analysis, affective content understanding of semantic events brings another dimension to content-based browsing resorting to the ideology of affective computing [35]. Affective computing seeks to provide better interaction with the user by understanding the user's emotional state and responding in a way which influences or takes into account the user's emotions [39]. Some methods based on affective computing for highlight extraction have been proposed [36]–[38]. With the application of affective computing in sports video, highlight ranking [20], [21] can be conducted. Highlight ranking extracts the affective cues from video to describe the excitement of content sequence. In [20], time duration of five semantic scenes in the broadcast soccer video were derived as affective features. In [21], six affective features of scoring event were extracted mainly from mid-level audio keywords and the supervised highlight model was exploited to evaluate the confidence of excitement.

B. Human Behavior Recognition in Broadcast Footage

Most existing approaches for moving object based sports video analysis have focused on tracking the movement trajectory of the player and ball [13]–[15]. Sudhir *et al.* [13] exploited the domain knowledge of tennis video to develop a court line detection algorithm and a player tracking algorithm for the purpose of automatically generating high-level annotation of play events. In [14], a real time tracking approach for player and ball in tennis game was presented. This work was based on specific-set camera system and cannot be straightforwardly applied to the broadcast video. Unlike object-based algorithm, Yu *et al.* [15] proposed a trajectory-based algorithm to detect and track the ball in the broadcast tennis video.

As another compelling behavior of players, their actions also attract the audience attention. However, little work has been concentrated on player action recognition and its applications to sports video analysis. Considering a cinematographic point of view, we can classify the shots in the broadcast racket sports video into three classes as shown in Fig. 2: close-up, medium shot, and long shot. Among these categories, medium shot and long shot indicate the playing process of the game being more focused by user. In medium shot, the magnitude of player figure is higher which is usually 300 pixels tall in CIF frame format (352 × 288). It is easy to segment and label human body parts resulting in a stick figure. Existing work [27], [28] has achieved good results on action recognition for medium shot. On the other hand, in a long shot, the camera covers a large part of the sports arena so that the player figure is smaller. It might be only 30 pixels tall in CIF frame. The action detail of the player is blurred due to the low figure resolution. It is very difficult to articulate the separate movements of different body parts, thus we cannot discriminate an action among many categories. To the best of our knowledge, there are few efforts devoted in the research of player action recognition in a long shot of the broadcast racket sports video.

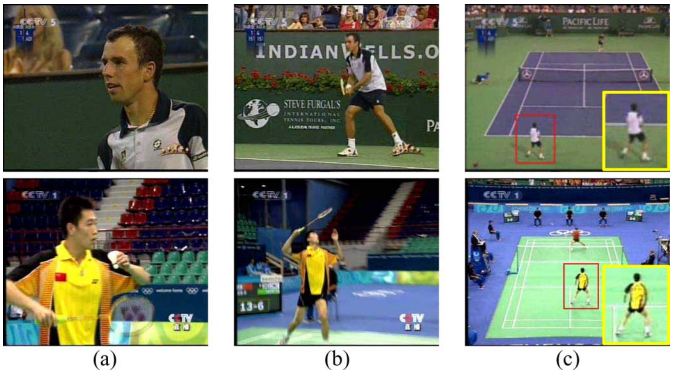

Fig. 2. Typical shots derived from broadcast racket sports video: (a) close-up shot, (b) medium shot, and (c) long shot. The zoomed picture is the player whose action to be recognized.

The key challenge of action recognition and classification is to find the proper motion representation. Various representations have been proposed such as motion history/energy image [29], spatial arrangement of moving points [30], etc. However, these motion descriptors require strict constraints such as multiple or/and stationary cameras, static background, and reasonable high resolution human figure in the video. These approaches are not suitable for the data considered in our work. Compared with the action recognition for the videos with high resolution figures, a little work [16], [17], [31]–[34] is attempting to analyze poor quality and nonstationary camera footage. Most of work is based on the appearance representation. The approach proposed in [31] modeled the structure of the appearance self-similarity matrix and was able to handle very small objects. Unfortunately, this method was based on periodicity and thus restricted to periodic motion. Efros *et al.* [32] developed a generic approach to recognize actions in the long shot. In their approach, a motion descriptor in a spatio-temporal volume was introduced and an associated similarity measure was utilized in the nearest neighbor classification framework for action categorization. However, the tennis videos they used are nonbroadcast which has less challenge for recognition. Roh *et al.* [33] presented a method for player gesture recognition using curvature scale space template of human silhouette and sequence matching algorithm. But only "serve" action in tennis game was involved in the experiment. A template-based algorithm to track and recognize player's actions in hockey and soccer sequences was proposed in [34] which the player was represented by the grids of histograms of oriented gradient and hidden Markove model was used as the classifier. For the research of sports video analysis, Miyamori *et al.* [16] developed an appearance-based annotation prototype system for tennis actions including overshoulder-, foreside-, and backside-swings. The recognition was based on silhouette transition. The original work was extended in [17] by using an integrated reasoning module with information about player and ball positions. However, since there are many variations of silhouette appearances in terms of the orientation of human posture, direction of camera, and insufficient resolutions of silhouette, etc., the appearance descriptor is not robust and discriminative for action recognition and classification. Moreover, the camera motion existing in the broadcast video can also make extracting player's appearance hard, and low resolution makes matching player's action to trained models unstable.

C. Our Contribution

In this paper, we present a new human behavior based approach for highlight analysis of racket sports video. For racket sports, players are easier to attract the attention of audience and their actions are more compelling. Compared with the simple temporal features [20] and single audio modality features [21] used as affective representation, player behavior such as action and trajectory is more objective to reflect human affective perception for the match to a large extent. However, player behavior has not been taken into account in the previous approaches of affective analysis for sports video. Furthermore, as an effective mid-level representation, player behavior provides an important cue for semantic analysis. Therefore, if player behavior can be incorporated into the affective analysis, it will not only improve the result of highlight extraction but also enhance the result of affective ranking.

Part of our work was published in [26]. Some prominent amelioration has been made in this paper. First, the motion representation scheme is improved and a more elaborate descriptor is proposed. Such improvement not only reduces the dimensionality of motion descriptor effectively, but also involves more semantic actions in the recognition framework. Our action recognition approach is consequently extended from original tennis to racket sports with the inclusion of badminton game. Secondly, an improved homography technique by integrating global motion estimation is employed to compute the real-world trajectory, which is one of the components of affective features for highlight analysis. Thirdly, audio mid-level features are exploited to represent the affective information of video content. The audio affective features are considered within the affective model to improve the ranking accuracy. Finally, a new subjective evaluation criterion is proposed for the guidance of ranking model construction and is exploited to compare the performance of subjective (human) and automatic (computer) highlight ranking.

The main contributions of our work include the following.

1) A novel player action recognition method is proposed based on motion analysis which is different from existing appearance-based approach. We characterize the motion within the human-centric figure by a new descriptor based on computing the optical flow, projecting it onto a number of motion channels, and smoothing each component.
2) In our approach, the optical flow is derived and treated as spatial patterns of noisy measurements instead of traditional precise pixel displacements [32]. Based on this idea, a new motion descriptor is proposed. Recognition is performed in a supervised learning framework.
3) We propose a new multimodal approach for highlight ranking of broadcast racket sports video. Different from previous work, our approach combines player action recognition with other multimodal features including trajectories of player and audio keywords of game-specific sounds.
4) Based on the video mid-level representation, we extract a set of affective features as the quantitative description for the attractive confidence of each in-play shot which is corresponding to the rally event at high-level semantic content. The affective features are inputted to nonlinear ranking model constructed by support vector regression. The highlight ranking for rally events is conducted to facilitate a scalable content summary of racket sports and provide a hierarchical browsing fashion for users.

III. Motion Based Player Action Recognition in Broadcast Racket Sports Video

Existing approaches for action recognition in broadcast racket sports video are based on appearance analysis [16], [33] or reasoning scenario with the integration of player and ball positions [17]. The primitive features exploited cannot be preserved among different videos. To solve feature-inconsistent problem, we develop a novel action recognition algorithm based on motion analysis. Two research challenges, motion representation and action recognition, are addressed. The algorithm contains three modules as shown in Fig. 3. 1) The algorithm starts by player tracking and human-centric figure computation. 2) Optical flow computed on the difference images of adjacent human-centric figures is derived as low-level feature. A novel motion descriptor is then calculated based on grid partition with the relationship between human body parts and optical flow field regions. 3) The action recognition is carried out in the framework of support vector machine. Voting strategy is utilized for action clip recognition by aggregating the frame classification over the temporal domain. Overhead-swing, left-swing and right-swing are recognized by our approach. Exampling with tennis which is one of the most popular racket sports, we will discuss the proposed approach in detail in the following subsections. Our action recognition approach can be easily extended to other similar racket sports.

