Modeling Spatiotemporal Relationships between Moving Objects for Event Tactics Analysis in Tennis Videos

Wei-Ta Chu and Wen-Ho Tsai
Department of Computer Science and Information Engineering
National Chung Cheng University, Taiwan
wtchu@cs.ccu.edu.tw, tsaiwenho@gmail.com

Abstract

Evolution of spatial relationships between objects often provides important clues for semantic video analysis. We present a symbolic representation that describes spatiotemporal characteristics and facilitates tactics detection based on string matching. To find typical spatiotemporal patterns of a targeted tactic, we organize training sequences as a tree, and effectively discover frequent patterns from the structure. Tactics detection is conducted by comparing a given test sequence with these frequent patterns. To realize the proposed idea, we develop elaborate audio/video processes to transform broadcasting tennis videos into symbolic sequences, and comprehensively tackle event detection and tactics analysis. We experiment on ten most important tennis championships in the year 2008, and report promising detection results on seven events/tactics. We demonstrate not only the effectiveness of the proposed methods, but also study the impacts brought by the results of tactics analysis.

Keywords: event detection, tactics analysis, spatiotemporal modeling, tennis video analysis

1. Introduction

Recently large amounts of studies have been conducted to extract object's movement for different purposes. Object tracking is performed by the computer vision society to facilitate video synthesis, object modeling, and applications related to surveillance. Motion magnitude and direction are predicted to facilitate efficient coding in video compression researches. In video analysis studies, object's movement is extracted to model semantic events or to describe attractiveness of video clips. Among various types of videos, sports video is the most prominent media where researchers study the role of motion information for object detection, event detection, video summarization, and etc. In this work, we investigate how to model spatiotemporal relationship of two moving objects and conduct tactics analysis for sports videos.

Sports video analysis is a flourish area in multimedia content researches. From the

viewpoint of research, sports games take place in a fixed space and convey clear structure, and therefore ease the development of analysis techniques. Significant amounts of studies for different sports have been conducted on automatic event detection, structure analysis, summarization/highlight extraction, three-dimensional (3D) visualization, and so on. Among these studies, object's movement provides important information from many perspectives. Motion information is fused with audio features to facilitate event model construction in sports videos [1][2][3]. Based on object motion, trajectories of players and ball are modeled as a function of time to describe tactics in soccer games [4]. The work in [5] further analyzes the pose of the player to discover highlights in racket sports videos. Many studies have been conducted on player tracking in broadcasting sports games [17][18], and some works focus on extracting ball trajectories [4][19]. Sophisticated models or tracking techniques have been developed to tackle with multiple objects tracking in complex environments. On the basis of camera motion, semantic shot classification [20] and sports video categorization [21] were developed.

Although motion parameters and object motion characteristics have been widely studied in sports video analysis, spatiotemporal relationships between objects draw relatively little attention before. How to describe and match spatiotemporal relationships between objects is still an ongoing research issue. In this paper, we take more emphasis on modeling the relative spatial relationship between objects along the time axis for detecting events and tactics in sports videos. Although the proposed method is not bounded to a specific sports game, we take tennis games as the main instance. We propose a two-level detection framework to comprehensively detect events and tactics, such as volley, passing ball and moon ball, in tennis matches. At the preprocessing stage, court lines and players are detected, and those positions are mapped to a real-world coordinate system to extract accurate spatiotemporal features. Audio effects, such as laughter and cheer, are also detected to facilitate event analysis. At the first-level event detection stage, a discriminative learning approach is used to detect events, such as ace and double fault, based on audiovisual features. For the events that contain rich interaction between players, such as passing ball and moon ball, the spatiotemporal relationships are transformed into symbolic sequences. Then a pattern mining and matching approach is exploited to accomplish the second-level tactics analysis.

Novelty and contribution of this paper are summarized as follows:

- We model the evolution of relative spatial information between two objects rather than absolute motion information. We demonstrate that evolution of relative movement provides informative clues for tactics analysis.
- Based on transformed symbol sequences, an effective pattern mining

approach is applied to characterize sports tactics. We develop a framework to perform tactics analysis for sports games with two opposite sides, such as tennis, badminton, and table tennis.

- We comprehensively study event detection and tactics analysis in tennis videos. We realize the proposed ideas with the help of elaborate audio/video analysis and effective tactics modeling. Very promising experimental results are reported.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 overviews the system framework, which is divided into two levels, i.e., the first-level event detection and the second-level tactics analysis. In Section 4, we describe video and audio processes for extracting audiovisual features, which are then fed to the discriminative learning module to perform event detection. Section 5 presents the proposed spatiotemporal modeling that describes evolution of spatial relationship, and how the model used in tactics analysis. Performance of the proposed method and comparison between different approaches are described in Section 6. Finally, we conclude this paper with future work in Section 7.

2. Related Work

2.1 Sports Video Analysis

For soccer videos, Ekin et al. [22] detect the goalmouth and analyze shot information to achieve automatic summarization. Yu et al. [23] propose a trajectory-based ball detection method and utilize ball trajectories to analyze soccer videos. To semantically segment soccer games, an approach based on HMMs is proposed to describe the context of shots and to find play and break segments [24].

For baseball videos, Rui et al. [25] extract game highlights through analyzing audience's or anchorpersons' sound reaction. To detect baseball events, a maximum entropy model is proposed to characterize events based on shot transition information [26]. From a different perspective, Zhang and Chang [27] utilize caption information and domain knowledge to detect events. More specifically, Chu and Wu [28] integrate rule-based and model-based methods to comprehensively detect events and develop realistic applications.

For tennis video analysis, Kolonias et al. [29] propose a generic architecture to describe the evolution of tennis events, but they only report limited experimental results. Rea et al. [30] track the player's position and use HMMs to characterize events. Similarly, Han et al. [31] utilize player's spatial information and integrate tennis heuristics to detect events. Kijak et al. [32] model shot transition patterns to detect specific scenes, such as rally and replay. Recently, Huang et al. [41] analyze

match structure and extract aural and visual information, such as audio excitement, close-ups, and slow-motion segments, to detect highlights of tennis matches. Although the aforementioned works conduct considerable studies on tennis video analysis, which is the main instance of this paper, few of them report comprehensive event detection results or fine tactics analysis.

For generic sports video analysis, a series of works [33][34] have been conducted to build a unified framework that fuses visual and aural information to detect events in various sports. The authors propose a mid-level representation framework to bridge the gap between low-level features and high-level events or highlights. Multimodal mechanisms, such as fusing motion vector model and audio keyword model [34], are also devised to achieve semantic analysis.

2.2 Spatiotemporal Modeling

Spatiotemporal characteristics of objects were mostly used to model video shots[35][36][37]. The VideoQ system [35] supports spatiotemporal queries that specify an object's moving trajectory. A pyramid-based structure was proposed to facilitate efficient spatiotemporal matching for sports videos [36]. The techniques of spatiotemporal modeling were also applied to other problems, such as salient frames detection [38] and facial expression modeling [40].

Temporal pattern mining is developed for finding the interactions and interrelations of spatial patterns in [39]. The most frequent spatial patterns are discovered for volleyball games, which initiates the possibility of using mining techniques to conduct sports video analysis. As compared with modeling video shots by object motion, describing objects' interaction and spatial evolution to conduct semantic video analysis draws relatively little attention in the past. In this paper, we would not only elaborate analysis by extracting real-world audiovisual features, but also emphasize the effects of modeling interactions between objects for semantic video analysis.

3. System Framework

3.1 Tennis Events and Tactics

In broadcasting tennis videos, the camera always captures the court view when two players combat against each other. We can focus on the video segments presenting the court and detect which kinds of events or tactics are invoked. In this work, we briefly call these kinds of video segments *plays*. According to tennis regulations, an event in a play would be one of the followings: net approach, rally, ace/unreturned serve, and double fault. More specifically, according to how a player gets his/her points, rally

and net approach can be further categorized into passing ball, moon ball, drop shot, unforced error, or volley.

Figure 1 shows the ontology of tennis events and tactics, and Figure 2 gives illustrated examples about spatiotemporal characteristics of four tactics.

