Player Action Recognition in Broadcast Tennis Video
with Applications to Semantic Analysis of Sports Game

Guangyu Zhu 1, Changsheng Xu 2, Qingming Huang 3, Wen Gao 1, and Liyuan Xing 3
1 School of Computer Science
Harbin Institute of Technology
Harbin, P.R. China, 150001
2 Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
3 Graduate School of Chinese
Academy of Sciences
Beijing, P.R. China, 100080
{gyzhu, qmhuang, wgao, lyxing}@jdl.ac.cn, xucs@i2r.a-star.edu.sg

ABSTRACT
Recognition of player actions in broadcast sports video is a challenging task due to low resolution of the players in video frames. In this paper, we present a novel method to recognize the basic player actions in broadcast tennis video. Different from the existing appearance-based approaches, our method is based on motion analysis and considers the relationship between the movements of different body parts and the regions in the image plane. A novel motion descriptor is proposed and supervised learning is employed to train the action classifier. We also propose a novel framework by combining the player action recognition with other multimodal features for semantic and tactic analysis of the broadcast tennis video. Incorporating action recognition into the framework not only improves the semantic indexing and retrieval performance of the video content, but also conducts highlights ranking and tactics analysis in tennis matches, which is the first solution to our knowledge for tennis game. The experimental results demonstrate that our player action recognition method outperforms existing appearance-based approaches and the multimodal framework is effective for broadcast tennis video analysis.

Categories and Subject Descriptors
I.4.8 [Scene Analysis]: Motion, Object Recognition, Tracking.
1.2.10 [Vision and Scene Understanding]: Video Analysis.

General Terms
Algorithms, Performance, Experimentation.

Keywords
Object Tracking, Motion Representation, Action Recognition, Semantic Analysis.

1. INTRODUCTION
Over the past decades, there is an explosive growth in the amount of available multimedia information in our daily lives. This trend necessitates the development of content-based video analysis, indexing and retrieval technology. In recent years, extensive research efforts have been devoted to sports video content analysis and applications due to their wide viewership and high commercial potentials. Technologies and prototypes have been developed for automatic or semi-automatic sports video content analysis [1]-[3], semantic event detection and personalization [4][5], virtual advertisements insertion [6], and multi-camera based 3-D virtual sports events generation [7]. For sports video analysis, there is a semantic gap between the richness of user semantics and the simplicity of available low-level perceptual visual and auditory features. To bridge this gap, we need to come up with a mid-level representation for sports video content. The objects in sports video can be considered as an effective mid-level representation to facilitate semantic analysis such as structure analysis, event detection, and tactics analysis. For example, the actions performed by players in tennis game reveal the process of the game and the tactics of the players. The movement of players in tennis video provides useful information for analysis and understanding of the game. Therefore, recognition of player actions is essential for the analysis of tennis matches and is desired by sports professionals and longtime famers for technical and tactic coaching assistance.

Figure 1. Two typical frames derived from broadcast tennis video. (a) Close-up, (b) Far-view, the zoomed picture is the player whose action is recognized.

Broadcast video is a post-edited video where the broadcast sequence feed is selected from frequent switches among multiple cameras according to broadcast director’s instruction. Considering the ratio of playfield pixels to the frame size as shown in Figure 1, the frames in broadcast tennis video can be divided into two classes: close-up where the magnitude of player figure is higher, and far-view where the magnitude is lower. In close-up frames, a player figure is usually 300 pixels tall. It is easy to segment and label human body parts such as the limbs, torso, and head. Existing work [8][9] has achieved good results on action recognition for close-up view. On the other hand, in far-view frames, a player figure might be only 30 pixels tall. 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 too many categories for far-view frames. To the best of our
knowledge, there are few efforts devoted in the research of player action recognition in far-view frames of the broadcast tennis video. Miyamori et al. [14] developed an appearance-based annotation system for tennis actions including foreside-swing, backside-swing and overshoulder-swing based on silhouette transitions. In [15], they improved original method by combining player and ball positions. However, the appearance is not necessarily preserved across different videos. Furthermore, ball detection and tracking in broadcast video is significantly difficult due to the poor quality of the footage.

In this paper, we propose a novel motion analysis approach to recognize player actions in far-view frames of the broadcast tennis video. Two basic actions, left-swing and right-swing which occupy more than 90% play behavior in tennis games, are recognized. The contributions of this paper include: 1) We propose a novel player action recognition method based on motion analysis, which is different from existing appearance-based approaches; 2) In our approach, we treat the optical flow as spatial patterns of noisy measurements instead of precise pixel displacements; 3) We propose a new motion descriptor, slice-based optical flow histograms abbreviated as S-OFHs, as the motion representation and carry out recognition using a supervised learning framework; 4) We propose a novel framework for automatic broadcast tennis video analysis by combining player action recognition with other multimodal features in sports video.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 presents the player action recognition approach in broadcast tennis video. Section 4 describes the multimodal framework for broadcast tennis video semantic and tactics analysis. Experiments are reported and analyzed in section 5. Finally, we conclude the paper with future work in section 6.