A. Player Tracking and Stabilization

The camera in broadcast video is not static and its movement will mislead the motion feature extraction. Our recognition algorithm starts by player tracking and human-centric figure com-

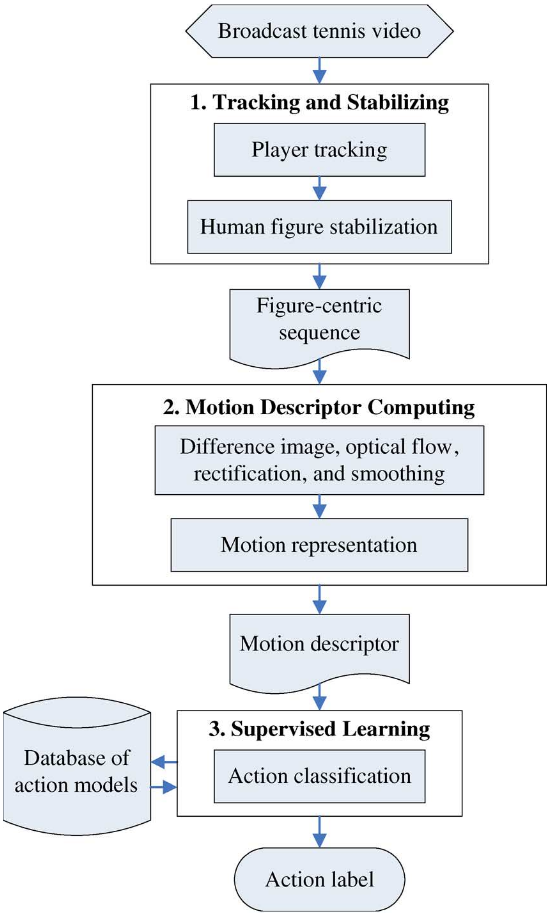
Broadcast tennis video
1. Tracking and Stabilizing
Player tracking
Human figure stabilization
Figure-centric sequence
2. Motion Descriptor Computing
Difference image, optical flow, rectification, and smoothing
Motion representation
Motion descriptor
Database of action models
3. Supervised Learning
Action classification
Action label

Fig. 3. Flow diagram of action recognition approach in broadcast racket sports video.

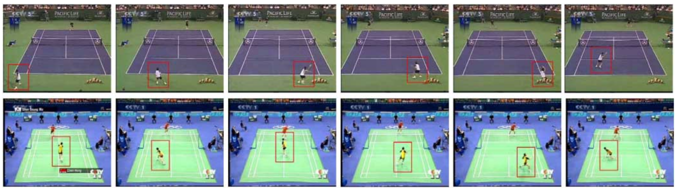

Fig. 4. Results of player tracking and stabilization, the rectangle region represents the stabilized human-centric figure.

putation for the purpose of eliminating the camera motion. Such process can be achieved by tracking the player region through the frame sequence. The detection algorithm in [13] is used to extract the player's initial position as the input of the tracker.

Existing methods for player tracking in racket game are based on template matching [13], [14], [16], [17], [32]. Such trackers are widely known for having the drifting problem because errors accumulate quickly over the time. These trackers are sensitive to the noise in the broadcast video such as player deformation and background clutter, which is unable to track the player for a long video sequence. This can be exemplified in [13] that the input video was first segmented into chunks of 30 frames and then tracking for each chunk was conducted separately. A sophisticated tracking strategy called support vector regression (SVR) particle filter [40] is employed in our approach. SVR particle filter enhances the performance of the classical particle filter with small sample set and is robust enough for the noisy circumstance in the broadcast video. More details about this tracker and its comparison with the classical particle filter can be found in [40].

To derive the human-centric figure, the tracking window around the player region is enlarged by a scale in pixel unit and a simple method of computing the centroid of player region is used. The centroid coordinates of the region are defined as follows:

$$m_x = \sum_{x\in R}\sum_{y\in R} x \cdot f(x,y) / \sum_{x\in R}\sum_{y\in R} f(x,y). \tag{1}$$

$$m_y = \sum_{x\in R}\sum_{y\in R} y \cdot f(x,y) / \sum_{x\in R}\sum_{y\in R} f(x,y). \tag{2}$$

where R is the region occupied by the object in the image plane and $f(x,y)$ the gray level at location (x,y). Then, the center of the window controlled by the tracker is shifted to the position (m_x, m_y).

Once the video sequence is stabilized, the motion in broadcast video caused by camera behavior can be treated as being removed. This corresponds to a skillful movement by a camera operator who keeps the moving figure in the center of the view. Any residual motion within the human-centric figure is due to the relative motion of different body parts such as limbs, head, torso and racket griped with player. Some results of tracking and stabilization are illustrated in Fig. 4.

B. *Motion Descriptor Computation*

We derive motion features from pixel-wise optical flow which is the most natural technique for capturing motion independent of appearance. The key challenge is that computation of optical flow is not very accurate, particularly on coarse and noisy data such as broadcast video footage. The insight of our approach is to treat optical flow field as spatial patterns of noisy measurements which are aggregated using our motion descriptor instead of precise pixel displacements at points. Within the human-centric figure, the motion is due to the relative movements caused by player's different body parts which are the different regions mapped into the image plane. These motion characteristics cannot be well captured by global features computed from the whole figure. A simple means of localizing the motion for recognition is to separately pay attention to different regions around the human torso. In our approach, we divide the optical flow field into various subregions called grids. The histogram is utilized to represent the spatial distribution for sub optical flow field in each grid.

1) Optical Flow Computation and Noise Elimination: The noise in figure background makes significant influence for the accurate computation of optical flow for the player region. It necessitates background subtraction. Considering the background of human-centric figure is a playfield, an adaptive method of playfield detection [41] is applied to subtract background with the post-processing of region growing [42]. The flowchart of the background subtraction is shown in Fig. 5.

The extraction of motion descriptor is based on the optical flow in the difference image of two adjacent human-centric figures. Although it is possible to directly use human-centric

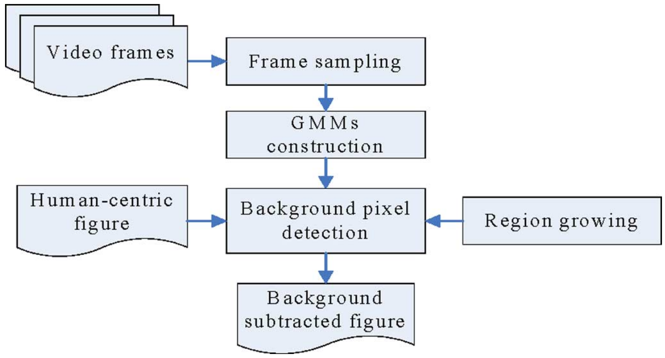

Fig. 5. Adaptive background subtraction for human-centric figure.

figures, it is undesirable to do so for a number of reasons. First, the human-centric figures might have significantly different brightness. This is the result of automatic camera gain, sudden camera flashes, or view point changes. Such factors will mislead the optimization process in the optical flow computation algorithm to output an unreliable vector field. The difference component of two adjacent figures eliminates the influence caused by these reasons. Secondly, the difference image reflects the relative moving parts in adjacent figures, which essentially represents the motion borders of the player. In biological vision, human neurons are more sensitive to direction and speed of motion border. Therefore, the optical flow field computed on difference image reflects the motion insight in human-centric figure more effectively. We compute optical flow using Horn-Schunck algorithm [43]. This process is described as follows:

$$\begin{aligned} DI_i &= HC_i - HC_{i-1}, \\ \mathbf{OFF}_i &= \mathrm{HS}(DI_i), \quad i = 2, \ldots, N \end{aligned} \tag{3}$$

where DI_i is the difference image computed from adjacent human-centric figures HC_i and HC_{i-1} after background subtraction, N is the total figure number, $\mathrm{HS}(\bullet)$ refers to Horn-Schunck algorithm, and $\mathbf{OFF}_i$ is the optical flow field of DI_i.

Half-wave rectification and Gaussian smoothing are applied to eliminate the noise in the optical flow field. The process is shown in Fig. 6. The optical flow magnitudes are first thresholded to reduce the effect of too small and too large motion probably due to the noise inside the human region. The optical flow vector field **OFF** is then split into two scalar fields corresponding to the horizontal and vertical components OFF_X and OFF_Y, which are then half-wave rectified into four nonnegative channels $OFF_X^+, OFF_X^-, OFF_Y^+$, and OFF_Y^-, where they satisfy $OFF_X = OFF_X^+ - OFF_X^-$ and $OFF_Y = OFF_Y^+ - OFF_Y^-$. They are each smoothed by a Gaussian filter. Thus, the noise in the original field is eliminated and the refined optical flow field is obtained.

The general idea of half-wave rectification is to make the data sparse so that significant smoothing can be applied. Smoothing the data is important for two major reasons. First, it reduces the amount of noise in the motion data. Secondly, it helps to match motion flow that is not perfectly aligned in position or normalized in scale.

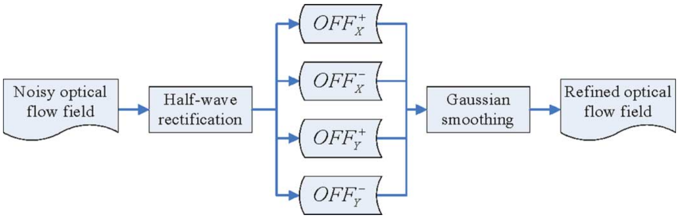

Fig. 6. Half-wave rectification and Gaussian smoothing for noise elimination of optical flow field.