1) Double fault: In double fault, the camera doesn't switch out of the court view after the first failed serve, and a player successively fails the second serve.
2) Ace or unreturned serve: A player successfully serves, and his/her opponent fails to return the ball. In an ace event, the opponent is not able to touch the ball and therefore fails to return. In an unreturned serve event, the opponent barely touches the ball but is still unable to successfully return (the returned ball can't cross the net or is out-of-court).
3) Volley: A player successfully serves and the opponent successfully returns. One of them approaches the net and volleys to get points. Figure 2(a) is a typical example about how players move in a volley tactic. Player A approaches the net and volleys, and player B quickly moves right but fails to return the ball.
4) Passing ball: A tactic to counter the net approach strategy. As shown in Figure 2(b), player A approaches the net to stress player B, but player B quickly hits a line drive so that player A can't even touch the ball.
5) Moon ball: Another tactic to counter the net approach strategy. As shown in Figure 2(c), player A tries to stress his/her opponent by approaching the net, but player B hits a lofty ball so that player A has to go towards the baseline.
6) Drop shot: A tactic to beat the player who stands far from the net. In Figure 2(d), player A hits softly such that the ball drops immediately after crossing the net, and player B is not able to approach the net and returns the ball in time.
7) Unforced error: All rally and net approach plays without passing ball, moon ball, or drop shot are categorized as unforced errors. A player is claimed to make an unforced error when the opponent doesn't approach the net to stress him/her, but he/she makes the ball out-of-court or fails to return the ball.

Note that only action that makes some player get the point in a play is considered. For example, player A may approach the net and smash the ball, but player B successfully hits back a line drive to get the point. We would say that this play contains a passing ball, since that's the factor player B gets his points.

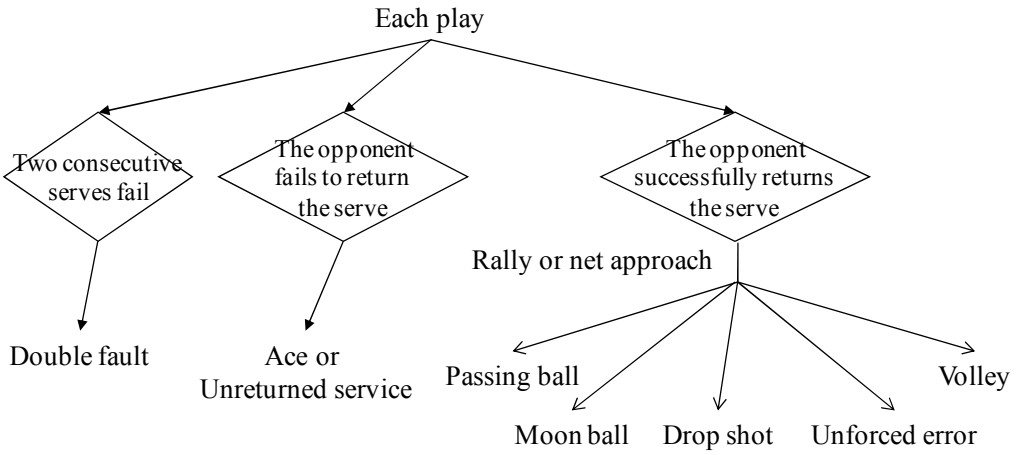


Figure 1. Ontology of tennis events and tactics.

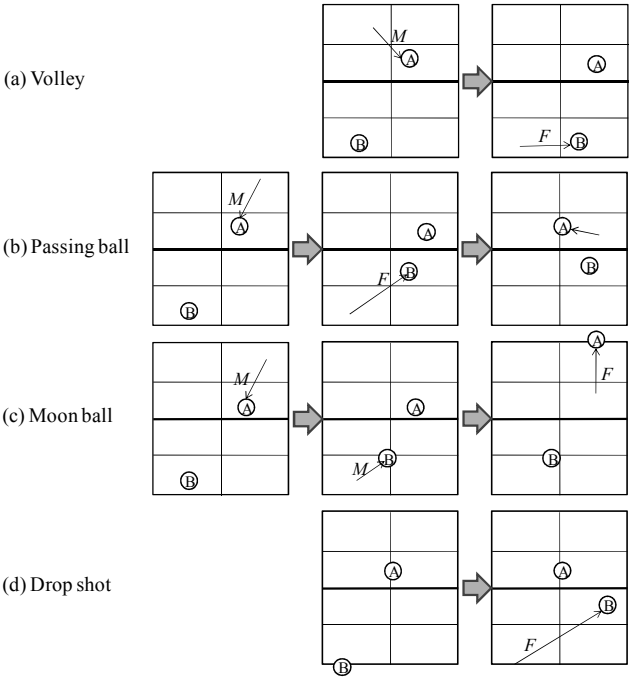


Figure 2. Illustrated examples of spatiotemporal characteristics of four tennis tactics.

3.2 Overview of Framework

Figure 3 shows the system framework that is composed of audiovisual feature extraction and two detection modules. From visual information, we segment videos into semantics-meaningful clips and extract high-level features, such as player's moving direction and speed. From aural information, we model special audio effects to be the basis for event detection.

A two-level detection approach is developed to detect the prescribed events and tactics. The audiovisual features extracted from each play are concatenated for discriminative learning in the first-level event detection. The events detected by the

first-level detection module either contain simple action, such as ace and double fault, or are needed to be discriminated further. There are complex interactions between players in different tennis tactics, and therefore the consideration of spatial evolution is needed at the second-level tactics analysis. Spatial evolution is transformed into symbolic representation, and spatiotemporal relationships between players are used to detect tactics.

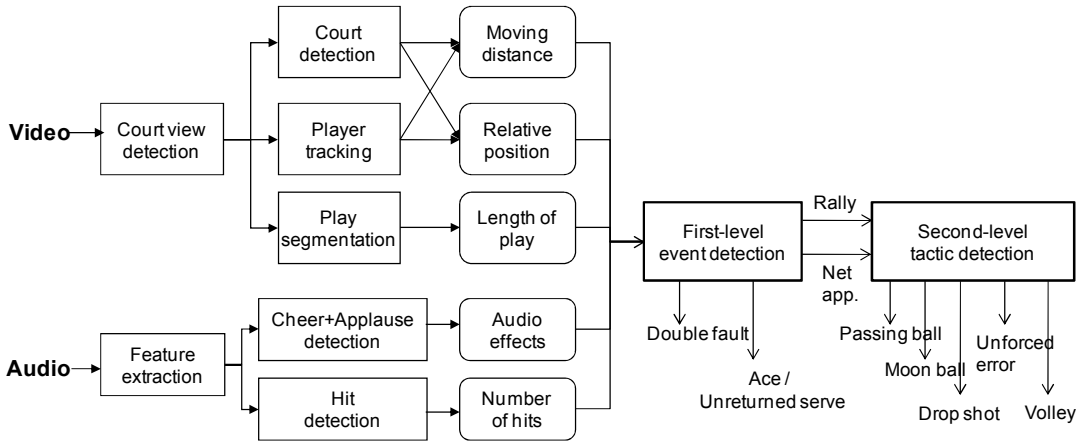


Figure 3. Tennis event tactics analysis framework.

4. First-Level Event Detection

In this section, we first describe audiovisual feature extraction and then develop a discriminative learning approach to detect four events, including ace/unreturned serve, double fault, rally, and net approach.

4.1 Video Processing

Because shot changes in broadcasting tennis matches are usually simple, we apply a typical shot change detection method [8] based on color histogram difference between adjacent video frames to detect shot boundaries.

4.1.1 Court view detection

Tennis videos are composed of court view shots and non-court view shots. Court view shots contain a large ratio of pixels with the court's colors, which are dominant colors in frames. In this work, the adaptive playfield detection method proposed in [9] is referred. We model the HSI (hue, saturation, intensity) histograms from training data by a Gaussian mixture model (GMM) λ, which consists of M Gaussian densities:

$$p(\xi|\lambda) = \sum_{i=1}^{M} w_i b_i(\xi), \quad (1)$$

where ξ is the color vector of a pixel and w_i is the weight of the ith mixture. The parameters of mixture components b_i and weights are estimated by the EM

(expectation maximization) algorithm [9].

Based on the obtained GMM, we determine two dominant color ranges. The reason for selecting two ranges is that a tennis court often has two dominant color ranges, such as the court of US open. Colors in the selected ranges are set as the initial guess of dominant colors. As the analysis proceeds, the ranges are adaptively adjusted according to newly-decoded video frames. In our work, we periodically adjust the dominant color ranges every ten minutes [6].