2. RELATED WORK

2.1 Related Work on Action Recognition

Most of existing work in human motion tracking and action recognition is focusing on close-up view with high resolution figures. In [8][9], the previous work on tracking and recognition at the level of human body parts (close-up) was reviewed, but those approaches were inapplicable for the far-view frames in broadcast tennis video. 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 [12], spatial arrangement of moving points [13], 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. Obviously, these approaches are not suitable for the type of video data considered in our work.

Compared with the action recognition for the videos with high resolution figures, a little work [10][11] is attempting to analyze poor quality and non-stationary camera footage. The approach proposed in [10] 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. [11] developed a generic approach to recognize action in “medium field” which is similar to “far-view” defined in our work. In their approach, a motion descriptor in a spatio-temporal volume was introduced and an associated similarity measure was utilized in a nearest neighbor classification (NNC) framework to perform action categorization.

However, the tennis videos they used are non-broadcast which has less challenge for tracking and recognition. Miyamori et al. [14] developed a prototype system of player action recognition in tennis video. The recognition was based on silhouette transition. The original work was extended [15] 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., appearance feature is not consistently preserved across different videos so that it is less robust for classification. Moreover, ball detection and tracking in broadcast video is also very difficult due to high speed of the ball and low resolution of the frames.

2.2 Related Work on Tennis Video Analysis

Since tennis is one of the most popular sports, tennis video analysis has attracted great attention. This interest is further motivated by the possible applications over a wide range of topics such as tactics analysis, indexing, retrieval, and summarization [14]-[23]. Most techniques available in the literature for tennis tactics analysis have focused on tracking the motion trajectories of the players and ball [16]-[18]. G. Sudhir et al. [16] exploited the domain knowledge of tennis video to develop a court line detection algorithm and a player tracking algorithm for the purpose of automatic generation of high-level annotations of play events. In [17], 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. [18] proposed a trajectory-based algorithm to detect and track the ball in the broadcast tennis video. The experimental results showed that the algorithm had obtained promising accuracy for summarization. In recent years, multimodal analysis has been widely employed to tennis video for structure parsing [19], event detection [20], video indexing [21], pattern discovery [22], and highlights ranking [23].

2.3 Our Approach

Existing approaches for action recognition in broadcast tennis video are based on appearance analysis [14] or reasoning scenario with the integration of player and ball positions [15]. 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. The algorithm contains three modules as shown in Figure 2. 1) The algorithm starts by player tracking and human-centric figure computation. 2) Optical flow is derived as low-level feature. A novel motion descriptor is then calculated based on slice partition with the relationship between locomotory body parts and optical flow field regions. 3) The action recognition is carried out in a supervised learning framework where support vector machine is employed in our approach. Voting strategy is utilized for action clip recognition by aggregating the frame classification over the temporal domain. Left-swing and right-swing performed by the player in far-view scenes in the broadcast video are recognized. By combining recognized player actions with other mid-level features, we propose a new multimodal framework as depicted in Figure 7 for broadcast tennis video semantic and tactics analysis including video annotation and retrieval based on action indexing, highlights ranking and hierarchical content browsing, tactics analysis and statistics.

3. PLAYER ACTION RECOGNITION

3.1 Player Tracking and Stabilization

Our recognition algorithm starts by player tracking and stabilization for human-centric figure computation. Such a process can be achieved by tracking the player candidate region through the frame sequence and then constructing a window in each frame centered at the player. The detection algorithm in [16] is used to extract the player’s initial position as the input of tracker.

The appropriate tracker utilized here is expected to be consistent, that is, the tracker should be robust enough for the noise in the video so as to always map the player in a particular body configuration to approximate the same stabilized image. Existing methods for tracking tennis players are based on template matching [14]-[17], which is similar to the correlation based algorithm [11]. These trackers are sensitive to the noise such as player deformation and background clutter caused by non-rigid object and low frame resolution, and cannot track player for a long video sequence. This can be exemplified in [16] that the input video was first segmented into chunks of 30 frames and then conducted tracking for each chunk separately. A sophisticated tracking strategy called SVR particle filter [24] is employed in our approach. SVR particle filter enhances the performance of classical particle filter with small sample set and is robust enough for the noise in broadcast video. The experimental result is very promising. More details about this tracker can be found in [24].

