Tennis Patterns: Player, Match and Beyond

Jonathan K. Vis [a] Walter A. Kosters [a] Antonio Terroba [b]

[a] *Leiden Institute of Advanced Computer Science, Universiteit Leiden, The Netherlands*
[b] *Telefónica I+D, Madrid, Spain*

Abstract

This paper proposes to apply the framework of pattern mining to the game of tennis. The method builds upon similarity measures for subrallies from a tennis match. Several mining problems are examined, where frequent subrallies (that have a high similarity to other subrallies from the same match) must be found. One can consider individual players or matches (involving rotation), and also general patterns for tennis. Experiments show some of the merits of the method.

1 Introduction

In the game of tennis two players compete, trying to defeat one another. Is it possible to discover interesting patterns from professional tennis matches? There are many ways in which this question can be addressed. In this paper we focus on the discovery of (partial) rallies that are characteristic for a player or for a match, or even for the game of tennis in general. In these rallies (unless they consist of one event) both players participate, so they might be seen as combined efforts. In another paper we discuss winning rallies [9], based on the so-called unbalancing event attribute, and also taking into account the opponent player's position.

The objective of the current paper is to find frequent patterns in tennis at different levels. We propose a method that is based on mining structured data, taking into consideration a comprehensive set of influencing variables captured during a tennis match. Human annotation yields a series of events, ordered in time, where each event consists of precisely described attribute values (not containing free text, etc.). We then try to find frequent patterns in tennis matches, and analyze these patterns in different ways. A game like tennis is quite complicated; the current paper must be seen as a modest attempt to approach the pattern mining problem in this field. Note that we do not give a statistical analysis, but rather generate hypotheses — as data mining often does.

The paper is organized as follows. Section 2 contains related work. In Section 3 we formalize a tennis match and present definitions for similarity measures and thresholds, as well as the mining problems to consider. We present experiments in Section 4 and the conclusions to the study in Section 5.

2 Related Work

The study of captured data from tennis matches in order to find patterns and relationships between variables [8, 2], so that tactical and strategic knowledge can be extracted, is relatively novel.

Wang et al. [11] treat the subject in a similar way, but they only consider relative player movements and no other variables are considered. Wang and Parameswaran [10] take into account 58 possible patterns and try to find them in the footage using Bayesian networks. Zhu et al. [13] propose a tactic representation based on temporal-spatial interactions in soccer. Lames [5] focuses on relative phases of lateral displacements, but it neither includes longitudinal movements nor represents the reality of a professional tennis match where players do not move back to the center of the court every time.

Schroeder et al. [7] use a framework based on short term and long term memory that allows an incremental processing of data streams. However, the tennis model used only includes one variable (the ball landing position) and only eight different locations. Chu and Tsai [1] use symbolic sequences to tackle tactics analysis. They use players location (four areas), players movement direction (up, down, left, right, still)

and players speed (fast, medium, still) to find frequent movement patterns. Jager et al. [4] analyze the players movements in volleyball using self-organizing artificial neural networks that allow them to cluster and visualize configurations and trajectories.

We incorporate more features, especially some that appear to be of great importance for gameplay. We do not focus exclusively on player movement or ball movement, but we aim at their intertwining. Furthermore, tactics are not used as input, but patterns are generated as output.

3 Formalization

In this section we explain how we formalize a tennis match between two players, 1 and 2. For the rules of tennis, the reader is referred to [3, 12].

3.1 Events and Rallies

For a tennis match, quite a lot of data is available. We will restrict ourselves to sequences of events (called (partial) rallies). A *rally* refers to a sequence that begins with a serve, and ends when the point is decided; in a sequence the two players alternate. A *partial rally* or *subrally* is a consecutive subsequence of a rally.

An event is a 5-tuple (pl, st, P_1, P_2, sb) describing a single stroke, its attributes being:

- pl: player hitting the ball, possible values: $\{1, 2\}$;
- st: stroke type, $\{FS, SS, FH, FHS, BH, BHS, VOL, SM, LOB, DSH\}$, corresponding to first serve, second serve, forehand, forehand sliced, backhand, backhand sliced, volley, smash, lob and drop shot, respectively;
- $P_1 = (x_1, y_1)$: position of the player when the ball is hit, C;
- $P_2 = (x_2, y_2)$: position of the ball when it bounces on the opponent's half of the court, C;
- sb: speed of the ball generated after the stroke, $\{slow, normal, fast\}$.

Here the set C is defined as $C = \{0, 1, \ldots, 316\} \times \{0, 1, \ldots, 768\}$, referring to points from the 2-dimensional tennis court (see Figure 1). Because the players change sides every couple of games, a transformation in the coordinates is needed so that the data is always coherent. In [9] more information on this (for instance, what happens if the ball does not bounce on the opponent's half of the court?), and on the way these attributes are gathered from footage of a match, is provided.

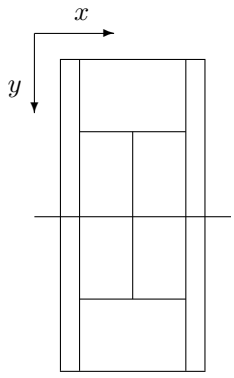

Figure 1: Tennis court reference model.