For each video shot, without loss of generality, we extract the first frame as its keyframe to do court view detection. According to the ratio of the number of dominant-color pixels to that of the whole frame, we can classify shots into court view or non-court view. Court view and non-court view shots respectively represent *plays* and *breaks* in tennis matches. Details of implementation please refer to [6].

4.1.2 Court line detection

For the (suspected) court view shots, we detect the court lines based on the techniques of line detection and camera calibration [10]. We first detect white pixels in frames, and then apply a standard Hough transform line detector to find the white lines. We can obtain the court position if all court lines are perfectly detected. However, because of noises caused by players or characteristics of different stadiums, many misses or false alarms may occur in line detection.

Fortunately, the specification of a tennis court is fixed in all matches. Based on the intersections of detected lines, we can map them with a predefined court model and find the parameters of camera. Since this mapping is plane-to-plane, the corresponding transformation can be seen as an eight-parameter perspective transformation [11]. To solve these eight parameters, we choose at least four line intersections in video frames to map the corresponding ones in the predefined court model. After calibration, we map the real-world court onto the displayed plane and obtain all court lines. Details of the court detection process please refer to [10].

4.1.3 Player detection/tracking

For court view shots, we detect and track player's positions to characterize each play. The essential idea of player detection is to find the region that has pre-defined uniforms. For a tennis match, we first manually select two players' uniform regions and calculate these regions' HSI histograms, respectively. From the samples that are taken every ten video frames, we track two players' positions by comparing the histograms of neighborhood in consecutive frames. The basic idea is the same as motion estimation in video compression. We find the minimum bounding box that

covers a player, and the midpoint of the bottom line of the bounding box is set as the player's position.

By combining the results of court detection and player detection, we can map the player's position onto a virtual map, which describes where the player is in the court. Figure 4 shows two sample results of player detection. The right bottom of the figures shows player's relative positions in the court.

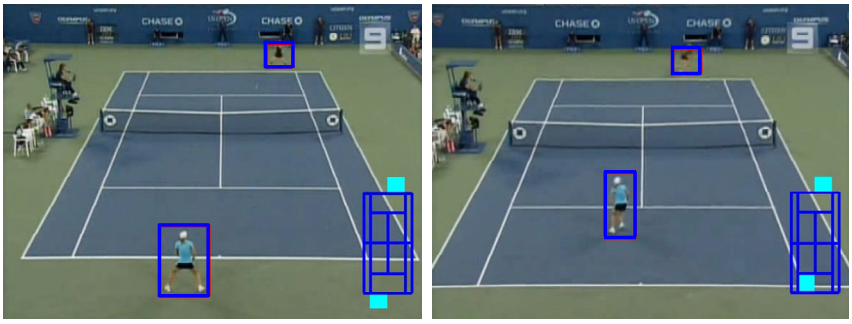


Figure 4. Two sample results of player detection.

4.2 Audio Processing

Aural information often provides significant cues for event detection. In this work, we detect the sound effect of applause mixed with cheer. Audiences are often kind to give applauses or cheers after good plays, such as ace or rallies. On the other hand, they often keep quiet if a player invokes a double fault. Sound effects recognition has been widely studied in recent years [12][13]. In this work, we apply an HMM-based method to model/detect applause mixed with cheer [13].

Two types of training data are collected from several tennis matches. The first dataset includes the clamor that is mixed by cheer and applause after plays. The second dataset consists of the sounds other than cheer and applause in playing, including quiet, anchorperson's speech, player's shouts, and sounds of racket hits. For these audio data, audio features are extracted for modeling, including energy, band energy ratio, zero-crossing rate, frequency centroid, bandwidth, and mel-frequency cepstral coefficient (MFCC) [14]. These features have been shown to be beneficial to sound effects recognition [12]. We respectively construct an HMM for these two types of audio data based on the Baum-Welch algorithm. After event modeling, how likely an audio segment belongs to an audio effect is evaluated. Details of sound effects modeling and testing please refer to [13].

4.3 First-Level Event Detection

We segment a tennis match into clips with each representing a single play. With the helps of audio/video processes described above, we extract high-level features from

each play to represent tennis matches:

- Player's relative position in the court: The court can be partitioned into the region near the net and the region near the baseline. If a player ever moves to the region near the net, this play most likely contains a net approach. A binary feature is extracted to represent whether players step into the region near the net or not.
- Moving distance of players: We map the player's position to the real-world coordinate system and calculate the moving distance between every ten frames. Generally, the player moves more in rallies than in aces. Overall, this feature is calculated by averaging the moving distances of two players.
- Length of play: It's apparent that the play length of a baseline rally is likely longer than that of an ace. With court view detection, we can easily calculate the length of a play.
- Applause/cheer sound effects: According to tennis etiquette, the audiences clamor after ace or unreturned serve but not after double fault. A binary feature is extracted to represent whether this kind of audio effect occurs.

Many studies have demonstrated that the discriminative learning approach achieves promising detection results [15]. On the basis of the features described above, we exploit a discriminative learning method in the first-level event detection module. Audiovisual features are concatenated as vectors to describe a play, and the problem of event detection is transformed into classifying a given test feature vector into one subspace, which is part of the hyperspace constructed from features representing a specific tennis event.

We exploit support vector machines (SVM) to construct a 4-class classifier. We construct the SVM classifier by using LIBSVM [16], which supports multiclass classification. At the detection stage, a play is represented as a feature vector and is detected as a double fault, ace/unreturned server, rally, or net approach.

5. Second-Level Tactics Analysis

Rally and net approach events are further examined in the second-level detection module. In rallies, players often have some strategies to beat opponents. For example, a player may suddenly hit the ball lightly so that the ball drops immediately after crossing the net. This makes his opponent barely catch the ball and cause an error. In net approaches, a player may approach the net to volley the ball or give a drop shot. On the other hand, the opponent can hit back quickly (if he can) to beat the net approach strategy and gets points by a passing shot.

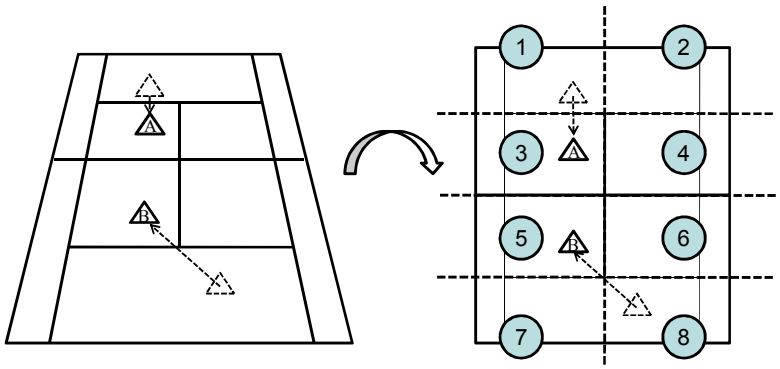
The tactics described above are involved with interaction between two players.

The evolution of movements reveals the occurrence of different tactics. In this section, we first transform video content into a symbolic representation, and then an effective structure is developed to find typical moving patterns of each targeted tactics. With typical moving patterns, tactics analysis is accomplished by performing approximate sequence matching.

5.1 Symbolic Representation

To describe spatial evolution, we uniformly examine one out of ten video frames. The characteristics of moving players are described in three aspects:

- Location: We segment the tennis court into eight regions, as shown in Figure 5. Note that the court on screen and the positions of players have been calibrated to the real-world coordinate, by the method described in Section 4.1.3. The position of each player in a sample is spatially quantized into one of the regions. The region where a player locates at the *t*th sample is denoted as R^t.
- Direction: Based on the positions of a player in the current sample and the previous sample, we can estimate his moving direction. Horizontal and vertical directions are described separately. Generally, we quantize the horizontal direction into "left", "right", and "still"; and quantize the vertical direction into "up", "down", and "still." The notations of D_h^t and D_v^t denote the horizontal and vertical direction at the *t*th sample, respectively.
- Speed: Based on the distance between a player in the current sample to the previous sample and the time difference between two samples, we can calculate this player's moving speed. Similarly, we quantize the moving speed into "fast", "medium", and "still." The notations of S_h^t and S_v^t denotes the horizontal and vertical speed at the *t*th sample, respectively.