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 is used to calculate the centroid of the player region. The centroid coordinates are defined as follows:

\[
\begin{align*}
mx &= \frac{\sum_{x=R} \sum_{y=R} xf(x, y)}{\sum_{x=R} \sum_{y=R} f(x, y)} \\
mn &= \frac{\sum_{x=R} \sum_{y=R} yf(x, y)}{\sum_{x=R} \sum_{y=R} f(x, y)}
\end{align*}
\]

where \( R \) is the region occupied by the object on the image plane and \( f(x, y) \) the gray level at location \((x, y)\). Then the center of the window controlled by tracker is shifted to 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 played with player.

3.2 Motion Descriptor Computation

In previous approaches, appearance-based feature is extracted for the representation of player action. However, in tennis games, different players may exhibit different postures for the same action and different postures may be recorded in different videos even for the same action. Thus, the appearance descriptor is not robust and discriminative for action recognition and classification.

In our approach, we extract features from pixel-wise optical flow which is the most natural technique for capturing the motion independent of appearance. The challenge is that the computation of optical flow is not very accurate, particularly on coarse and noisy data such as broadcast video footage. Our insight 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 sub-regions called slices. The histogram is utilized to represent the spatial distribution for each sub optical flow field in slices.

3.2.1 Optical Flow Computation and Noise Elimination

The extraction of motion descriptor is based on the optical flow in the human-centric figure. We compute optical flow at each figure using Horn-Schunck algorithm [25]. The motion induced by background is not needed for our research because it does not reflect the real movement information about player action. Moreover, the noise in figure background makes significant influence for the accurate computation of optical flow for the player region. It necessitates background subtraction before computing optical flow. Considering the background of human-centric figure is playfield, an adaptive method of playfield detection [26] is applied to subtract background. After background pixels detection, we employ the region growing algorithm in [27] to perform post-processing to connect background pixels into regions, eliminate noises, and smooth the boundaries. The flowchart of the background subtraction is shown in Figure 3.

Half-wave rectification and Gaussian smoothing is applied to eliminate the noise in optical flow field. The method is shown in Figure 4. The optical flow magnitudes are first thresholded to reduce the effect of too small and too large motion probably due to 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 \( OFF_y \), each of which is then half-wave rectified into four non-negative channels \( OFF_x^+, \, OFF_x^-, \, OFF_y^+, \, \text{and} \, OFF_y^- \), 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.
motivated by the kernel density estimation for color distribution [28], we derive a group of slice based optical flow histograms (S-OFHs). First, we define \( \{1, \ldots, N\} \) for the action clip. The window length is empirically set to 25 frames.

Based on frame recognition and voting strategy, the action clips in tennis video are classified into two categories: left-swing and right-swing. We use audio features [20] to detect the sound of hitting ball to locate the action clip in the video. As shown in Figure 6, 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 25 frames.

Figure 3. Adaptive background subtraction for human-centric figure.

Figure 4. Half-wave rectification and Gaussian smoothing for noise elimination of optical flow field.

3.2.2 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 are exhibited in the different figure regions. This can be demonstrated by observing the optical flow field computed from the processed human-centric figure (see Figure. 5(a)). For left-swing, the optical flow field in the left figure region 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. Based on this observation, we adopt a simple but effective region style called slice in our approach. The whole optical flow field after noise elimination is split into three slices along the width orientation as shown in Figure 5(b). The height of the slice is equal to the height of the figure and the width can be set adaptively in accordance with the object spatial structure. Here, we set even width for each slice.

Histogram based methods are widely used for spatial recognition. Motivated by the kernel density estimation for color distribution [28], we derive a group of slice based optical flow histograms (S-OFHs). First, we define \( b(p) \in \{1, \ldots, m\} \) as the bin index of histogram associated with the optical flow vector \( f \) at location \( p \). For each position \( q \) inside the optical flow field \( OFF \), considering a grid region \( R(q) \) centered at \( q \), the probability of bin \( u = 1, \ldots, m \) in the histogram of \( OFF \) is then calculated as

\[
h_u = C \sum_{q \in OFF} \sum_{p \in R(q)} k(||q - p||) \delta[b(p) - u]
\]

where \( \delta \) is the Kronecker delta function, \( C \) is the normalization constant ensuring \( \sum_u h_u = 1 \), \( k \) is a convex and monotonic decreasing kernel profile which assigns a smaller weight to the locations that are farther from the center \( q \).