2) *Local Motion Representation:* With the context of action recognition, the motion in the human-centric figure is due to the relative movement of different body parts which is exhibited in the different figure regions. Thus, the distribution of optical flow field in the figure is pockety for each swing type. This can be demonstrated by observing the optical flow computed from the processed human-centric figure [see the sub-insets in Fig. 7(a)]. For overhead-swing, the dense regions of the optical flow field are mainly distributed in the top part of the human-centric image. For left-swing, the optical flow field in the left figure is much denser than the field in the right region. Contrarily, the field in the right region is denser than that in the left region for right-swing. Such evidence can be further distinctly represented by our motion descriptor—grid based optical flow histograms. We adopt a simple but effective region style called grid in our approach. The whole optical flow field is divided into nonoverlapping grids along the width and height orientation respectively as shown in Fig. 7(b). Considering the structure of the player occupied in the human-centric figure, the 3 × 3 grid style is employed in our approach which is enough to reflect the distribution of different body parts in the spatial space. More complex partition (e.g., 5 × 5, 7 × 7) will make each grid too small and result in the sparse histogram representation.

Motivated by grids of histograms of oriented gradient [44] and kernel density estimation for color distribution [45], we derive a group of grid based optical flow histograms (G-OFHs) as the motion descriptor for swing action. Let $G_x(\mathbf{p})$ and $G_y(\mathbf{p})$ denote the x (horizontal) and y (vertical) components of the optical flow vector $\mathbf{f}$ at location $\mathbf{p}$, respectively. The magnitude $M(\mathbf{p})$ and the orientation $\theta(\mathbf{p})$ of the flow vector are computed by

$$M(\mathbf{p}) = \sqrt{G_x^2(\mathbf{p}) + G_y^2(\mathbf{p})}. \tag{4}$$

$$\theta(\mathbf{p}) = \tan^{-1}\left[G_y(\mathbf{p})/G_x(\mathbf{p})\right]. \tag{5}$$

The essence of G-OFHs is that, for each grid, we quantize the orientation $\theta(\mathbf{p})$ for all the flow vectors into m orientation bins weighted by their magnitude $M(\mathbf{p})$ meanwhile considering the information of their neighboring vectors with the assistance of kernel function.

We define $b(\mathbf{p}) \in \{1, \ldots, m\}$ as the bin index of histogram associated with the optical flow vector $\mathbf{f}$ at location $\mathbf{p}$. For each position $\mathbf{q}$ inside the optical flow field **OFF**, considering a block region $R(\mathbf{q})$ centered at $\mathbf{q}$, the probability of bin $u = 1, \ldots, m$ in the histogram of **OFF** is then calculated as

$$h_u = C \cdot \sum_{\mathbf{q} \in \mathbf{OFF}} \sum_{\mathbf{p} \in R(\mathbf{q})} k\left(\|\mathbf{q} - \mathbf{p}\|\right) \cdot M(\mathbf{p}) \cdot \delta\left[b(\mathbf{p}) - u\right] \tag{6}$$

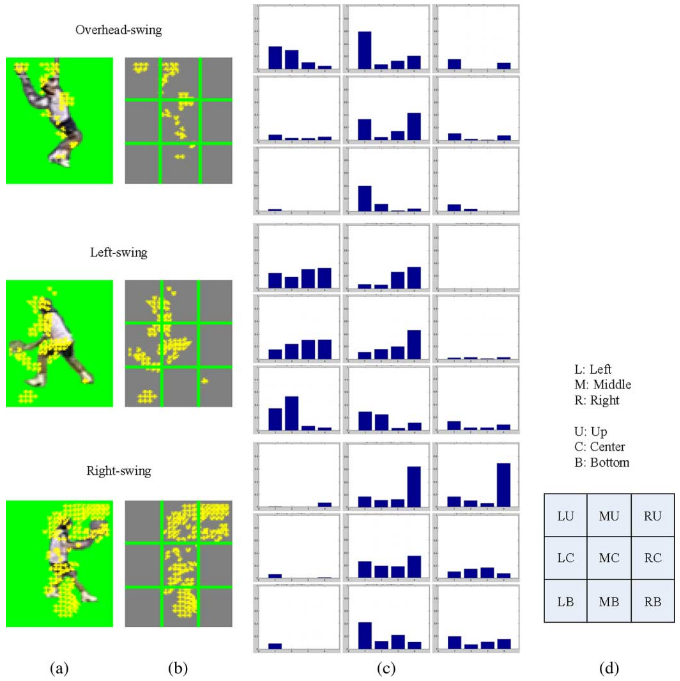

Fig. 7. Grid partition and grid based optical flow histograms (G-OFHs) for three swing actions. In (c), the position of histograms in sub-inset for each swing action is corresponding to the layout of grid partition in (b). Sub-figure (d) shows the abbreviated label for each region of grid partition. (a) Refined optical flow; (b) grid partition; (c) grid based optical flow histograms (G-OFHs); and (d) label for each grid region.

where δ is the Kronecker delta function, C is the normalization constant ensuring $\sum_{u=1}^{m} h_u = 1$, and k is a convex and monotonic decreasing kernel profile which assigns a smaller weight to the locations that are farther from the center $\mathbf{q}$. Given a refined optical flow $\mathbf{OFF}_i$ for figure F_i in the human-centric figure sequence, $i = 1, \ldots, N$ where N is the total figure number, $\mathbf{OFF}_{i,j}$ is the sub optical flow field in the grid, $j = 1, \ldots, L$ and here $L = 9$. The $\mathrm{G{-}OFH}_{i,j}$ is defined as follows according to (6)

$$h_u^{i,j} = C \cdot \sum_{\mathbf{q}\in\mathbf{OFF}_{i,j}} \sum_{\mathbf{p}\in R(\mathbf{q})} k\left(\|\mathbf{q}-\mathbf{p}\|\right) \cdot M(\mathbf{p}) \cdot \delta\left[b(\mathbf{p}) - u\right]. \tag{7}$$

With our empirical observation, the histogram can be built without explicitly computing the orientation of the optical flow vector. Instead we use the normalized horizontal and vertical strengths $g_x(\mathbf{p}) = G_x(\mathbf{p})/M(\mathbf{p})$ and $g_y(\mathbf{p}) = G_y(\mathbf{p})/M(\mathbf{p})$ to index the optical flow vector into m bins. The algorithm gives satisfactory result even when m is as small as 4. Thus, nine optical flow histograms are constructed for each human-centric figure. Fig. 7(c) shows the G-OFHs for overhead-swing, left-swing, and right-swing, respectively.

Most of the motion in the human-centric figure is caused by the relative movement of human body parts around the torso such as limbs and racket griped by the player. Using the abbreviate labels for grid regions as shown in Fig. 7(d), the region MC is corresponding to the torso of human body where the G-OFH does not reflect the insight of motion distribution. Therefore, the grid MC is not used to calculate the motion descriptor. In our approach, eight grids including LU, LC, LB, MU, MB, RU, RC, and RB are selected for computing the G-OFHs. Eight G-OFHs for one figure are ultimately utilized as the motion descriptor. From Fig. 7(c), we can see that G-OFHs can effectively capture the discriminative features for different actions in spatial space.

In our previous work [26], we employed slice based optical flow histograms (S-OFHs) as the motion descriptor. The optical flow field was partitioned into three slices only along the width orientation. For each slice, two histograms were constructed over the quantized vector magnitudes for horizontal and vertical orientation respectively. Left and right slices were selected for computing the S-OFHs. Four S-OFHs were ultimately utilized as the motion descriptor for one human-centric figure. The number of quantization level was 15 in our previous work, which resulted in a 60-dimension descriptor for one action figure. Compared with S-OFHs, G-OFHs descriptor exploits a more elaborate partition scheme to take the local motion spatial patterns into account. As above presentation, 32-dimension vector is generated as the motion descriptor. Such improvement not only reduces the dimensionality of motion descriptor effectively, but also involves more semantic actions in the recognition framework. Consequently, our action recognition approach is able to extend from original tennis to racket sports with the inclusion of badminton game.

C. *Supervised Learning Based Action Classification*

We formulate action recognition as a classification task. Supervised machine learning can be exploited as the solution for this problem. Various supervised learning algorithms can be employed to train an action recognizer. Support vector machine (SVM) [46] is used in our approach. SVM has been successfully applied to a wide range of pattern recognition and classification problems. The advantages of SVM over other methods consist of: 1) providing better prediction on unseen test data, 2) providing a unique optimal solution for a training problem, and 3) containing fewer parameters compared with other methods. We employ v-Support Vector Classification (v-SVC) [47]. Parameter v is the lower bound on the fraction of support vectors or the upper bound on the fraction of margin errors equivalently. Compared with traditional SVC algorithms, parameter v is more convenient and intuitive to be initialized. The concatenation of eight G-OFHs for each optical flow field in the figure of one video frame is fed into support vector machine. The radial basis function (RBF) kernel $K(x, y) = \exp(-\lambda \cdot \|x - y\|)$ is utilized to map training vectors into a high dimensional feature space for classification. In the experiments, we set $v = 0.3$ and $\lambda = 1/dim$ where $dim = 32$ is the dimension of the input feature vector.

Based on frame recognition and temporal voting strategy, the action clips are classified into three categories: overhead-swing, left-swing and right-swing. We use audio features [19] to detect the sound of hitting ball to locate the action clip in the video. As shown in Fig. 8, the frame corresponding to the occurrence of hitting ball is called hitting point. The adjacent window before hitting point is selected as the action clip. The window length is empirically set to be 25 frames in our experiments. First, we define set $T = \{1, \ldots, M\}$ to label each action category with $t \in T$, where M is the number of categories and here $M = 3$.

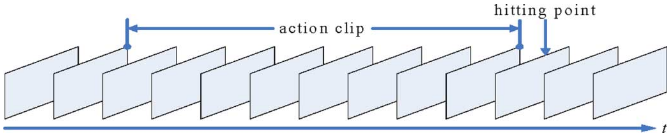

Fig. 8. Location of action clip in broadcast racket sports video.