(a) 2D court image (b) Real-world court coordinate system

Figure 5. (a) The 2D court image; (b) player movements mapping into the real-world coordinate system.

Based on the information extracted above, the symbol sequence of each sample is described in six parts: (1) position of the object in the previous frame; (2) horizontal moving direction; (3) horizontal moving speed; (4) vertical moving direction; (5) vertical moving speed; and (6) position of the object in the current frame. The symbol sequence for the *i*th player at the *t*th sample is represented as $Q_i^t = \langle i, R^{t-1}, D_h^t, S_h^t, D_v^t, S_v^t, R^t \rangle$. At each sample, the moving characteristics of both players are transformed and concatenated as $\langle Q_A^t, Q_B^t \rangle$, in which the player IDs are *A* and *B*, respectively.

In tennis video analysis, the locations of each player are quantized into four regions, their speeds are quantized into three levels (fast, medium, slow), and their vertical/horizontal movements are also described in three directions (up/left, down/right, still). From the example of Figure 5(b), solid-line triangles and dash-line triangles denote the locations of players in current and previous samples, respectively. Because locations of players are mapped to the real-world coordinate system, we can precisely identify the location, moving speed, and moving direction of players. In this example, we obtain that player A moves from region 1 to region 3 in medium speed, and player B moves from region 8 to region 5 in fast speed.

Table 1 shows symbol definitions of the proposed tennis analysis system. By these definitions, the symbol pair $\langle Q_A^t, Q_B^t \rangle = \langle A, 1, R, F, D, F, 4 \rangle \langle B, 8, L, S, N, S, 7 \rangle$ means that player *A* moves right-down quickly from region 1 to region 4, and player *B* moves left slowly from region 8 to region 7.

Table 1. Definitions of symbols used in tennis tactics analysis.

Types	Meaning (Symbols)		
Object	Object ID (A, B)		
Location	Regions (1,2,…,8)		
Horizontal moving direction	Still (N)	Left (L)	Right (R)
Vertical moving direction	Still (N)	Up (U)	Down (D)
Speed of horizontal & vertical moving	Fast (F)	Medium (M)	Slow (S)

To further describe relative moving characteristics between two players, their locations, moving speeds, and moving directions are jointly considered. The symbol pair $\langle Q_A^t, Q_B^t \rangle$ at each sample is further transformed into an appropriate meta-symbol. Figure 6 shows some examples of relative moving patterns in tennis videos. In Figure 6(a), player *A* stands in the left-top region, and player *B* stands in the right-bottom region. Both may move in three directions, in different speeds. For example, in the

moving situation as illustrated by the bold lines, player *A* moves down quickly and player *B* moves left-top quickly, and this symbol pair $\langle Q_A^t, Q_B^t \rangle$ is further denoted by a meta-symbol. If player *A* moves down quickly but player *B* moves left-top slowly, another meta-symbol will be used to describe this case. Therefore, the number of meta-symbols that represent possible moving relationships between two players in Figure 6(a) is 3 (player *A*'s moving directions) × 3 (player *B*'s moving directions) × 3 (player *A*'s moving speed) × 3 (player *B*'s moving speed) = 81. Similarly, each possible relationship in different cases (Figure 6(b), (c), (d), and others) is represented by a meta-symbol.

For a video clip, we uniformly take samples and transform each sample into a meta-symbol. Spatial evolution between two players is represented by a meta-symbol sequence, which we briefly call it a *moving sequence* in the following.

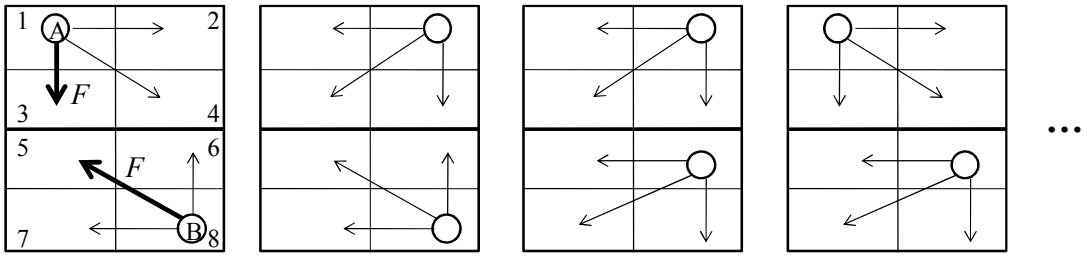


Figure 6. Examples of moving patterns between two objects in tennis videos.

5.2 Moving Pattern Mining

For each targeted tactic, we collect thirty plays and transform them into meta-symbol sequences as training data, which provide the templates of each tactic. Note that we only consider the last 5-second segment in each play, because only the actions in the last part of a play indicate what tactic drives the point. At the stage of tactics detection, the meta-symbol sequence corresponding to a test video clip is compared with templates, and is claimed to contain a specific tactic if it has the closest relationship to some template.

The method described above is not efficient and is easily affected by variations of spatiotemporal relationships related to a tactic. Therefore, we introduce the idea of mining frequent patterns from meta-symbol sequences. The mined frequent patterns are viewed as canonical representations of a tactic, and are used as the foundation of tactics detection.

The essential idea of finding frequent patterns is that specific moving patterns would occur in a tactic. For example, the "passing ball" tactic often occurs when player *A* approaches the net, and player *B* hits the ball quickly to pass through player

A so that he/she cannot touch the ball. Player *A* often has to move back quickly to save this point. The idea of finding typical moving pattern is also motivated from the 58 winning patterns recommended by United States Tennis Association [7].

To efficiently find typical representations, we construct a moving sequence tree for each targeted tactic, based on the meta-symbol sequences transformed from training data. The node in a moving sequence tree corresponds to a meta-symbol, and each node is associated with a number that indicates the times this meta-symbol appears in moving sequences. The procedure of moving sequence tree construction is described as follows.

Algorithm 1: Moving Sequence Tree Construction

Input: A set of meta-symbol sequences $O = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_N\}$ corresponding to a specific tactics. Let $\boldsymbol{o}_i(j)$ denote the *j*th meta-symbol of the sequence $\boldsymbol{o}_i$, *j* = 1, 2, …, *M*.

1 For *i* = 1 to *N*

2 If $\boldsymbol{o}_i(1)$ is not a root of any existing moving sequence tree, then

Initialize a moving sequence tree rooted by $\boldsymbol{o}_i(1)$. The following procedures will act on this newly-initiated tree. Set the appearance count associated with the root node as 1.

Otherwise, increase the appearance count associated with the node by 1. The following procedures will act on this existing tree.

3 For *j* = 2 to *M*

4 If $\boldsymbol{o}_i(j)$ is not a child of $\boldsymbol{o}_i(j-1)$, create a child node $\boldsymbol{o}_i(j)$ for $\boldsymbol{o}_i(j-1)$, and set the associated appearance count as 1. Otherwise, increase the appearance count associated to $\boldsymbol{o}_i(j)$ by 1.

Table 2 shows a set of meta-symbol sequences, and Figure 7 illustrates the corresponding moving sequence tree. In each node, the first item denotes the meta-symbol, and the number indicates times the meta-symbol being visited by the tree construction process.

After constructing the moving sequence tree, we traverse this tree by the depth-first-search algorithm and calculate "support" for each path from the root to a leaf. The value of support is calculated by summing the number associated with each node on a path. For example, support value of the leftmost path in Figure 7 is 5+4+2+2+2 = 15. Based on this moving sequence tree, the paths with support values larger than a minimum support threshold *min_sup* are extracted as frequent moving patterns. There are totally 30 meta-symbols (6 sequences, 5 meta-symbols in each

sequence) in Table 2. If the value of *min_sup* is set as 0.3, the sequence "a,b,c,d,e" with support value larger than 30×0.3=9 is claimed as a frequent moving pattern. On the contrast, the sequence "a, h, f, h, i" in the rightmost of the tree doesn't count to be frequent.