Given a refined optical flow field \( OFF \) for the figure \( F_i \) in the human-centric figure sequence, \( i = 1, \ldots, N \) where \( N \) is the total figure number, \( OFF_{ij} \) is the sub optical flow field in the slice, \( j = 1, \ldots, L \) and here \( L = 3 \). The \( S-OFH_{ij} \) is defined as following according to Eq. (2).

Thus, for each \( OFF_{ij} \), two S-OFHs annotated as \( S-OFH_{ij}^h \) and \( S-OFH_{ij}^v \) are constructed for horizontal and vertical orientation of the optical flow field, respectively.

In our approach, left and right slices are selected for computing the S-OFHs. Four S-OFHs for one figure are ultimately utilized as the motion descriptor. Figure 5(c) shows the S-OFHs for corresponding action. We can see that S-OFHs can effectively capture the discriminative features for different actions in spatial space.

3.3 Action Classification

Various supervised learning algorithms can be employed to train an action pattern recognizer. Support vector machine (SVM) [29] is used in our recognition approach. SVM has been successfully applied to a wide range of pattern recognition and classification problems. The advantages of SVM consist of: 1) providing better prediction on unseen test data, 2) providing a unique optimal solution for a training problem, 3) containing fewer parameters compared with other methods.

In our approach, we formulate action recognition as a classification task. The concatenation of four S-OFHs for each optical flow field in the figure of one video frame is fed as feature vector into support vector machine. The radial basis function (RBF) kernel \( K(x, y) = \exp(-\lambda \|x - y\|) \) is utilized to map training vectors into a high dimensional feature space for classification.

Based on frame recognition and voting strategy, the action clips in tennis video are classified into two categories: left-swing and right-swing. We use audio features [20] to detect the sound of hitting ball to locate the action clip in the video. As shown in Figure 6, 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 25 frames.

Figure 5. Slice partition and slice based optical flow histograms (S-OFHs) of left and right swing.

\[
h_i^l = C \sum_{q \in OFF} \sum_{p \in R(q)} k(||q - p||) \delta[b(p) - u]
\]

Thus, for each \( OFF_{ij} \), two S-OFHs annotated as \( S-OFH_{ij}^h \) and \( S-OFH_{ij}^v \) are constructed for horizontal and vertical orientation of the optical flow field, respectively.

\[
Vote(f_i) = \begin{cases} 
1 & \text{if Reg}(hc_i) = \text{left} - \text{swing} \\
-1 & \text{if Reg}(hc_i) = \text{right} - \text{swing}
\end{cases}
\]
Far-view alignment 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 8 frames.

Because there are two categories, the equal sign is just assigned for left-swing so as to avoid the occurrence of marginal classification. 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 8 frames.

4. TENNIS VIDEO ANALYSIS BASED ON MULTIMODAL FRAMEWORK

In this section, we propose a multimodal framework for semantic indexing, highlights ranking and tactics analysis of broadcast tennis video by integrating action recognition with audio analysis and real-world trajectory computation. The proposed framework extracts features from player action, audio keywords, and real-world trajectory computation to form a mid-level representation of the video, which is superior to low-level features for semantic analysis and is able to bridge the gap between low-level features and high-level semantics.

4.1 Framework Overview

As illustrated in Figure 7, the multimodal framework consists of four modules.

- **Low-level analysis**: The dominant color-based algorithm in [16] was used to identify all the in-play shots which are also the rally scenes screened from far-view point. In this way, we discarded all those sequences not representing game actions such as close-up of player, audience shots or commercials. Texture features of playfield were also extracted to improve the accuracy of original color-based algorithm.

- **Mid-level analysis**: A set of mid-level features were extracted from each in-play shot. 1) Action recognition provided movements annotation as the indispensable information for player behavior analysis during the game. To assist the improvement of action based analysis, audio analysis and real-world trajectory computation were incorporated. 2) We used audio keywords [20] as a mid-level representation, which refer to a set of game-specific sounds with strong relationship to the behavior of players. We created three domain-specific audio keywords for tennis games: silence, hitting ball, and applause, which have strong hints to event. These audio keywords were used to detect the occurrence time of the player’s swing. 3) Player’s position was extracted in each video frame using object tracking approach in [24]. With the implementation of homography based planar projection from image plane to planform [30], we figured out the player’s real-world trajectory which is viewed in actual tennis court to correct the projective distortion as shown in Figure 10 due to the angular location of cameras when game was recorded.

- **Fusion of mid-level features**: Compared with low-level features, the mid-level features are able to facilitate the high-level analysis from the semantic concept point of view. We analyzed and recognized collected mid-level features within each in-play shot and mapped mid-level features to high-level semantic video attributes with the assistance of domain knowledge.