3.2 Similarity Measures

In this section we propose several similarity measures for events and partial rallies. We deviate from the approach taken in [9], where the unbalancing stroke attribute plays a crucial role in the pattern mining.

First, we define a similarity measure sim between individual events. In this case, when we have events $e = (pl, st, P_1, P_2, sb)$ and $e' = (pl', st', P_1', P_2', sb')$, a first proposal for a similarity measure between these events is (all functions used will be explained in the sequel):

$$\begin{aligned} sim(e, e') = \; & simplayer(P_1, P_1') + simball(P_2, P_2') \\ & + \delta(pl, pl') + simstroke(st, st') + \delta(sb, sb') \end{aligned} \tag{1}$$

where each separate function determines the similarity between the corresponding attributes. With $dist(P, Q)$ representing the Euclidean distance between points P and Q (in pixel width; one pixel width corresponds to (1/6)-th of a foot, so roughly 5 cm), we define:

$$simplayer(P, Q) = \begin{cases} 1 & \text{if } dist(P, Q) < 20 \\ 0.7 & \text{if } 20 \leq dist(P, Q) < 30 \\ 0.5 & \text{if } 30 \leq dist(P, Q) < 40 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$simball(P, Q) = \begin{cases} 1 & \text{if } dist(P, Q) < 10 \\ 0.7 & \text{if } 10 \leq dist(P, Q) < 20 \\ 0.5 & \text{if } 20 \leq dist(P, Q) < 30 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$simstroke(st, st') = \delta(st, st') + \epsilon(st, st') \tag{4}$$

$$\delta(u, v) = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The function ϵ allows for additional weight in the case of near equal stroke types, and is currently defined as

$$\epsilon(st, st') = \begin{cases} 0.7 & \text{if } \{st, st'\} \text{ equals } \{FHS, FH\} \text{ or } \{BHS, BH\} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

So to speak, the "ball similarity" is more sensitive than the "player similarity": for a high similarity, the ball bouncing positions must be more alike. All of the individual terms can get their own weight, if necessary. Note that $0 \leq sim(e, e') \leq 5.0$; often, a linear scaling is used in order to obtain values between 0 and 1. And of course, the step functions used could be smoothened in appropriate ways, e.g., applying a continuous function. The current choice seems to perform better in the experiments done so far, providing more stable results. A possible explanation might be that it compensates for inaccuracies due to the human annotation of the matches.

Now that we have defined the similarity between events, we can easily determine $sim(seq, seq')$, the similarity between same-length sequences (or partial rallies) seq and seq' of single events. First, if the sequences are of unequal length, we define the similarity to be 0. If the length of both sequences equals n and $seq = \langle e_1, \ldots, e_n \rangle$ and $seq' = \langle e_1', \ldots, e_n' \rangle$, then:

$$sim(seq, seq') = \sum_{i=1}^{n} sim(e_i, e_i') \tag{7}$$

In order to obtain a high value for sim, it is quite natural to suppose that the two sequences are aligned in such a way that they both begin with the same player (and of course, all subsequent event pairs are then also with the same player). The current definition does not enforce this, but will most likely punish misaligned combinations with a very low similarity. It would also have been possible to force the condition $pl \neq pl'$ into equation (1).

In order to prepare for a possible "exchange of players", we define $R_{180}((x, y)) = (316 - x, 768 - y)$ for $(x, y) \in C$, a $180°$ rotation around the center of the court; we further put

$$R_{180}((pl, st, P_1, P_2, sb)) = (3 - pl, st, R_{180}(P_1), R_{180}(P_2), sb) \tag{8}$$

and extend this definition to sequences in the natural way.

3.3 Similarity Thresholds and Support

Once we know the similarity value between events $sim(e, e')$ and sequences $sim(seq, seq')$, we need to establish the criteria by which we will consider two events or two sequences as similar in order to proceed

with the pattern mining. We will use the thresholds $event_{\text{thr}}$ and $series_{\text{thr}}$ for this matter. Note that we are defining several thresholds to allow for greater flexibility. This way, two events e and e' will be considered similar if and only if $sim(e, e') \geq event_{\text{thr}}$. Likewise, two sequences seq and seq' of length n will be considered *similar* if and only if

$$sim(seq, seq') \geq n \times series_{\text{thr}} \tag{9}$$

where we also demand all concurrent event pairs from the sequences to be similar, i.e.,

$$sim(e_i, e'_i) \geq event_{\text{thr}} \quad i = 1, 2, \ldots, n \tag{10}$$

with $seq = \langle e_1, \ldots, e_n \rangle$ and $seq' = \langle e'_1, \ldots, e'_n \rangle$. Sequences of unequal length are considered dissimilar.

The concept of a series threshold makes it possible to find long similar sequences which might present higher event similarity at some points than others, and that a simple event threshold filter would block. This is of particular interest if inequality (10) is in some way weakened, or even skipped for some values of i. We leave this as a suggestion for further research.