Table 2. Examples of meta-symbol sequences

Video ID	Meta-symbol sequence
S1	a, b, c, d, e
S2	a, b, c, d, e
S3	a, b, f, g, e
S4	a, a, f, h, e
S5	a, b, f, d, e
S6	a, h, f, h, i

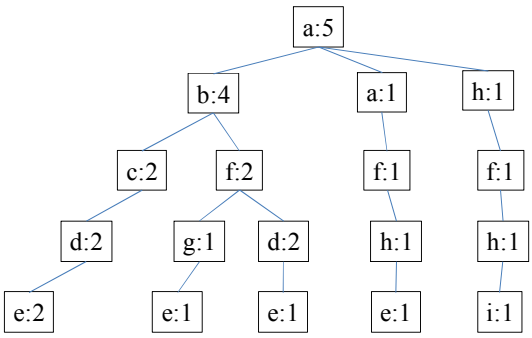


Figure 7. The constructed moving sequence tree corresponding to Table 2.

Table 3. Number of frequent moving patterns determined by the processes with different *min_sup* values.

min_sup	Passing ball	Moon ball	Drop shot	Volley
10%	24	8	10	16
30%	18	7	6	12
50%	6	2	2	7

Larger *min_sup* value means that higher thresholds are set to determine frequent moving patterns, and therefore fewer frequent patterns are obtained. Because passing ball and volley are involved with more complex interactions between players, generally more frequent moving patterns are found. Table 3 shows the number of frequent moving patterns determined by the processes with different *min_sup* values. In the experiment section, we will further study the influence of different *min_sup* values on the performance of tactics detection.

The essence of utilizing frequent moving patterns to characterize tennis tactics is

that similar spatial evolution will occur in the same tactic. Figure 8 illustrates a frequent moving pattern mined from training data of passing balls. Note that this moving pattern is represented as a meta-symbol sequence in our system, and Figure 8 just visualizes the corresponding spatiotemporal relationship. From Figure 8(c) to (e), it's likely that the following actions took place: Figure 8(c) – player A approaches the net, while player B stands near the baseline; Figure 8(d) – player A hits softly to make the ball drop immediately after crossing the net, and player B approaches the net fast and save this point; Figure 8(e) – player A returns the ball quickly to through the coverage of player B such that player B has to move towards the baseline to save the ball. This play ends immediately after the last action. We can imagine that this is a typical case of a passing ball.

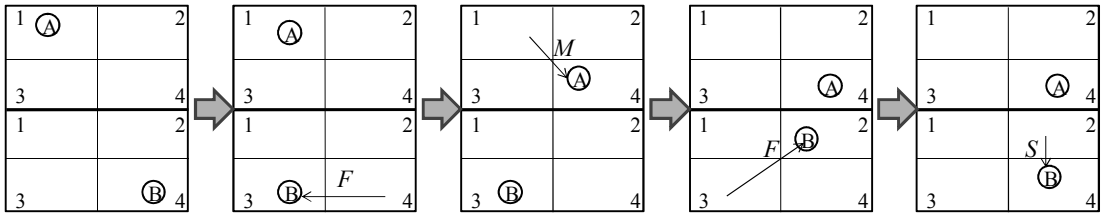


Figure 8. A frequent moving pattern mined from training data of passing balls.

5.3 Pattern Matching

After finding frequent patterns for each tactic, we compare a test meta-symbol sequence with them to accomplish detection. Similarity between sequences is typically evaluated based on string matching algorithms. Among the widely-studied matching algorithms, finding the longest common subsequence (LCS) is one of the methods that jointly finds the optimal matching between two sequences and evaluates the extent of matching. Therefore, a dynamic programming approach is used to find LCS in this work.

For an unknown video clip that conveys the meta-symbol sequence $\boldsymbol{Y}$, we compare $\boldsymbol{Y}$ with the mined patterns $\boldsymbol{X}=\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N\}$. The play with the meta-symbol sequence $\boldsymbol{Y}$ is identified as the class corresponding to the pattern $\boldsymbol{X}_{i^*}$ if

$$i^* = \arg\max_i |LCS(\boldsymbol{Y}, \boldsymbol{X}_i)|, \quad (2)$$

where $|LCS(\boldsymbol{Y}, \boldsymbol{X}_i)|$ denotes the length of the longest common subsequence between $\boldsymbol{Y}$ and $\boldsymbol{X}_i$.

The description above reveals conventional string-based matching, in which a symbolic sequence represents just spatial relation or temporal relation. However, to evaluate the sequences describing spatiotemporal relationship, we not only have to find the LCS, but also have to constrain the distance between matched symbols. A

targeted tactic should contain continuous spatiotemporal evolution similar to typical patterns. From the example in Figure 9, we see that the test pattern has the same-length LCSs to two different frequent patterns. However, the test pattern continuously matches with the first frequent pattern at the first four meta-symbols, while it matches with the second frequent pattern at the first two and the last two meta-symbols. The breach of matched sequence in the second case means that the test pattern is similar to a tactic at the beginning, but breaks down in the middle of the progress of a play. On the contrary, the test pattern continuously matches with the first frequent pattern, and only differs with it at the end, which is often a short period before the camera switches out of the court view. According to the description above, intervals between matched symbols degrade the degree of correspondence between two sequences. We should further take this characteristic into account in tactics detection.

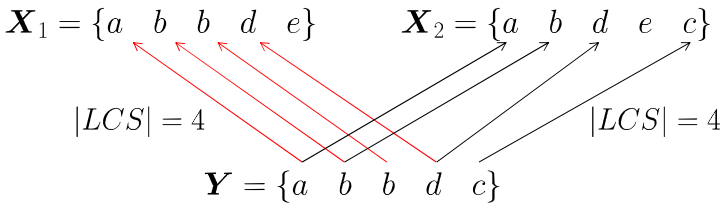


Figure 9. An example of pattern matching.

For the meta-symbol sequence $\boldsymbol{Y}=\{y_1,y_2,\ldots,y_m\}$ and a frequent pattern $\boldsymbol{X}=\{x_1, x_2, \ldots, x_n\}$, assume that the LCS between them is $\boldsymbol{Z}=\{z_1, z_2, \ldots, z_{|LCS(\boldsymbol{Y},\boldsymbol{X})|}\}$. If $z_1 = y_i$ and $z_2 = y_j$, the number of j-i means the distance between two matched symbols in the test meta-symbol sequence, and is denoted by g_1. Similarly, we can calculate the distances $g_2, g_3, \ldots, g_{|LCS(\boldsymbol{Y},\boldsymbol{X})|-1}$ based on $\boldsymbol{Z}$ and $\boldsymbol{Y}$. With the sequence $\boldsymbol{Z}$, the probability of the meta-symbol sequence $\boldsymbol{Y}$ corresponding to the class with the pattern $\boldsymbol{X}$ is defined as

$$P(C|LCS(\boldsymbol{Y},\boldsymbol{X})) = \prod_{i=1}^{\ell-1} e^{-(g_i-1)}, \tag{3}$$

where ℓ is $|LCS(\boldsymbol{Y},\boldsymbol{X})|$, and C is the tactic class corresponding to the pattern $\boldsymbol{X}$. With this definition, the play with the meta-symbol sequence $\boldsymbol{Y}$ is identified as the class C_{i^*} corresponding to the pattern $\boldsymbol{X}_{i^*}$ if

$$i^* = \arg\max_i P(C_i|LCS(\boldsymbol{Y},\boldsymbol{X}_i)). \tag{4}$$

Corresponding to the example in Figure 9, the probabilities corresponding to the tactic presented by two frequent patterns are:

$$P(C|LCS(\boldsymbol{Y},\boldsymbol{X}_1)) = e^{-(1-1)} \times e^{-(1-1)} \times e^{-(1-1)} = 1, \tag{5}$$

$$P(C|LCS(\boldsymbol{Y},\boldsymbol{X}_2)) = e^{-(1-1)} \times e^{-(2-1)} \times e^{-(1-1)} = e^{-1}. \tag{6}$$

Therefore, the test pattern is viewed as conveying the tactic as that in the first

frequent pattern.