- **High-level analysis**: High-level analysis includes semantic content indexing, highlights ranking and tactics analysis. Video content was indexed and annotated based on recognized player actions. Action-trajectory based subjective ranking for highlights was evaluated. Affective features extracted from action and trajectory were exploited as the input of support vector regression ranking model. A tactics summarization scheme was proposed based on action recognition.

4.2 Action Based Tennis Video Indexing

Annotating digital tennis footage with metadata such as description of the stroke types (e.g. forehand-stroke and backhand-stroke) will provide us an easy access to footage segments in a video archive. The tennis coaches will also benefit from the annotation to retrieve training video clips of certain stroke type for the player to analyze his/her technical weaknesses. Another potential application is the automatic collection of game information, which facilitates match commentators or home viewers with direct access to current match statistics.

Primitive features such as color, shape, texture and motion are not able to achieve this purpose because they just represent low-level information of video content. We propose a new indexing method for tennis video based on action recognition and domain knowledge. The objective is to annotate tennis video by actions of forehand-stroke and backhand-stroke displayed by players in the game. By predefining the domain knowledge that provides the information about which hand grips the racket for specific player, for instance, Agassi is right-handed whereas Nadal is left-handed, we can estab-
lish a mapping between swing types and stroke types. The steps of proposed indexing method are illustrated in Figure 8.

**Predefine:**
- set of swing type $A = \{\text{left-swing, right-swing}\}$
- set of stroke type $S = \{\text{forehand-stroke, backhand-stroke}\}$
- mapping function $f_{st} : A \rightarrow S$

**Input:** broadcast tennis video $T$

**Output:** indices of stroke $IS = (s_1, \ldots, s_n)$, $s_i \in S$

1. Using audio analysis, the sequence $(V_1, \ldots, V_n)$ for all the action clips in $T$ is obtained.
2. Using discriminative function defined by Eq. (5), we get swing indices $IA = (a_1, \ldots, a_n)$, $a_i \in A$.
3. Using mapping function $f_{st}$, we get the indices of stroke $IS = (s_1, \ldots, s_n)$, $s_i \in S$.

**Figure 8. Video indexing based on stroke action.**

### 4.3 Highlights Ranking and Browsing

Inside the category of ball-type sports games, a distinction can be made between time-constrained games (soccer, basketball, etc.) and score-constrained games (tennis, badminton, etc.). Time-constrained games have relatively loose structure and distinct highlights such as shot, goal in soccer. Contrastively, score-constrained games have few distinct exciting events. Such kind of sports video is composed of a restricted number of typical scenes producing a repetitive pattern. Therefore, in contrast with soccer video analysis, we cannot define all the scoring events (in-play shots) in tennis video as the highlights. With the combination of action recognition and real-world trajectory computation, we propose a highlights ranking scheme for tennis game as shown in Figure 9. The ranking scheme facilitates the hierarchical browsing and customized content delivery according to the retrieval confidence input by users.

**Figure 9. Highlights ranking for tennis game based on action-trajectory representation.**

In tennis game broadcasting, the cameras are usually located at the two ends of the court above the central line. Thus, the tennis court is projected to be trapezoidal in the video frames as shown in Figure 10. Such projection leads to the distortion of player’s movement in the court. To calculate the real-world trajectory, the player’s positions in all frames within a shot are first obtained by tracking module in the action recognition method. Figure 11(a) shows such results aggregated in one representative frame. We then use homography technique [30] to calculate real-world trajectory which is the locus of the player viewed from planform as shown in Figure 11(b). The final trajectory is smoothed by Gaussian filter.

**Figure 10. Projective distortion in broadcast tennis video.** (a) The tennis court in video sequence, (b) The actual tennis court.

**Figure 11. Real-world trajectory computation.** (a) Trajectory in frames, (b) Corresponding real-world trajectory computed based on homography.

Highlights ranking model is an automatic evaluation system for the attractive confidence contained in sports footage and expected to arise in users while watching that footage. The input of ranking model is the features extracted from sports video like motion and sound which can reflect the human perception of highlights degree for a holonomic event. Such features are called affective features. The output is the automatic estimation of impressive confidence. In our ranking approach, we extract four affective features from action-trajectory 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 between left-swing and right-swing occurred in an in-play shot. Considering the sequence of action clips $(V_1, \ldots, V_n)$ in shot and Eq. (5) where $n$ is the total number of the action clips, we define the indicator function $SL$ as

$$SL(V) = \begin{cases} +1 & \text{if Category}(V) = \text{left-swing} \\ -1 & \text{if Category}(V) = \text{right-swing} \end{cases}$$

(6)

Then, swing switching rate is calculated as follows:

$$SSR = \sum_{i=1}^{n-1} |SL(V_i) - SL(V_{i+1})| / 2(n-1)$$

(7)

where $|x|$ means the absolute value of $x$.