Given f_i which is the ith frame in the action clip V, the corresponding human-centric figure is hc_i, $i = 1, \ldots, N$, N is the total number of frames in the clip and here $N = 25$. Then, the final recognized action category of V is determined as

$$Vote(t) = \sum_{i=1}^{N} \delta\left[\mathrm{Reg}(hc_i) - t\right]. \tag{8}$$

$$Category(V) = \underset{1 \le t \le M}{\arg\max}\left[Vote(t)\right]. \tag{9}$$

where δ is the Kronecker delta function, and $\mathrm{Reg}(\bullet) \in T$ refers to our recognition approach worked on the frame. Note that sound itself has a continuous existence and possible misalignment between video and audio stream may occur, thus we exploit sliding window technique to vote the action type from a sequence of frame-based classification results. The skip of sliding window is set to be eight frames in the experiments.

IV. Highlight Ranking for Broadcast Racket Sports Video

In this section, we propose a novel multimodal approach of highlight ranking for the broadcast racket sports video by integrating action recognition, trajectory computation, and audio analysis. The affective features are extracted from action-trajectory-audio mid-level representation to describe the excitement degree for each in-play shot. A nonlinear model is applied to evaluate the impressive confidence of rally events. With the ranking result, scalable summary and hierarchical video browsing are achieved.

A. Mid-Level Representation for Video Content

The approach of player action recognition in the broadcast racket sports video has been described in the previous section. Badminton game has similar video structure and court configuration with tennis game. Taking broadcast tennis video as an example, we present the player real-world trajectory computation and audio keywords generation.

1) Real-World Trajectory Computation in Broadcast Racket Sports Video: In racket game broadcasting, the cameras are usually located at the two ends of the court above the central line. Thus, the game court is projected to be trapezoidal in the video frames as shown in Fig. 9. Such projection leads to the distortion of player's movement in the court. Here we use an improved planar projection method based on homography technique [48] to correct the projective distortion and calculate the player real-world trajectory.

Using homogeneous representation, a three-dimension vector $\mathbf{x} = (x, y, w)^T$ can be used to represent a point $(x/w, y/w)^T$ in Euclidean two-dimension space. Since sport court can be assumed to be a planar, the mapping from the world coordinate system to the image coordinate system can be described by a plane-to-plane mapping [48] called homography $\mathbf{H}$ which is a 3 × 3 matrix. This is an eight-parameter perspective transformation, mapping a position $\mathbf{M}$ in the court model coordinate system to the image coordinate $\mathbf{m}$. Representing position as homogeneous coordinate, the transformation is

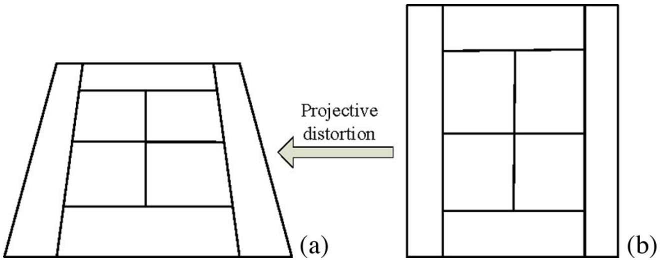

Fig. 9. Projective distortion in broadcast tennis video: (a) the tennis court in video sequence and (b) the actual tennis court.

$$\mathbf{m} = \mathbf{H} \cdot \mathbf{M}. \tag{10}$$

Equation (10) is satisfied only for the first frame in a shot and the subsequent frames when the camera is static. The mapping matrix $\mathbf{H}$ has to be updated in case of the occurrence of camera motion. Considering the camera motion in long shots of the broadcast racket video, pan is the most frequent behavior and the acceleration of motion between two successive frames is small. Therefore, we use an improved homography algorithm by integrating an update scheme for matrix $\mathbf{H}$ with the assistance of global motion estimation (GME) [49].

As shown in Fig. 10, let $\mathbf{H}_{t-1}$ and $\mathbf{H}_t$ be the homography mapping matrix for frame $t-1$ and t respectively, $\mathbf{P}_{t-1,t}$ be the GME transform matrix from frame $t-1$ to frame t. Considering the arbitrary image coordinate $\mathbf{m}_{t-1}$ in frame $t-1$, $\mathbf{m}_t$ and $\mathbf{M}$ are its correspondences in frame t and court model respectively. Without losing generality, $\mathbf{m}$ is selected from the intersections of lines in the court for the convenience of correspondence identification. Based on the theory of homography and global motion estimation, we have

$$\begin{cases} \mathbf{m}_{t-1} = \mathbf{H}_{t-1} \cdot \mathbf{M} \\ \mathbf{m}_t = \mathbf{H}_t \cdot \mathbf{M} \\ \mathbf{m}_t = \mathbf{P}_{t-1,t} \cdot \mathbf{m}_{t-1} \end{cases}. \tag{11}$$

With the elimination of $\mathbf{m}_{t-1}$, $\mathbf{m}_t$, and $\mathbf{M}$, we obtain the update function of mapping matrix $\mathbf{H}$ as

$$\mathbf{H}_t = \mathbf{P}_{t-1,t} \cdot \mathbf{H}_{t-1}. \tag{12}$$

To calculate the real-world trajectory, the player's positions in all frames within an in-play shot are first obtained by tracking module in the action recognition method. Fig. 11(a) shows such results aggregated in one representative frame. Then, we utilize improved homography algorithm to calculate the real-world trajectory which is the locus of the player viewed from planform as shown in Fig. 11(b). The final trajectory is smoothed by Gaussian filter to eliminate the noisy positions.

2) Audio Keywords Generation for Broadcast Racket Sports Video: Fig. 12 shows the flowchart of audio keywords generation for the broadcast racket sports video. Three domain-specific

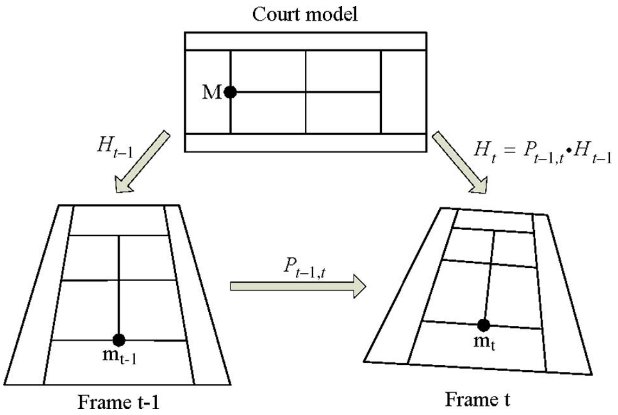

Fig. 10. Update scheme for homography mapping matrix using global motion estimation.

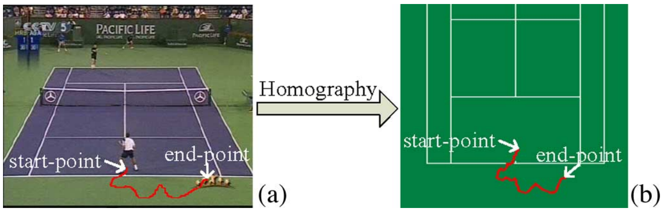

Fig. 11. Real-world trajectory computation: (a) trajectory in frames and (b) corresponding real-world trajectory.

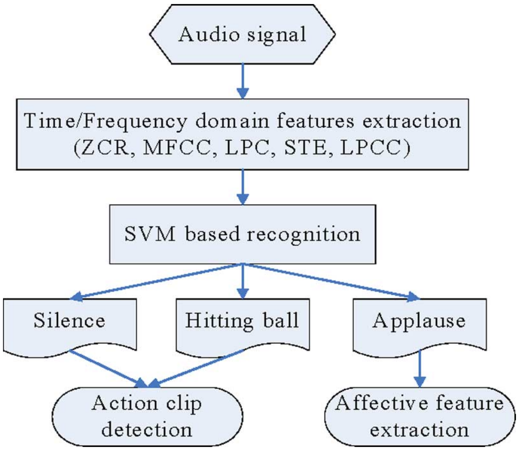
Audio signal
Time/Frequency domain features extraction
(ZCR, MFCC, LPC, STE, LPCC)
SVM based recognition
Silence
Hitting ball
Applause
Action clip detection
Affective feature extraction

Fig. 12. Audio keywords generation for broadcast racket sports video.

audio keywords are generated for racket games: silence, hitting ball, and applause. We make use of representations of the audio signal in terms of time-domain and frequency-domain measurements to construct the sound recognizer by support vector machine. These measurements include zero-crossing rate (ZCR), Mel-frequency cepstral coefficients (MFCC), linear prediction coefficient (LPC), short time energy (STE), and linear prediction cepstral coefficients (LPCC). More details about audio keywords generation can be found in [19].

The generation of audio keywords has two purposes as shown in Fig. 12. First, we use the keywords, silence and hitting ball, to detect the action clip in the in-play shots. To improve the detection accuracy, silence recognition is incorporated with hitting ball in our approach. Secondly, applause is exploited as one selection for the affective feature extraction because the audience applause happening after a score event reflects the human perception for excitement degree of the entire event.

B. Affective Attributes Production

We extract six affective features from action-trajectory-audio representation for an in-play shot.