With the determined moving patterns, rally and net approach are further categorized into passing ball, moon ball, drop shot, volley, and unforced error. Each of the first four tactics can be described by several frequent moving patterns. Unforced errors have no specific pattern, and the plays that have very small probability to other four tactics are claimed to contain unforced errors. According to the descriptions above, we can more precisely express tactics detection in the following:

Assume that $\boldsymbol{X}_{1,j}$ is the *j*th frequent moving pattern corresponding to the tactic C_1. The targeted tactics are {C_1: passing ball, C_2: moon ball, C_3: drop shot, C_4: volley}. The tactic in the play with a meta-symbol sequence $\boldsymbol{Y}$ is identified as

$$\begin{cases} C_{i^*}, & \text{if } (i^*, j^*) = \arg\max_{i,j} P(C_i | LCS(\boldsymbol{Y}, \boldsymbol{X}_{i,j})) \text{ and} \\ & \quad P(C_{i^*} | LCS(\boldsymbol{Y}, \boldsymbol{X}_{i^*,j^*}) > \epsilon, \\ \text{unforced error}, & \text{otherwise}, \end{cases} \tag{7}$$

where the parameter ϵ is the threshold of the minimal required probability for a play to be claimed as one of the four targeted tactics. In this work, the parameter ϵ is set as e^{-10}, which means that only the meta-symbol sequence has less than ten time gaps to the most similar frequent moving pattern can be categorized into one of the four tactics.

6. Experimental Results

6.1 Performance of Event Tactics Detection

Ten important tennis matches in the year 2008 are used to evaluate the proposed event tactics detection methods. Table 4 shows the detailed information of evaluation data. There are totally about 15 hours of videos, including more than 4400 plays in the evaluation data. The courts in these matches include grass (Wimbledon), clay (French Open), and hardcourt (Others).

- Overall detection performance

The number of frequent moving patterns is determined by the value of *min_sup*. When the value of *min_sup* is set higher, fewer moving patterns will be claimed as being frequent. We study the variation of detection performance when different *min_sup*s are set, say 0%, 10%, 30%, and 50%.

Figure 10 shows overall detection performance in terms of F-measure with different *min_sup* values. F-measure jointly considers precision and recall values, and is defined as follows. Larger F-measure value means better performance in both precision and recall values.

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

When larger *min_sup* values are set, fewer moving patterns are determined to be frequent, and fewer plays are correctly detected. Although precision value of detecting tactics increases when larger *min_sup* values are set, those plays that don't match any frequent moving pattern are detected as unforced errors, which degrades the precision of detecting unforced errors. Therefore, F-measure values for larger *min_sup* cases decrease. Overall, F-measure values over ten tennis matches and seven events/tactics are 0.74, 0.71, 0.64, and 0.57, for *min_sup* = 0%, 10%, 30%, and 50%, respectively.

From Figure 10, we can see the detection performances for M9 and M10 are especially worse than others. The main reason for this is that French Open is held in clay tennis courts, and white court lines are often annoyed by red dust. Errors in count line detection cause the errors of court segmentation, and therefore degrade accuracy of events/tactics detection.

Table 4. Detailed information of evaluation data.

ID	Tennis match	Duration	Number of plays
M1	2008 Beijing Olympics Men's Single – Finals R. Nadal vs. F. Gonzalez: 6-3, 7-6, 6-3	1 hr 29 min	407
M2	2008 Beijing Olympics Women's Single – Finals E. Dementieva vs. D. Safina: 3-6, 7-5, 6-3	1 hr 18 min	393
M3	2008 US Open Men's Single – Finals R. Federer vs. A. Murray: 6-2, 7-5, 6-2	1 hr 15 min	389
M4	2008 US Open Women's Single – Finals S. Williams vs. J. Jankovic: 6-4, 7-5	1 hr 23 min	415
M5	2008 Australian Open Men's Single – Finals N. Djokovic vs. J.-W. Tsonga: 4-6, 6-4, 6-3, 7-6	1 hr 50 min	573
M6	2008 Australia Open Women's Single – Finals M. Sharapova vs. A. Ivanovic: 7-5, 6-3	1 hr 17 min	395
M7	2008 Wimbledon Men's Single – Finals R. Nadal vs. R. Federer: 6-4, 6-4, 6-7, 6-7, 9-7	2 hr 5 min	587
M8	2008 Wimbledon Women's Single – Finals V. Williams vs. S. Williams: 7-5, 6-4	1 hr 53 min	498
M9	2008 French Open Men's Single – Finals R. Nadal vs. R. Federer: 6-1, 6-3, 6-0	1 hr 8 min	377
M10	2008French Open Women's Single – Finals A. Ivanovic vs. D. Safina: 6-4, 6-3	1 hr 46 min	412
Total		15 hr 24 min	4446

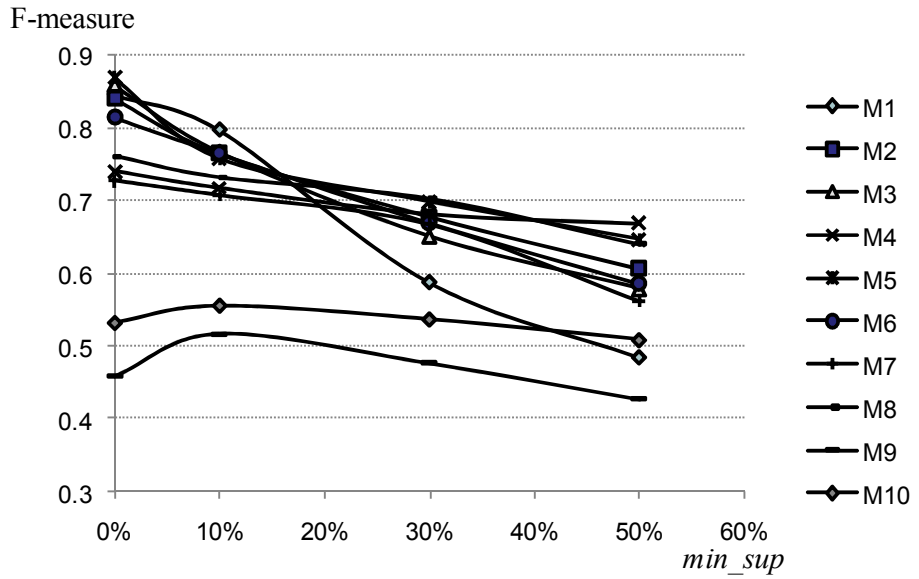


Figure 10. Overall detection performance with different *min_sup* values.

- Evaluation of execution time

Although increasing *min_sup* values degrades detection performance, we can largely save time for events/tactics detection. Table 5 shows comparison of the ratios of execution time and F-measure values. The ratios are calculated by dividing the execution time (F-measure) with frequent moving patterns derived from the case of *min_sup*=10% by that from the case of *min_sup*=0%. The case of *min_sup*=0% corresponds to that all meta-symbol sequences in training data are exhaustively matched with a given test sequence.

$$R_t = \frac{\text{execution time of the case of } min\_sup = 10\%}{\text{execution time of the case of } min\_sup = 0\%}, \tag{9}$$

$$R_f = \frac{\text{F-measure of the case of } min\_sup = 10\%}{\text{F-measure of the case of } min\_sup = 0\%}. \tag{10}$$

From the tennis match M1 in Table 5, for example, we see although the F-measure performance in the case of *min_sup* = 10% degrades to 94% of the exhaustive approach, we only need half of execution time. Overall, only 58% of execution time is needed in detection with frequent moving patterns derived from *min_sup*=10%, while 96% accuracy relative to the exhaustive approach can be maintained.

Table 5. Relationships between execution time ratio and F-measure ratio.

ID	Ratio of execution time (R_t)	Ratio of F-measure (R_f)
M1	0.5	0.94
M2	0.52	0.91
M3	0.58	0.89
M4	0.53	0.97
M5	0.47	0.87
M6	0.56	0.94
M7	0.60	0.97
M8	0.67	0.96
M9	0.64	1.13
M10	0.66	1.04
Overall	0.58	0.96

- Detailed events/tactics detection performance

To see the variation of detection performance, we show confusion matrices of detection results. Due to space limitation, we only show confusion matrices for M2 and M3, in the case of *min_sup*=10%. We think that, when *min_sup* is set as 10%, better balance between execution time and detection performance is achieved.

In Tables 6 and 7, columns mean the truth events/tactics, and the rows mean the detection results. We can see that most plays are correctly detected, and the detection performance is very promising. The worst detection performance in M2 lies on passing balls. Ten passing balls are erroneously detected as unforced errors. Plays with a passing ball tactic are often involved with relatively complex interaction between two players. Thus, the detection performance for passing balls is slightly worse than others. For the same reason, plays with volley have slightly worse performance. There are similar trends in Table 7.

Table 6. Confusion matrix of events/tactics detection for M2.