- **Features on trajectory:** Three features are derived from trajectory description.
  1) **Speed of Player (SOP)** which is calculated based on the length of real-world trajectory and shot duration.
  2) **Maximum Covered Court (MCC)** which is the area of rectangle shaped with the leftmost, rightmost, topmost, and bottommost points on real-world trajectory.
  3) **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

$$DSR = \left(|IH| + |IV|\right) / \|PS\|$$

(8)

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

436
The affective feature vector comprised of (SSR, SOP, MCC, DSR) is fed into the ranking model constructed by support vector regression (SVR) [23]. Based on our experimental results, SVR ranking model has been demonstrated that its performance is consistent with subjective evaluation [23]. Ultimately, the ranking system provides a hierarchical browsing fashion for users with the impressive confidence which is a customized parameter defined by user’s preference. The returned ranking list is represented as

$$U_i = \{u_i \mid H(u_i) \geq \text{coef}, i = 1, \ldots, c\}$$  \hspace{1cm} (9)

where $u_i$ is an in-play shot, $H(\cdot)$ is our ranking method, coef is the highlight threshold for impressive confidence defined by user.

### 4.4 Tactics Analysis and Statistics

Most current research efforts are focused on characterization and summarization of the semantic content of sports video. Tactics analysis, which is to explore the match strategies in sports video, can potentially offer assistance to coaches and players during training. Existing techniques available for tennis tactics analysis have focused on tracking the motion trajectories of the players and ball [16]-[18]. Although these methods can provide elaborate annotations to tennis matches, the ability is limited due to the event-driven indexing strategy which is inflexible and intractable for tactics analysis. We propose a novel action-driven tactics analysis approach for tennis game which is able to discover the insight of the stroke performance of players.

![Figure 12: Approach of tactics analysis for broadcast tennis game.](image)

Our tactics analysis approach shown in Figure 12 can evaluate the performance of player strokes and give the statistics of the scoring corresponding to forehand-stroke and backhand-stroke. We first use audio keywords to detect the last stroke action of a rally round. If the last stroke is played by the interested player, the stroke type is identified by our action recognition method. Then, the scoreboard region is extracted from the video. The identification and recognition of the information inside the scoreboard region in the video has been widely studied [31]. Therefore, based on the recognized result, we can know whether this stroke is scored or not. Making use of two-modality information, we can consequently give the statistics of the success rate and failure rate for forehand-stroke and backhand-stroke performed by the interested player. The success rate (SR) and failure rate (FR) are defined as

$$SR = \frac{\# \text{forehand-stroke (or backhand-stroke) with score}}{\# \text{forehand-stroke (or backhand-stroke)}}$$

$$FR = \frac{\# \text{forehand-stroke (or backhand-stroke) without score}}{\# \text{forehand-stroke (or backhand-stroke)}}$$

### 5. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed approaches, we carried out extensive experiments. The test data we used are obtained from the videos recorded from live broadcast television program for three matches of Pacific Life Open 2004, French Open 2005, and Australian Open 2005, respectively. The videos are compressed in MPEG-2 standard with the frame resolution of 352×288.

#### 5.1 Results of Action Recognition

To verify the effectiveness of our action recognition algorithm, 6 sequences are extracted from the whole videos with 194 action clips including 99 left-swing actions and 95 right-swing actions. Table 1 shows the detail of the test data.

**Table 1. Detail of video data for action recognition**

<table>
<thead>
<tr>
<th>Match</th>
<th># left-swing clips</th>
<th># right-swing clips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pacific_Open_1</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>Pacific_Open_2</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>Pacific_Open_3</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>French_Open_1</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Australian_Open_1</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>Australian_Open_2</td>
<td>7</td>
<td>18</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>99</td>
<td>95</td>
</tr>
</tbody>
</table>

To quantitatively evaluate the performance, we calculate Recall (R) and Precision (P) for each action category. The Accuracy (A) metric is employed to evaluate the holistic performance. Additional 71 action clips are utilized to train the two-action SVM model. Table 2 shows the experimental results with the Accuracy for all the action clips 90.21%. The reason resulting in incorrect recognition is that 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 S-OFHs misreport the motion characteristics in the human-centric figure.