- *Features on action*: We extract Swing Switching Rate (SSR) as an affective feature on action. Swing switching rate gives the estimation of frequency for switching of player actions among overhead-swing, left-swing, and right-swing occurred in an in-play shot. We first define the indicator function SL as

$$SL(x) = \begin{cases} 1, & \text{if } x \neq 0 \\ 0, & \text{if } x = 0. \end{cases} \tag{13}$$

Then, considering the sequence of action clips $(V_1, \ldots, V_n)$ in the shot and (9), swing switching rate is calculated as follows:

$$SSR = \sum_{i=2}^{n} SL\left[Category(V_i) - Category(V_{i-1})\right]/(n-1) \tag{14}$$

where n is the total number of the action clips.

- *Features on trajectory*: Three features are derived from trajectory description.
  - Speed of Player (SOP) which is calculated based on the length of real-world trajectory and shot duration.
  - Maximum Covered Court (MCC) which is the area of rectangle shaped with the leftmost, rightmost, topmost, and bottommost points on real-world trajectory.
  - Direction Switching Rate (DSR) which is the switching frequency for movement direction of the player in the court. For a real-world trajectory, let PS be the set of position points. We first scan the trajectory from horizontal and vertical direction to find the sets of inflexion IH and IV respectively. Then DSR is defined as

$$DSR = (\|IH\| + \|IV\|)/\|PS\| \tag{15}$$

where $\| \bullet \|$ is the cardinality of a set.

- *Features on audio*: 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. Therefore, the audio keywords—Applause is exploited to extract affective features as the response of audience from audio component of sports video.
  - Duration of Applause (DOA) which is the time duration of audience applause.
  - Applause Average Energy (AAE) which is the energy measurement for applause signal.

The affective feature vector for an in-play shot comprised $(SSR, SOP, MCC, DSR, DOA, AAE)$ is fed into the ranking model.

C. Ranking Model Construction

The objective of ranking model is to build the relationship between affective features and impressive confidence. Different from the method in [36] which used the criterion of comparability, compatibility, and smoothness for model construction, we exploit a subjective evaluation criterion to guide the ranking modeling inspired by the pairwise comparison method [50]. The model constructed by our method is nonlinear which can reflect human perception more reasonably and is more general than the observation based linear model [36].

In order to compare the subjective (human) evaluation and the automatic (computer) estimation, Peker *et al.* proposed a pairwise comparison method in [50]. However, it is not feasible to quantitatively observe the effect of each affective feature and ranking model on individual event if applying pairwise comparison straightforwardly to our work. Furthermore, the number of quantization level for subjective experiment in [50] should be defined manually whereas it is hard to be initialized in advance. We propose a subjective evaluation criterion which is superior to the pairwise comparison method. The major improvement lies in introducing the adaptive quantized highlight rank R into pairwise comparison. Different from the discrete levels of subjective experiment in [50], subjects are left free to give evaluation value for each in-play shot between 0 and 1 according to its excitement degree in our method. Instead of using fixed quantization level in [50], we define the rank candidate set $K = \{2, \ldots, M\}$, and the adaptive quantized highlight rank $R \in K$ is automatically decided by an optimal quantization process Q which maps a continuous value in [0,1] to an integer of set $\{0, \ldots, R\}$ by minimizing the quantization error.

We define the continuous ground truth of shot S_i to be $v_i \in [0,1], i = 1, \ldots, N$ where N is the number of in-play shots. Its corresponding discrete level is integer $r_i \in \{0, \ldots, R'\}$, $R' \in K$. The highlight rank R is decided by the optimal quantization process Q with the principle of minimizing the error between v_i and r_i. Such quantization process is described as follows:

$$Q(v_i) = r_i, \text{ if } r_i/R' \leq v_i < (r_i + 1)/R'. \tag{16}$$

$$err_{R'} = R' \cdot \sum_{i=1}^{N} |v_i - r_i|. \tag{17}$$

$$R = \underset{2 \leq R' \leq M}{\arg\min}(err_{R'}) \tag{18}$$

where $err_{R'}$ is the quantization error for rank R'. As long as R is determined, the continuous score for each in-play shot can be converted to the discrete highlight level with the lowest quantization error. Consequently, the subjective evaluation criterion—highlight ranking accuracy (HRA) is defined

$$HRA = \frac{1}{N} \sum_{i=1}^{N} \frac{R - |Q(v_i) - Q(c_i)|}{R} \times 100\% \tag{19}$$

where c_i is the impressive confidence scored by ranking model corresponding to v_i of shot S_i. The component $|Q(v_i) - Q(c_i)|$ in (19) represents the relative bias between highlight ranked by human and computer. Evaluation criterion (19) shows that the accuracy is obtained by averaging the human-computer rank bias. The difference of 1% in HRA means a difference of 1% in relative bias. If HRA is 80%, there is 20% difference on ranking between human and computer relatively. Therefore, the more effective features and reasonable ranking model are, the higher the ranking accuracy is.

Based on the guidance of proposed subjective evaluation criterion, support vector regression is exploited to train the ranking model. The reasons why we use SVR are due to following considerations. 1) SVR has the advantages of kernel based learning algorithms, such as minor training data needed and better expansibility for unseen test data, 2) SVR is more robust for the nonlinear/noisy data and can provide more powerful prediction capacity, and 3) This is also the most important reason that there is no research effort that can demonstrate that the model of human perception is linear. As the linear modeling is the special case of nonlinear computing technique, nonlinear SVR model is trained as the ranking model with consideration of its generality.

D. Scalable Summarization and Hierarchical Browsing of Broadcast Racket Sports Video

With the ranking result, we organize the highlight content with a hierarchical structure. Scalable content summarization and hierarchical browsing fashion are achieved.

1) Scalable Content Summarization: For a broadcast racket video, we represent arbitrary highlight h_i as a three-tuple

$$h_i = \langle id_i, td_i, hc_i \rangle \tag{20}$$

where id_i denotes the highlight index in the video, td_i is the highlight length (time duration), and hc_i corresponds to the highlight confidence. Therefore, we obtain the set of highlight segments H sorted by confidence value in descending order

$$H = \{h_i | hc_i > hc_j, \forall i < j, 1 \leq i, j \leq n\} \tag{21}$$

where n is the total number of highlights. Therefore, two kinds of scalable content summary can be constructed according to the query requirement from summary duration and summary excitement, respectively.

Due to the time limitation, the user only wants to obtain the game summary within the time duration T involving the most exciting segments. The feedback set is constructed by selecting the top m highlight segments as

$$U_T = \left\{ u_i | u_i \in H, \sum_i td_i \leq T, i = 1, \ldots, m \right\}. \tag{22}$$

On the other hand, some users want to summarize the game by the highlights with the customized impressive confidence not less than the threshold C. Then, the returned result is represented as

$$U_C = \{u_i | u_i \in H, \forall hc_i \geq C, i = 1, \ldots, k\} \tag{23}$$

where k is the number of selected highlight segments.

2) Hierarchical Video Browsing: As illustrated in Fig. 13, the highlights (rally scenes) of the rackets sports video can be formatted as a hierarchical structure according to the ranking level. Suppose that the overall highlight rank is $(0, 1, \ldots, r, \ldots, R)$ which rank 0 represents the least exciting whereas rank R represents the most exciting, and the customized browsing parameter $L = r$ is defined by user's preference, the set of selected browsing content is then represented as

$$U_L = \{u_i | u_i \in H, \forall Q(hc_i) \geq L, i = 1, \ldots, p\} \tag{24}$$

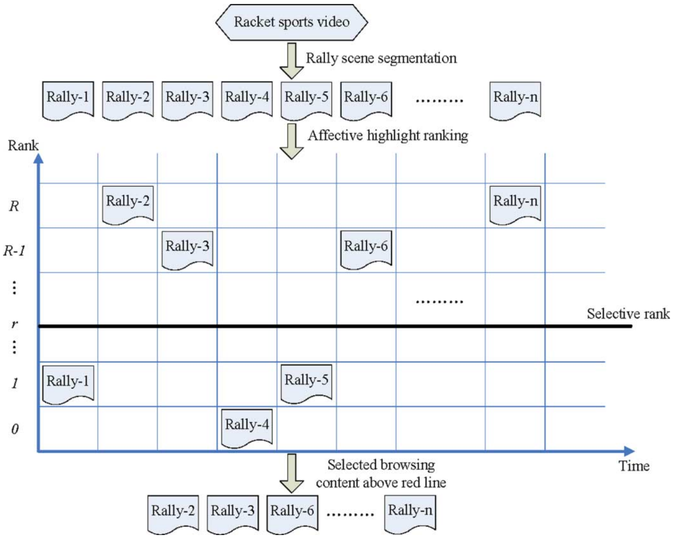

Fig. 13. Hierarchical structure for video browsing of racket sports video.

TABLE I
INFORMATION OF THE EXPERIMENTAL VIDEO DATA FOR HIGHLIGHT RANKING

Game	Data	Players	Duration (h/m/s)	# In-play shot
Tennis	Pacific_Open_2004	Agassi vs. Hrbaty	0:39:38	38
	French_Open_2005	Nadal vs. Puerta	3:34:14	243
	Australian_Open_2005	Safin vs. Hrbaty	2:12:0	207
	Dubai_Championship_2006	Nadal vs. Schuettler	1:25:29	144
Badminton	Olympic_Game_2004_A	Zhang Ning vs. Mia Audina	1:21:32	113
	Olympic_Game_2004_B	Shon Seung Mo vs. Chen Hong	1:33:52	132
Total			10:46:45	877

where H is the set of highlight segments defined in (21), $Q(\bullet)$ is the optimal quantization process defined in (16), and p is the number of selected highlight segments. Exampling with Fig. 13, all the rally scenes above the bold line are selected as the browsing content. The user can access any rally scene by its index.

V. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed approaches, we carried out experiments on tennis and badminton videos which are the most two representative games of racket sports. The test data were recorded from live broadcast television program. The data for tennis game were captured from the matches of Pacific Life Open 2004, French Open 2005, Australian Open 2005, and Dubai Championship 2006. The data for badminton were captured from Olympic Game 2004. The videos were compressed in MPEG-2 standard with the frame resolution of 352× 288. The total duration of test data is more than 10 h. As illustrated in Fig. 1, all the in-play shots were detected using the algorithm in [13], and made manual correction if necessary. The detail about the video data is listed in Table I including the opponent players, the duration, and the number of in-play shots.

A. Results of Action Recognition

To verify the effectiveness of our action recognition algorithm, we manually annotated the ground truth of the player actions in four matches with 3128 action clips including 724 overhead-swing, 1183 left-swing, and 1221 right-swing. Table II shows the detail of the test data. To quantitatively evaluate the performance, we calculated Recall (R) and Precision (P) for each action category, which are defined as follows:

TABLE II
DETAIL OF VIDEO DATA FOR ACTION RECOGNITION

Game	Data	# Overhead-swing	# Left-swing	# Right-swing
Tennis	Pacific_Open_2004	41	91	136
	French_Open_2005	249	629	270
	Austrilian_Open_2005	209	327	645
Badminton	Olympic_Game_2004_A	225	136	170
Total		724	1183	1221

$$R = n_c/(n_c + n_m). \tag{25}$$

$$P = n_c/(n_c + n_f) \tag{26}$$

where for an action category, n_c is the number of clips correctly recognized, n_m is the number of missed clips, and n_f is the number of clips false-alarmed. The Accuracy (A) metric was employed to evaluate the holistic performance, which is defined as

$$A = (n_{c-overhead} + n_{c-left} + n_{c-right})/n_{total} \tag{27}$$

where $n_{c-overhead}$, n_{c-left}, and $n_{c-right}$ are the number of clips correctly recognized of overhead-swing, left-swing, and right-swing respectively, n_{total} is the total number of action clips. The data of training and testing sets were selected using three-fold cross validation from the whole dataset with the ratios of 2/3 and 1/3. Therefore, three runs were carried out and the average performance was used as the final evaluation. Two action classifiers were constructed for tennis player and badminton player, respectively.

Table III shows the experimental results. For tennis video, our method achieves the Accuracy of 90.7%. For badminton video, the Accuracy for all the clips is 87.6%. The holistic evaluation is 90.2% with Accuracy metric. Fig. 14(a) illustrates some representative action clips accurately recognized by our approach and Fig. 14(b) shows some failure examples. The reason resulting in incorrect recognition is because the player is a deformable object of which the limbs make free movement during the action displaying. This will disturb the regular optical flow distribution to make the G-OFHs misreport the motion characteristics in the human-centric figure. In the test video data, the player figures are generally about 30 to 40 pixels tall and the detail of swing action is blurred severely. Since our proposed descriptor represents the spatial patterns of blurred movement, it is robust for the poor quality of broadcast video.

A comparison with the existing appearance-based work was also carried out. The algorithm in [16] was evaluated using the same training and testing data. Because there is no open source code for this method that can be found, we implemented the algorithm by ourselves strictly obeying the original description in [16]. Silhouette transitions are extracted from the human-centric figures. KL transform is utilized to expand silhouette features to a certain eigenspace. Different numbers of high-ranked features in the eigenspace are selected as the discriminative parameters to identify action categories in the nearest neighbor framework. We employed different percentages of eigenfeature basis and obtained the best Accuracy results when 80% is used. Table III

TABLE III
EXPERIMENTAL RESULTS OF ACTION RECOGNITION USING OUR APPROACH AND APPEARANCE-BASED APPROACH

Game	Action	# Clip	Our approach			Appearance-based approach		
			R (%)	*P* (%)	*A* (%)	*R* (%)	*P* (%)	*A* (%)
Tennis	Overhead-swing	499	90.4	93.2	90.7	76.6	69.0	71.5
	Left-swing	1047	87.8	89.9		73.4	70.5	
	Right-swing	1051	93.8	91.7		67.1	75.4	
Badminton	Overhead-swing	225	87.6	85.3	87.6	57.3	62.3	59.1
	Left-swing	136	85.3	91.3		64.0	75.0	
	Right-swing	170	89.4	92.7		57.6	66.7	
Total		3128			90.2			69.4

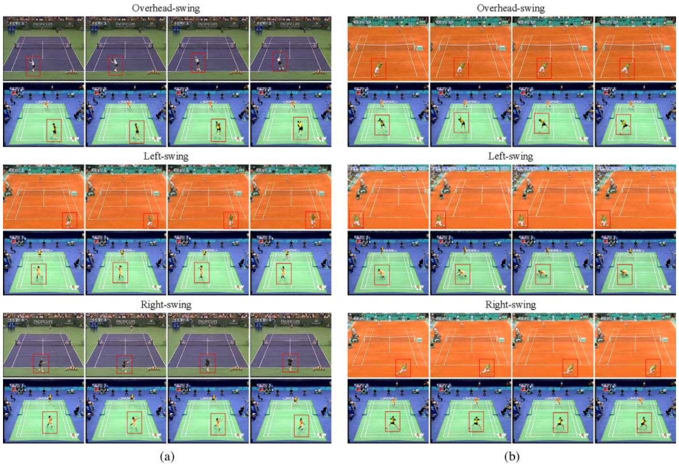

Fig. 14. Representative recognition results for overhead, left and right swing: (a) samples of correct recognition and (b) samples of failure recognition.

summarizes the comparative results. From this comparison, it can be concluded that our method significantly outperforms appearance-based algorithm because the motion descriptor is preserved better than the appearance representation and more robust for classification and recognition.

B. Results of Highlight Ranking

We conducted the experiments on five complete broadcast videos of tennis matches (French Open 2005, Australian Open 2005, and Dubai Championships 2006) and badminton matches (Olympic Game 2004 A and B).

1) User Study for Ground Truth Preparation: There is no objective measure available today to evaluate the excitement degree of highlight. To evaluate the highlights and obtain the ground truth, we employed subjective user study [51] to label the highlight confidence of in-play shots. Highlight ranking is subjective and the evaluation result depends on the subjects involved in the study. This can be demonstrated by two cases we conducted in user study. In the first case, we invited six people including three male and three female aging from 26 to 34. All the people are sports amateurs. In the second case, we invited six new people in the study including three male and three female aging from 23 to 30. Among these six people, three are tennis fans and three are badminton fans. All the people were naïve to the purpose of study and the only thing they need to do was to rate each in-play shot using a score between 0 and 1 according to its exciting degree they feel. The people themselves were free to present exact definition and scale of the highlights. It will not be confused as human-beings are good at comparison especially in a continuous match [50], and they are able to automatically adjust the exciting degree value to a reasonable state according to the whole video.

TABLE IV
SUBJECTIVE USER STUDIES FOR HIGHLIGHT EVALUATION

Data	Case 1		Case 2 (Selected as ground truth)	
	SD (%)	*SC* (%)	*SD* (%)	*SC* (%)
French_Open_2005	27.7	72.3	13.5	86.5
Australian_Open_2005	29.1	70.9	12.2	87.8
Dubai_Championship_2006	24.0	76.0	11.3	88.7
Olympic_Game_2004_A	21.5	78.5	10.8	89.2
Olympic_Game_2004_B	19.6	80.4	10.1	89.9
Mean	24.4	75.6	11.6	88.4

For shot S_i, the mean value of the scores rated by all people is defined as the ground truth g_i. Then, the subjective deviation (SD) for each match is calculated as

$$SD = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M}\sum_{j=1}^{M}(v_{i,j} - g_i)^2\right)^{1/2} \tag{28}$$

where $v_{i,j}$ represents the score given by subject j for shot S_i, N and M are the number of shots and people respectively. Table IV lists the result of subjective evaluation. Based on the deviation, we can define the subjective consistency (SC) as

$$SC = 1.0 - SD \tag{29}$$

which can be regarded as the ranking accuracy of human. From Table IV, we can see that the means of subjects' deviation are 24.4% and 11.6% with 75.6% and 88.4% subjective consistency, respectively.

The significant difference between two user studies is mainly because the background of the subjects participating in the study. In the first case, the people in the test have different definitions or concepts for the highlight excitement degree. Therefore, the distribution of impressive scores is diverse and results in the deviation being more than 20%. However in the second case, the people have similar definition of highlight concept and impressive degree, the deviation is thus smaller which is about 12%. In our ranking experiment, we exploited the result of second case as the ground truth to compare with the automatic estimation.

2) Verification of Effectiveness for SVR Highlight Ranking: We employed the data listed in Table V to train the ranking model and verify the effectiveness of our SVR based approach. We used the subjective criterion—highlight ranking accuracy (HRA) defined in (19) as the evaluation metric. In order to determine the quantized highlight rank R in (19), the rank candidate set was initialized $K = \{2, \ldots, 20\}$. Then, all the shots were split into two subsets by equal partition of the game duration. The shots in the subset which belong to the first half were used to optimize the quantization process. The final computed number of rank R is 10. The same data set was selected to train two SVR highlight ranking models for tennis and badminton respectively, and then evaluation was conducted on all the data.