	Unforced error	Passing ball	Moon ball	Drop shot	Volley	Ace	Double fault
Unforced error	**27**	10	4	2	4	0	0
Passing ball	0	**28**	4	0	0	0	0
Moon ball	0	2	**30**	1	0	0	0
Drop shot	0	1	0	**46**	5	0	0
Volley	6	3	5	2	**53**	0	0
Ace	0	0	0	0	0	**10**	0
Double fault	0	0	0	0	0	0	**9**

Table 7. Confusion matrix of events/tactics detection for M3.

	Unforced error	Passing ball	Moon ball	Drop shot	Volley	Ace	Double fault
Unforced error	**50**	10	5	5	3	0	0
Passing ball	0	**47**	4	0	0	0	0
Moon ball	0	2	**20**	1	0	0	0
Drop shot	0	1	0	**23**	5	0	0
Volley	6	3	5	2	**31**	0	0
Ace	0	0	0	0	0	**15**	0
Double fault	0	0	0	0	0	0	**6**

- Comparison

Comprehensive comparison of tactics detection performance is important but is hardly to be achieved, because there is no standard dataset and the targeted tactics are different in different works. Therefore, we are just able to compare the most relevant works in terms of methodology.

Wang and Parameswaran [42] proposed a Bayesian network approach to classify tennis matches into 58 winning patterns based on ball movement. Theoretically, if we can detect all landing positions of the ball in the whole progress of a tennis match, we can infer almost all actions that would occur. However, detecting the tennis ball is difficult, especially when there is camera motion and video quality degradation due to compression. That's why most tactics analysis works focus on player's movement.

Zhu et al. [5] track player's trajectory and recognize player's action as forehand-stroke or backhand-stroke. Based on player's trajectory, game highlight is generated. Based on action recognition results, interesting statistics were demonstrated to show the relationship between the ratio of forehand to backhand strokes and game results. However, the reported results only show rough statistics, and their targeted tactics are not clearly defined. Moreover, interaction between players is not considered.

Sudhir et al. [44] developed models to detect court lines and players. They claim that this information facilitates detection of specific tactics, such as volley and passing shot. However, they focused on court line detection and player tracking, and just describe the possibility of tactics detection without real implementation.

Wang et al. [43] jointly consider movements of two players, and discover salient moving patterns from games. Although they detect frequent moving patterns and mention that they would be caused by specific tactics, the mined frequent moving patterns are not used for tactics detection.

Table 8 summarizes comparison of different methods in terms of (1) what kind of

object is detected; (2) whether interaction between players is used; and (3) whether targeted tactics are clearly defined and detected. As compared to other methods, we model spatiotemporal relationships between players and clearly detect events and tactics in tennis videos.

Table 8. Comparison of different tennis analysis methods.

	[42]	[5]	[44]	[43]	Ours
Object	Ball	Player	Player	Player	Player
Interaction	No	No	No	Yes	Yes
Definition and detection of tactics	No	Subtle	No	No	Clear

6.2 Discussion

- Game analysis based on results of tactics detection

Results of events/tactics detection can be applied to many aspects. For example, with the detected boundaries of plays and the associated events/tactics, event-on-demand services can be provided. For game abstraction, exciting events, such as ace and passing ball, can be especially selected to generate game highlights.

The major contribution of this work is that we model spatiotemporal relationships between players and therefore detect tennis tactics that have not been studied well before. We argue that deeper analysis can be achieved when we achieve such elaborate tactics detection. A good indicator about a player's performance is the number of unforced error. By manually selecting uniforms of players and continuously tracking them in a play, we can determine whether the player in white or the player in black, for example, draws an unforced error. With this information, we can calculate the number of unforced errors issued by two players and study the correspondence between unforced errors and game results.

Table 9 shows the number of unforced errors in seven tennis matches and the corresponding game results. Surprisingly, although the number of detected unforced errors may not exactly match game ground truth, we can see high correlation between it and the corresponding game result. The player who issued fewer unforced errors won the game. This correspondence shows that detecting tactics like unforced error largely approaches semantic video analysis. These detection results provide a different perspective for semantic analysis other from that in [5], which conducted analysis based on player action recognition.

Table 9. The correspondence between the number of unforced errors and game results.

ID	Number of unforced error	Result
M1	Red (R. Nadal) vs. White (F. Gonzalez) Number of unforced error – 6 : 21	R. Nadal Won
M2	White (E. Dementieva) vs. Black (D. Safina) Number of unforced error – 11 : 16	E. Dementieva Won
M3	Red (R. Federer): White (A. Murry) Number of unforced error – 16 : 34	R. Federer Won
M5	Blue (N. Djokovic) vs. Black (J.-W. Tsonga) Number of unforced error – 39 : 45	N. Djokovic Won
M6	White (M. Sharapova) vs. Blue (A. Ivanovic) Number of unforced error – 23 : 30	M. Sharapova Won
M9	Green (R. Nadal) vs. Black (R. Federer) Number of unforced error – 9 : 23	R. Nadal Won
M10	Red (A. Ivanovic) vs. White (D. Safina) Number of unforced error – 16 : 27	A. Ivanovic Won

- Limitation of the framework

Recently, the proposed spatiotemporal modeling and matching methods can only be applied to two opposite objects. More objects with more complex interaction would be described in more sophisticated approaches. Even so, we demonstrate that the proposed method is effective in detecting tennis tactics.

In real implementation, the major shortage lies on manual selection of players' uniforms. It seems easy to detect players by finding the moving objects on screen. However, size of the player at upper part of the court is small, and is often mixed with the audience or advertisement boards. Detecting this player solely based on frame difference or block-based motion estimation doesn't work well. In this work, we manually select the initial position of players' uniform and then perform tracking. More elegant approaches that automatically detect players based on statistical information of motion and color will be developed in the future.

- Generality of spatiotemporal modeling

Although we focus on tennis tactics analysis, the proposed spatiotemporal modeling method is not limited to this domain. It should be able to be applied to other domains in which two objects' movements represent targets of interest. For example, badminton and table tennis have very similar court settings to tennis, and how players or ball move reveal the progress of games.

In badminton games, we can detect players as done in tennis, segment the court

into appropriate regions, transform players' movement into symbols, and then model spatial evolution by the proposed method. Players' moving patterns corresponding to important tactics, such as smash and cross net shot, are likely to be modeled.

Note that "object" in the spatiotemporal modeling is not limited to human. If the placement of the pingpong in table tennis can be detected, we can describe the spatial evolution of the pingpong. Which object should be detected depends on domain knowledge of specific domains and what kind of goals are targeted. In Sections 4 and 5, we specially develop modules and extract features for tennis videos, which would be replaced by other modules when other domains of videos are analyzed.

7. Conclusion

We have presented a two-level detection framework to comprehensively detect events and tactics in tennis matches. At the preprocessing stage, court lines and players are detected, and those positions are mapped to a real-world coordinate system. Audio effects are also modeled to facilitate event detection. At the first-level event detection stage, a discriminative learning approach is used to detect events, such as ace and double fault, based on audiovisual features. At the second-level tactics detection, plays that contain rich interaction between players are transformed into symbolic sequences. Frequent moving patterns are mined effectively based on a tree structure, and the probability of a test symbolic sequence corresponding to a specific tactic is evaluated based on the idea of longest common subsequence. Comprehensive experiments were conducted, and the results not only show that the proposed method is promising, but also provide some extensive impacts on semantic video analysis.

In the future, we will extend the proposed approach to modeling the interaction of multiple objects. We also plan to utilize the spatiotemporal modeling to detect events in surveillance videos.

Reference

[1] Sadlier, D.A., and O'connor, N.E. (2005) Event detection in field sports videos using audio-visual features and a support vector machine. IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 10, pp. 1225-1233.

[2] Leonardi, R., Migliorati, P., and Prandini, M. (2004) Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains. IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 5, 634-643.

[3] Xu, H., and Chua, T.-S. (2006) Fusion of AV features and external information sources for event detection in team sports video. ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 2, no. 1, pp.

44-67.

[4] Zhu, G., Huang, Q., Xu, C., Rui, Y., Jiang, S., Gao, W., and Yao, H. (2007) Trajectory based event tactics analysis in broadcast sports video. In Proceedings of ACM Multimedia, pp. 58-67.

[5] Zhu, G., Huang, Q., Xu, C., Xing, L., Gao, W., and Yao, H. (2007) Human behavior analysis for highlight ranking in broadcast racket sports video. IEEE Transactions on Multimedia, vol. 9, no. 6, pp. 1167-1182.