**Table 2. Experimental results of action recognition**

<table>
<thead>
<tr>
<th>Stlew</th>
<th># clip</th>
<th>R (%)</th>
<th>P (%)</th>
<th>A (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-swing</td>
<td>99</td>
<td>89.89</td>
<td>90.82</td>
<td>90.21</td>
</tr>
<tr>
<td>Right-swing</td>
<td>95</td>
<td>90.53</td>
<td>89.58</td>
<td></td>
</tr>
</tbody>
</table>

A comparison with the existing appearance-based work is also carried out. The algorithm in [14] is evaluated using the same training and test data. Because there is no open source code for this method that can be found, we implemented the algorithm by ourselves strictly obey the original description in [14]. Silhouette transitions are extracted from the human-centric figures. KL transform is utilized to expand silhouette features to certain eigenspace. Different numbers of high-ranked features in eigenspace are selected as the discriminative parameters to identify action category in the nearest neighbor framework. We employed different percentages of eigen-feature basis and obtained the best Accuracy results when 73% is used. Table 3 gives the comparative results.

**Table 3. Comparative results for appearance-based method**

<table>
<thead>
<tr>
<th>Stlew</th>
<th># clip</th>
<th>R (%)</th>
<th>P (%)</th>
<th>A (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-swing</td>
<td>99</td>
<td>77.78</td>
<td>65.81</td>
<td>68.04</td>
</tr>
<tr>
<td>Right-swing</td>
<td>95</td>
<td>57.89</td>
<td>71.43</td>
<td></td>
</tr>
</tbody>
</table>
Comparing Table 2 with Table 3, it can be concluded that our method outperforms appearance-based algorithm because the motion descriptor is preserved much better than appearance representation and more robust for classification and recognition. Figure 13 illustrates some representative frames from the videos for the actions accurately recognized by our approach.

Figure 13. Representative recognition results for left-swing and right-swing.

Figure 14. Retrieval result for “backhand-stroke” using indexing technique based on action recognition.

5.2 Results of Tennis Video Analysis

We use three and an half hours video data for the experiments of broadcast tennis video semantic and tactics analysis. The various experiments are carried out within the multimodal analysis framework illustrated in Figure 7.

5.2.1 Results of Video Indexing

Using proposed action-based video indexing, we extract a sequence from one match of Pacific Life Open 2004 between Agassi and Hrbaty as test data to demonstrate the validity of our method. The video is about one-hour-long. Using the scheme illustrated in Figure 8, stroke indexing is first constructed and actions are annotated. Two players are both right-handed gripping. Therefore, the mapping function \( f_{AS} \) between swing types and stroke types is defined as

\[
\begin{align*}
\text{left-swing} & \rightarrow \text{backhand-stroke} \\
\text{right-swing} & \rightarrow \text{forehand-stroke}
\end{align*}
\]

(12)

Then, the indexing production is displayed to users in the form of retrieval result according to the scenario of user’s query. To vividly exhibit such a process, we have developed a prototype of action-based retrieval system. Figure 14 shows the result of retrieving “backhand-stroke”.

5.2.2 Results of Highlights Ranking

Three different video sequences are extracted from one match of French Open 2005 for this experiment. In order to make comparison, we asked six subjects to give the exciting degree for each in-play shot within three sequences. The six subjects are naive to the purpose of the experiment and the only thing they need to do is to rate each shot using a score between 0 and 1 according to its exciting degree they feel. The subjects themselves are 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, and they are able to automatically adjust the exciting degree value to a reasonable state according to the whole video. Then, the mean value of the scores rated by all subjects for each in-play shot is defined as the ground truth. Table 4 lists the detail of each video and its average deviation of subjective evaluation on shots. It can be seen that the mean of subjects’ deviation is 21.7%. Based on the deviation for each sequence, we can calculate the ranking accuracy \( A \) from human perception as

\[
A = 1.0 – \text{Deviation}
\]

(13)

which is the guideline in the latter evaluation experiments.

Table 4. Data for highlights ranking and subjective evaluation

<table>
<thead>
<tr>
<th></th>
<th>time duration</th>
<th># in-play shot</th>
<th>Deviation (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>French_Open_1</td>
<td>32 min</td>
<td>78</td>
<td>20.8</td>
<td>79.2</td>
</tr>
<tr>
<td>French_Open_2</td>
<td>25 min</td>
<td>54</td>
<td>26.1</td>
<td>73.9</td>
</tr>
<tr>
<td>French_Open_3</td>
<td>20 min</td>
<td>41</td>
<td>18.2</td>
<td>81.8</td>
</tr>
<tr>
<td>Mean value</td>
<td></td>
<td>21.7</td>
<td>78.3</td>
<td></td>
</tr>
</tbody>
</table>