We first conducted the automatic ranking with manually annotated action data regardless of the existing error recognition. Then, we conducted the experiment using the recognized result which involves false recognized action clips. To compare such

TABLE V
TRAINING AND TESTING DATA FOR EFFECTIVENESS VERIFICATION OF HIGHLIGHT RANKING APPROACH

Data	# In-play shot	# Training shot	# Test shot
French_Open_2005	243	116	243
Australian_Open_2005	207	128	207
Olympic_Game_2004_A	113	47	113
Total	563	291	563

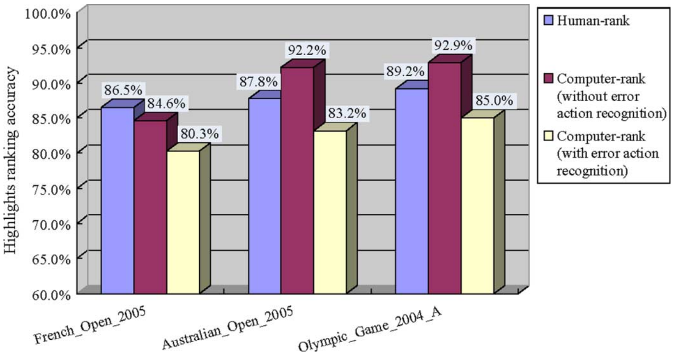

Fig. 15. Results of effectiveness verification for automatic highlight ranking and comparison between human and computer.

two results, it can demonstrate that the ranking model based on support vector regression is robust to the errors of input features. Fig. 15 shows the two computer ranking results and the comparison between manual and automatic results. The mean value of HRA for computer with manually annotated data is 89.9%. Note that about 90% accuracy is a remarkable result automatically obtained by computer since there is still 12.2% mean deviation (SD) of three videos for subjective ranking from human perception. This result meanwhile demonstrates that the extracted affective features can reflect human perception to a large extent. On the other hand, the mean HRA value for computer with falsely recognized result is 82.8%. This is satisfactory because the recognized data involve about 10% false result which is the noise of the ranking model input. Since the noise-insensitive loss function is exploited by support vector regression [46], our SVR based ranking model is robust to small errors induced by noise.

3) Verification of Generalization for SVR Highlight Ranking: In this experiment, we aim to verify the generalization of the proposed ranking approach. We conducted the verification on two complete videos of Dubai Championship 2006 for tennis and Olympic Game 2004 B for badminton. The trained ranking models in the abovementioned experiment were employed. All the in-play shots in these two videos are not involved in the training set for model construction. Therefore, the performance evaluated on these two videos can be exploited as the demonstration for the expansibility of our highlight ranking approach.

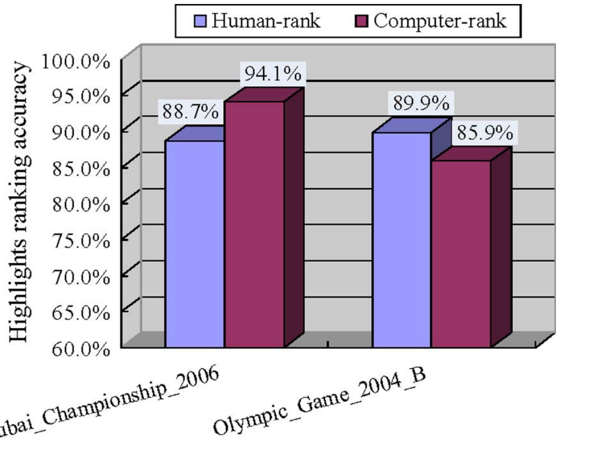

Fig. 16. Results of generalization verification for automatic highlight ranking and comparison between human and computer.

Fig. 16 shows the experimental result. It can be seen from the figure that the trained models achieve a satisfied ranking accuracy with the average error between subjective evaluation and computer estimation being about 4.7%. This verifies that our SVR highlight ranking approach has powerful expansibility and is general for new video data. The reason leading to such result can be summarized from two aspects of affective feature extraction and ranking model construction. First, the affective features employed in our approach are extracted from the mid-level representation of the video content in terms of player actions, trajectories and audio keywords. The mid-level representation is persistent across different sports videos as long as the games are of the same genre. Thus, our affective features are general for different videos of racket sports. Secondly, the ranking model simulates the mechanism of human perception for the evaluation of highlight impression. It is mostly relative with the subjects included in the user study and the affective features used. The model construction is independent of the video's physical attributes such as the TV channel for game broadcasting and the opponent players in the match. Consequently, the ranking model is general for different racket sports videos.

With the ranking results, scalable summarization and hierarchical browsing are easy to be achieved according to (22), (23), and (24). For the novice user, our approach first presents the highlight segments with the initial retrieval parameter, e.g., $T = 1\ \mathrm{min}, C = 0.5$, or $L = 5$. Then, the user is able to tune the parameter to obtain the improved feedback according to his/her feeling for current result. Because the ranking model is constructed by learning the evaluation from sports professionals, the retrieval result is suitable for most of the common users.

4) Essential Affective Feature Analysis: We exploited forward search algorithm [52] to perform the essential analysis. The objective of this analysis is to evaluate the role of each affective feature in ranking approach and obtain the essential affective feature set for highlight ranking. Moreover, this work can sort all the features according to their effectiveness in ranking performance. This will benefit for the further affective feature extraction, which guides the ranking system to select the features most reflecting the human perception. The affective feature set was first divided into selected set F_S and unselected set F_U, and then selected the feature from F_U one by one using the procedure shown in Fig. 17.

Fig. 18 presents the analysis results. We sort all the affective features in descending order of ranking accuracy and find the essential feature sets for highlight ranking of tennis and badminton respectively. As shown in Fig. 18(a), the set of (DSR, SSR, DOA, MCC, SOP) is more effective for highlight ranking of tennis game. From Fig. 18(b), we can see all the six features

```
(1) Set $F_S = \phi$ and $F_U = F$ ;
(2) Label all the features in $F_U$ untested;
(3) Select on untested feature $f$ from $F_U$ and label it as tested;
(4) Put $f$ and $F_S$ together to form the temporary testing feature set $\tilde{F}_S$ ;
(5) Evaluate the highlight ranking accuracy (HRA). In this procedure, 3-fold cross validation on all the in-play shots is conducted. The average HRA for all
    iterations is set as the estimated accuracy for the testing feature set;
(6) If there are still untested features in $F_U$ , goto (3);
(7) Find the feature $\hat{f}$ such that when we add it into the feature set $\tilde{F}_S$ , the highest HRA will be obtained $\hat{f} = \arg\max[HRA(\tilde{F}_S)]$ and then move $\hat{f}$ from
    $F_U$ to $F_S$ ;
(8) If there are still untested features in $F_U$ , goto (2). And if $F_U$ is empty, the procedure exists.
```

Fig. 17. Essential affective features analysis using forward search algorithm.

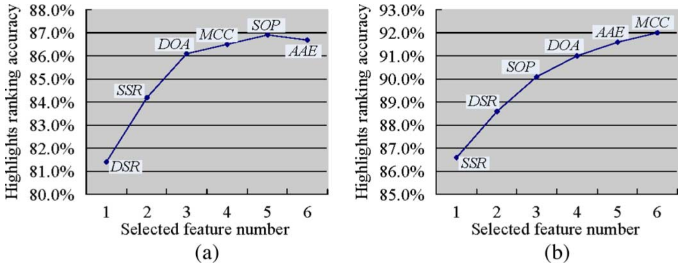

Fig. 18. Essential affective feature analysis: (a) tennis and (b) badminton.

(SSR, DSR, SOP, DOA, AAE, MCC) are important for the ranking of badminton game. With the analysis result, we investigate the six people involved in the user study: all the people stated that trajectory and action are the most two preferred criteria in ranking behavior and the audience response represented by applause is utilized as the complementary reference. Such investigation is consistent with our analysis result. This demonstrates that the extracted affective features in our approach can properly reflect the human perception.

VI. Conclusion

In this paper, we present a novel multimodal approach to rank the highlights extracted from broadcast racket sports video in the affective context. Two challenges, affective feature extraction and ranking model construction, are addressed. We extract the affective features from player behavior and audience response. The nonlinear highlight ranking model is constructed based on support vector regression.

We propose a new motion analysis method to recognize the player actions in broadcast racket sports video. The proposed method achieves satisfied result for the recognition of three basic player actions. With the comparison, our method significantly outperforms the existing appearance-based algorithm. The highlight ranking approach combines the player action recognition with real-world trajectory computation and audio keywords generation to establish the mid-level representation of video content. The affective features are extracted from player actions, trajectories and game-specific audio keywords. With the consideration of generality, support vector regression is employed to construct the nonlinear ranking model. Experimental results demonstrate that the affective features properly reflect the human perception and the ranking model is effective for automatic highlight evaluation.

To the best of our knowledge, our affective highlight ranking approach is the first proposed solution for racket sports based on broadcast video. Several issues will be further studied in our future work. As the primary work, action recognition, player tracking and audio keywords generation will be generalized to more sports domains straightforwardly or with suitable changes. Since our proposed nonlinear ranking approach relies on the statistical model, it can be easily extended to other sports video for affective analysis with more representative features coming from various channels, such as automatic speech recognition, closed caption, and text web broadcasting. In addition, the human-computer interaction is not involved in the current ranking approach. Considering the feedback information from user is in the list of future work to improve the personalized capacity of ranking approach.