[6] Chu, W.-T., Tien, M.-C., Wang, Y.-T., Chou, C.-W., Hsieh, K.-Y., and Wu, J.-L. (2007) Event detection in tennis matches based on real-world audiovisual cues. In Proceedings of the 20th Computer Vision, Graphics, and Image Processing Conference, pp. 541-548.

[7] United States Tennis Association. (1996) Tennis Tactics – Winning Patterns of Play. Human Kinetics Publishers.

[8] Hanjalic, A. (2002) Shot-boundary detection: unraveled and resolved? IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 2, 90-105.

[9] Liu, Y., Jiang, S., Ye, Q., Gao, W., and Huang Q. (2005) Playfield detection using adaptive GMM and its applications. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 421-424.

[10] Farin, D., Krabbe, S., de With, P.H.N., and Effelsberg, W. (2004) Robust camera calibration for sport videos using court models. In Proceedings of SPIE Storage and Retrieval Methods and Applications for Multimedia, vol. 5307, 80-91.

[11] Hartley, R., and Zisserman, A. (2000) *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[12] Cai, R., Lu, L., Zhang, H.-J., and Cai, L.H. (2003) Highlight sound effects detection in audio stream. In Proceedings of IEEE International Conference on Multimedia & Expo, vol. 3, 37-40.

[13] Cheng, W.-H., Chu, W.-T., and Wu, J.-L. (2003) Semantic context detection based on hierarchical audio models. In Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval, 109-115.

[14] Wang, Y., Liu, Z., and Huang, J.C. (2000) Multimedia content analysis using both audio and visual cues. IEEE Signal Processing Magazine, vol. 17, no. 6, 12-36.

[15] Snoek, C.G.M., Worring, M., Geusebroek, J.-M., Koelma, D.C., Seinstra, F.J., and Smeulders, A.W.M. (2006) The semantic pathfinder: using an authoring metaphor for generic multimedia indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, 1678-1689.

[16] Chang, C.-C., and Lin, C.-J. (2001) LIBSVM: a library for support vector

machine. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] Pallavi, V., Mukherjee, J., Majumdar, A.K., and Sural, S. (2008) Graph-based multiplayer detection and tracking in broadcast soccer videos. IEEE Transactions on Multimedia, vol. 10, no. 5, pp. 794-805.

[18] Zhu, G., Xu, C., Huang, Q., and Gao, W. (2006) Automatic multi-player detection and tracking in broadcast sports video using support vector machine and particle filter. In Proceedings of IEEE International Conference on Multimedia & Expo, pp. 1629-1632.

[19] Zhu, G., Xu, C., Huang, Q., and Liu, H. (2008) Event tactic analysis based on player and ball trajectory in broadcast video. In Proceedings of ACM International Conference on Image and Video Retrieval, 515-523.

[20] Duan, L.-Y., Xu, M., Tian, Q., Xu, C.-S., and Jin, J.S. (2005) A unified framework for semantic shot classification in sports video. IEEE Transactions on Multimedia, vol. 7, no. 6, pp. 1066-1083.

[21] Tagagi, S., Hattori, S., Yokoyama, K., Kodate, A., and Tominaga, H. (2003) Sports video categorizing method using camera motion parameters. In Proceedings of IEEE International Conference on Multimedia & Expo, vol. 2, pp. 461-464.

[22] Ekin, A., Tekalp, A.M., and Mehrota, R. (2003) Automatic soccer video analysis and summarization. IEEE Transactions on Image Processing, vol. 12, no. 7, pp. 796-807.

[23] Yu, X., Xu, C., Leong, H.W., Tian, Q., Tang, Q., and Wan, K.W. (2003) Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In Proceedings of ACM Multimedia, pp. 11-20.

[24] Xie, L., Xu, P., Chang, S.-F., Divakaran, A., and Sun, H. (2004) Structure analysis of soccer video with domain knowledge and hidden Markov models. Pattern Recognition Letters, vol. 26, no. 7, pp. 767-775.

[25] Rui, Y., Gupta, A., and Acero, A. (2000) Automatically extracting highlights for TV baseball programs. In Proceedings of ACM Multimedia, pp. 105-115.

[26] Han, M., Hua, W., Xu, W., and Gong, Y. (2002) An integrated baseball digest system using maximum entropy method. In Proceedings of ACM Multimedia, pp. 347-350.

[27] Zhang, D., and Chang, S.-F. (2002) Event detection in baseball video using superimposed caption recognition. In Proceedings of ACM Multimedia, pp. 315-318.

[28] Chu, W.-T., and Wu, J.-L. (2008) Explicit semantic events detection and development of realistic applications for broadcasting baseball videos.

Multimedia Tools and Applications, vol. 38, no. 1, pp. 27-50.

[29] Kolonias, I, Christmas, W., and Kittler, J. (2004) Automatic evolution tracking for tennis matches using an HMM-based architecture. In Proceedings of IEEE Workshop on Machine Learning for Signal Processing, pp. 615-624.

[30] Rea, N., Dahyot, R., and Kokaram, A. (2005) Classification and representation of semantic content in broadcast tennis videos. In Proceedings of IEEE International Conference on Image Processing, vol. 3, pp. 1204-1207.

[31] Han, J., Farin, D., and de With, P.H.N. (2006) Multi-level analysis of sports video sequences. In Proceedings of SPIE Conference on Multimedia Content Analysis, Management, and Retrieval.

[32] Kijak, E., Gravier, G., Oisel, L., and Gros, P. (2006) Audiovisual integration for tennis broadcast structuring. Multimedia Tools and Applications, vol. 30, pp. 289-311.

[33] Xu, M., Duan, L.-Y., Xu, C.-S., and Tian, Q. (2003) A fusion scheme of visual and auditory modalities for event detection in sports video. In Proceedings IEEE International Conference on Acoustics, Speech, & Signal Processing, vol. 3, pp. 189-192.

[34] Duan, L.-Y., Xu, M., Chua, T.-S., Tian, Q., and Xu, C.-S. (2003) A mid-level representation framework for semantic video analysis. In Proceedings of ACM Multimedia, pp. 33-44.

[35] Chang, S.-F., Chen, W., Meng, H.J., Sundaram, H., and Zong, D. (1998) A fully automated content-based video search engine supporting spatiotemporal queries. IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, no. 5, pp. 602-615.

[36] Choi, J., Joen, W.J., and Lee, S.-C. (2008) Spatio-temporal pyramid matching for sports videos. In Proceedings of ACM International Conference on Multimedia Information Retrieval, pp. 291-297.

[37] Galmar, E., and Huet, B. (2008) Spatiotemporal modeling and matching of video shots. In Proceedings of 1st ICIP workshop on Multimedia Information Retrieval: New Trends and Challenges.

[38] Song, X., and Fan, G. (2007) Selecting salient frames for spatiotemporal video modeling and segmentation. IEEE Transactions on Image Processing, vol. 16, no. 2, pp. 3035-3046.

[39] Lan, D.-J., Ma, Y.-F., Ma, W.-Y., and Zhang, H.-J. (2004) Spatio-temporal pattern mining in sports video. In Proceedings of Pacific-Rim Conference on Multimedia, pp. 306-313.

[40] Pantic, M., Patras, I., and Valstar, M.F. (2005) Learning spatio-temporal models of facial expressions. In Proceedings of International Conference on Measuring

Behaviour.

[41] Huang, Y.-P., Chiou, C.-L., and Sandnes, F.E. (2009) An intelligent strategy for the automatic detection of highlights in tennis video recordings. Expert Systems with Applications, vol. 36, pp. 9907-9918.

[42] Wang, J.R., and Parameswaran, N. (2005) Analyzing tennis tactics from broadcasting tennis video clips. In Proceedings of International Multimedia Modelling Conference, pp. 102-106.

[43] Wang, P., Cui, R., and Yang, S.-Q. (2004) A tennis video indexing approach through pattern discovery in interactive process. In Proceedings of Pacific-Rim Conference on Multimedia, pp. 49-56.

[44] Sudhir, G., Lee, J.C., and Jain, A.K. (1998) Automatic classification of tennis videos for high-level content-based retrieval. In Proceedings of International Workshop on Content-Based Access of Image and Video Databases, pp. 81-90.