To compare the performance of subjective and automatic highlights ranking, evaluation criterion proposed in [23] is utilized for measurement. The highlights ranking accuracy \( HRA \) is defined

\[
HRA = \frac{1}{M} \sum_{m=1}^{M} \left[ R \left( r_m \right) - Q \left( c_m \right) \right] \times 100\%
\]

(14)

where \( M \) is the number of in-play shots, \( r_m \) is the ground truth value for the \( m \)th in-play shot and \( c_m \) is the corresponding value scored by SVR ranking model, \( Q \) represents an optimal quantization process mapping a continuous value in \([0,1] \) to an integer of set \( \{0,\ldots,R\} \) with minimizing the quantization error, \( R \) is the maximum integer level which can be adaptively decided by \( Q \) according to the value distribution of ground truth. Eq. (14) 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.

We use another 85 in-play shots to train the SVR ranking model and then evaluate on test data. Figure 15 shows the comparison between manual and automatic ranking results. The mean value of highlights ranking accuracy for computer is 79.3% in terms of ground truth and evaluation criteria. It must be noted that about 80% accuracy is a remarkable result obtained by computer automatically as there is still 21.7% deviation for subjective ranking from human perception. This result, meanwhile, shows that the selected affective features can reflect human perception to a large extent. With the ranking results, hierarchical browsing is easy to be achieved as described in section 4.3.
5.2.3 Results of Tactics Analysis and Statistics

We conduct this experiment on a complete broadcast video of 1/4 final of Australian Open 2005 (Safin vs. Hrbaty). The duration of video is more than one hour. Because two players are both right-handed gripping, the mapping function $f_{AS}$ between swing type and stroke type is the same with Eq. (12). As described in section 4.4, we identify the last stroke type if it is played by our interested player using proposed action recognition, audio analysis and mapping function. Based on the text analysis, the scoring result of the stroke is obtained. Therefore, we can evaluate the performance of the players and give the statistics for the success rate and failure rate of forehand-stroke and backhand-stroke played by the interested player.

To more vigorously demonstrate our proposed scheme, we analyze both players’ stroke performance in the experiment respectively. Figure 16 shows the success rate of stroke performance for two players, whereas Figure 17 shows the failure rate. Based on two figures, we discover that the misplay of two players is both serious and most of scores are obtained from the failure of the opponents. The most serious weakness of Hrbaty is backhand-stroke compared with Safin. The forehand-stroke of Safin is better than Hrbaty. Therefore, considering the performance of two players as a whole, Safin is superior to Hrbaty. This analysis is demonstrated to be felicitous by the competition result that Safin won the match ultimately.

6. CONCLUSIONS AND FUTURE WORK

Player actions in tennis video reveal the process and tactics of the match. It is an important clue for analysis and understanding of the tennis game. We first present a new method to recognize the basic player actions in broadcast tennis video. Then we integrate action recognition into a multimodal framework and propose a novel approach for semantic and tactics analysis of broadcast tennis video.

Most of existing methods for action recognition are concentrated on the analysis of player appearance. By extensive observations, appearance-based features are not preserved across different videos. Contrastively, from the motion analysis point of view, our method considers the mapping relationship between relative movements of different body parts and the regions in the image plane. A new motion descriptor, which is a group of histograms based on optical flow, is proposed. As our fundamental contribution, we treat optical flow field as spatial patterns of noisy measurements instead of precise pixels displacements. We formulate action recognition as a classification task and employ support vector machine to train the action pattern recognizer.

Our multimodal framework for broadcast tennis video analysis facilitates not only video indexing and retrieval based on players’ stroke types, but also highlights ranking and hierarchical content browsing. Furthermore, tactics analysis and statistics for the success and failure of the stroke performance of players are also implemented in our multimodal framework. To the best of our knowledge, this is the first solution for tennis game tactics analysis based on broadcast video. The proposed analysis schemes are conducted based on the mid-level representation for tennis video rather than the low-level features. We carry out extensive experiments. The results demonstrate that our analysis approach is effective.

For the future work in action recognition, more effective slice partition and elaborate S-OFHs description will be investigated. Our ongoing work is focused on involving more semantic actions in the recognition framework, e.g. overhead-swing in tennis or similar game video such as badminton. As a primary work, action recognition will be applied to more applications such as 3-D scene reconstruction and match content enrichment. For tennis game analysis, future work includes improving the ranking accuracy with the reference of human evaluation by combining audio affective features into the highlights ranking model. More features coming from various channels such as automatic speech recognition, closed caption, and text web broadcasting will be considered to enhance the current framework.

7. REFERENCES