References

[1] Y. Gong, T. S. Lim, H. C. Chua, H. J. Zhang, and M. Sakauchi, "Automatic parsing of TV soccer programs," in *Proc. Int. Conf. Multimedia Computing and System*, Washington, DC, 1995, pp. 167–174.

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

[3] L. Xie, P. Xu, S. F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recognit. Lett.*, vol. 25, no. 7, pp. 767–775, 2004.

[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] J. Assfalg, M. Bertini, C. Colombo, A. Delbimbo, and W. Nunziati, "Semantic annotation of soccer video: Automatic highlights identification," *Comput. Vis. Image Understand.*, vol. 92, no. 2–3, pp. 285–305, 2003.

[6] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted American football video by highlight selection," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 575–586, Aug. 2004.

[7] K. Wan, J. Wang, C. Xu, and Q. Tian, "Automatic sports highlights extraction with content augmentation," in *Proc. Pacific-Rim Conf. Multimedia*, Tokyo, 2004, vol. 3332, pp. 19–26.

[8] D. Liang, Y. Liu, Q. Huang, and G. Zhu, "Video2Cartoon: Generating 3-D cartoon from broadcast soccer video," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 217–218.

[9] G. S. Pingali, Y. Jean, A. Opalach, and I. Carlbom, "LucentVision: Converting real world events into multimedia experiences," in *Proc. Int. Conf. Multimedia & Expo*, New York, 2000, vol. 3, pp. 1433–1436.

[10] C. C. Cheng and C. T. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 585–599, 2006.

[11] H. L. Wang and L. F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, 2006.

[12] [Online]. Available: http://en.wikipedia.org/wiki/List-of-sports#Racket-sports

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

[14] G. S. Pingali, Y. Jean, and I. Carlbom, "Real time tracking for enhanced tennis broadcasts," in *Proc. Conf. Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998, pp. 260–265.

[15] X. Yu, C. H. Sim, J. R. Wang, and L. F. Cheong, "A trajectory-based ball detection and tracking algorithm in broadcast tennis video," in *Proc. Int. Conf. Image Processing*, Singapore, 2004, vol. 2, pp. 1049–1052.

[16] H. Miyamori and S. Iisaku, "Video annotation for content-based retrieval using human behavior analysis and domain knowledge," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp. 320–325.

[17] H. Miyamori, "Improving accuracy in behavior identification for content-based retrieval by using audio and video information," in *Proc. Int. Conf. Pattern Recognition*, 2002, vol. 2, pp. 826–830.

[18] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues," in *Proc. Int. Conf. Multimedia Expo*, Baltimore, MD, 2003, vol. 3, pp. 309–312.

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

[20] X. Tong, Q. Liu, Y. Zhang, and H. Lu, "Highlight ranking for sports video browsing," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 519–522.

[21] L. Xing, H. Yu, Q. Huang, Q. Ye, and A. Divakaran, "Subjective evaluation criterion for selecting affective features and modeling highlights," *Proc. SPIE Multimedia Content Analysis, Management, and Retrieval*, vol. 6073, 2006.

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

[23] Q. Ye, Q. Huang, W. Gao, and S. Jiang, "Exciting event detection in broadcast soccer video with mid-level description and incremental learning," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 455–458.

[24] D. Tjondronegoro, Y. P. P. Chen, and B. Pham, "Content-based video indexing for sports applications using integrated multi-modal approach," in *Proc. ACM Multimedia*, Singapore, 2005, pp. 1035–1036.

[25] Q. Ye, Q. Huang, S. Jiang, Y. Liu, and W. Gao, "Jersey number detection in sports video for athlete identification," in *Proc. SPIE Visual Communications & Image Processing*, Beijing, China, 2005, pp. 1599–1606.

[26] G. Zhu, C. Xu, Q. Huang, W. Gao, and L. Xing, "Player action recognition in broadcast tennis video with applications to semantic analysis of sports game," in *Proc. ACM Multimedia*, Santa Barbara, CA, 2006, pp. 431–440.

[27] M. Shah and R. Jain, *Motion-Based Recognition*. Norwell, MA: Kluwer, 1997.

[28] D. M. Gavrila, "The visual analysis of human movement: A survey," *Comput. Vis. Image Understand.*, vol. 73, no. 1, pp. 82–98, 1999.

[29] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[30] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 7, pp. 814–827, 2003.

[31] R. Cutler and L. S. Davis, "Robust real-time periodic motion detection, analysis, and application," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 8, pp. 781–796, 2000.

[32] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proc. Int. Conf. Computer Vision*, Nice, France, 2003, vol. 2, pp. 726–733.

[33] M. C. Roh, B. Christmas, J. Kittler, and S. W. Lee, "Robust player gesture spotting and recognition in low-resolution sports video," in *Proc. Eur. Conf. Computer Vision*, 2006, pp. 347–358.

[34] W. L. Lu and J. J. Little, "Tracking and recognizing actions at a distance," in *Proc. Eur. Conf. Computer Vision Workshop on Computer Vision Based Analysis in Sport Environments*, 2006, pp. 49–60.

[35] R. W. Picard, *Affective Computing*. Cambridge: MIT Press, 2000.

[36] A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[37] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1114–1122, 2005.

[38] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework," in *Proc. Int. Conf. Image Processing*, Barcelona, Spain, 2003, vol. 1, pp. 5–8.

[39] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput., Commun., Applicat.*, vol. 2, no. 1, pp. 1–19, 2006.

[40] G. Zhu, D. Liang, Y. Liu, Q. Huang, and W. Gao, "Improving particle filter with support vector regression for efficient visual tracking," in *Proc. Int. Conf. Image Processing*, Genova, Italy, 2005, vol. 2, pp. 422–425.

[41] S. Jiang, Q. Ye, W. Gao, and T. Huang, "A new method to segment playfield and its applications in match analysis in sports video," in *Proc. ACM Multimedia*, New York, 2004, pp. 292–295.

[42] Q. Ye, W. Gao, and W. Zeng, "Color image segmentation using density-based clustering," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, 2003, vol. 3, pp. 345–348.

[43] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.

[44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Conf. Computer Vision and Pattern Recognition*, San Diego, CA, 2005, pp. 886–893.

[45] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 5, pp. 564–577, 2003.

[46] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[47] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithm," *Neural Comput.*, vol. 12, pp. 1083–1121, 2000.

[48] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[49] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 497–501, 2000.

[50] K. A. Peker and A. Divakaran, "Framework for measurement of the intensity of motion activity of video segments," *Proc. SPIE*, vol. 4862, pp. 126–137, 2002.

[51] J. Chin, V. Diehl, and K. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proc. SIGCHI on Human Factors in CS*, Washington, DC, 1988, pp. 213–218.

[52] A. K. Jain, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4–37, 2000.

[53] C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan, "Live sports event detection based on broadcast video and web-casting text," in *Proc. ACM Multimedia*, Santa Barbara, CA, 2006, pp. 221–230.

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, where he is currently pursuing the Ph.D. degree.

His research interests include image/video processing, multimedia content analysis, computer vision and pattern recognition, and machine learning.

Qingming Huang (M'04) 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 School of Chinese Academy of Sciences. He has published over 80 scientific papers. His current research areas are image processing, video analysis, video coding, and pattern recognition.

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

From 1996 to 1998, he was with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. He joined the Institute for Infocomm Research (I2R), Singapore, in March 1998. His research interests include multimedia content analysis, indexing and retrieval, digital watermarking, computer vision, and pattern recognition. He has published over 150 papers in those areas. He is an Associate Editor of *Multimedia Systems Journal*.

Dr. Xu served as Program Co-Chair of 2006 Asia-Pacific Workshop on Visual Information Processing (VIP2006) and 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 and Expo, Pacific-Rim Conference on Multimedia, and International Conference on Multimedia Modeling.

Liyuan Xing received the B.S. degree in computer science from Southern Yangtze University, Wuxi, China, in 2003 and the M.S. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2006.

She has worked on medical image processing at the R&D Center, Toshiba (China) Co., Ltd., since July, 2006. Her research interests include image/video processing, pattern recognition, machine learning, and multimedia content analysis.

Wen Gao (M'92-SM'05) received the M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1985 and in 1988, respectively, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

He was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo, in 1992, and a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, in 1993. From 1994 to 1995, he was a Visiting Professor at the AI Lab, Massachusetts Institute of Technology, Cambridge. Currently, he is a Professor with the School of Electronic Engineering and Computer Science, Peking University, Peking, China, and a Professor in computer science at Harbin Institute of Technology. He is also the Honor Professor in computer science at the City University of Hong Kong and the External Fellow of International Computer Science Institute, University of California, Berkeley. He has published seven books and over 200 scientific papers. His research interests are in the areas of signal processing, image and video communication, computer vision, and artificial intelligence.

Dr. Gao is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, Editor-in-Chief of the *Journal of Computer* (in Chinese), and Editor of the *Journal of Visual Communication and Image Representation*.

Hongxun Yao (M'00) received the B.S. and M.S. degrees in computer science from the Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and in 1990, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology in 2003.

Currently, she is a Professor with School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include pattern recognition, multimedia processing, and digital watermarking. She has published three books and over 80 scientific papers.