Given these thresholds, we can now define the so-called *support* in a match M for a sequence seq:

$$support_M(seq) = \sum_{\substack{seq' \text{ in } M \\ seq' \text{ similar to } seq}} sim(seq, seq') \tag{11}$$

Being similar requires the rallies to satisfy inequalities (9) and (10). Note that it is not necessary that seq is a subrally from M itself, it can even be any synthetic sequence.

If we want to find common patterns for both players, as opposed to the normal support, the rotated sequences must also be taken into account. In that case the most similar sequence, i.e., the one having the maximum similarity, contributes to the support computation. So we define:

$$support\_rot_M(seq) = \sum_{\substack{seq' \text{ in } M \\ seq' \text{ similar to } seq \text{ or } R_{180}(seq)}} \max(sim(seq, seq'), sim(R_{180}(seq), seq')) \tag{12}$$

3.4 Mining Problems

We are now able to define our two mining problems. Given a match between two players, we want to determine the *frequent* subrallies, i.e., those subrallies for which there are many similar subrallies or rather for which the support is high. If we concentrate on the player identity, we just use $support_M$ (so similar subrallies are necessarily from the same player), and we are looking for so-called *personal frequent subrallies*; if we are more interested in general patterns, we use $support\_rot_M$, also incorporating rotated sequences, and we then get so-called *anonymous frequent subrallies*. In the second case, for the *support* computation, we both consider the subrally itself and its rotated version. Note that in this case we still get patterns that are (or rather, might be) connected to the match under consideration, and are dependent on the two players. We will return to this issue in Section 4.

More precisely, we define:

MINING PROBLEM 1 — PERSONAL FREQUENT SUBRALLIES

Given a minimum support threshold $min\_support$, an event threshold $event_{\text{thr}}$ and a series threshold $series_{\text{thr}}$, determine those partial rallies seq of length n in the match M for which $support_M(seq) \geq min\_support$.

---

and

MINING PROBLEM 2 — ANONYMOUS FREQUENT SUBRALLIES

Given a minimum support threshold $min\_support$, an event threshold $event_{\text{thr}}$ and a series threshold $series_{\text{thr}}$, determine those partial rallies seq of length n in the match M for which $support\_rot_M(seq) \geq min\_support$.

Given these two problems, it seems that there are two supports that are of interest for any given sequence. However, in fact there are three: for any event sequence seq we are interested in $support_M(seq)$,

$support_M(R_{180}(seq))$ and $support\_rot_M(seq)$. But since — except for rare cases, if both a subrally and its rotated version are similar to the given one —

$$support\_rot_M(seq) = support_M(seq) + support_M(R_{180}(seq)) \tag{13}$$

we will indeed restrict our attention to two of these.

It would also be possible to take into account the outcome and the relative order of the rallies, as well as the score within the match. And of course, additional attributes to those mentioned in Section 3.1 could be included in the analysis.

As we will see in Section 4, the tennis matches we consider have in the order of magnitude of 1,000 events. If we restrict ourselves for the moment to subrallies of length $n = 3$ (the "3-rallies"), there will be at most approximately 1,000 of these. The mining problems defined here can then be solved without too much computing effort. It is easily possible to use a brute-force algorithm. Therefore, we will not further elaborate on this; indeed, a straightforward computation was used to generate the results in Section 4. However, in some situations it might be useful to rely on more powerful (i.e., efficient) algorithms for frequent itemset mining, like the one from [6].

4 Experiments

In this section we will describe several experiments, highlighting some examples. We are interested in patterns for players and for matches, and finally we will examine more general patterns. But first we describe our datasets.

4.1 Datasets

Over 3,000 events with all their attributes (player hitting the ball, stroke type, player position, ball bouncing position and speed of the ball), and more than seven hours of recordings where captured and analyzed, covering men's and women's matches in both hard courts and clay courts. Table 1 shows information on the three matches analyzed.

Match	Tournament	Remarks	Some statistics	Players	Result
I	Australian Open 2010, Women, Hard	SemiFinal, 122 minutes, 831 events, 181 rallies	424 3-rallies, 117 start with serve, average rally length 4.2, standard deviation 3.1	S. Williams vs Na Li	7–6, 7–6
II	Roland Garros 2009, Men, Clay	Round 16, 210 minutes, 1,628 events, 271 rallies	1,024 3-rallies, 221 start with serve, average rally length 5.7, standard deviation 4.0	Söderling vs Nadal	6–2, 6–7, 6–4, 7–6
III	ATP MS Paris 2007, Men, Hard	Final, 70 minutes, 518 events, 88 rallies	310 3-rallies, 66 start with serve, average rally length 5.4, standard deviation 4.1	Nalbandian vs Nadal	6–4, 6–0

Table 1: Datasets: three tennis matches.

The longest rallies have length 18, 25 and 17, respectively. Note that match II is much larger than match I (that has the shortest rallies), which is in turn larger than match III. The different surfaces of the courts provide an explanation for this.

4.2 Patterns for Players

In our first experiment we addressed MINING PROBLEM 1 for the three matches. The thresholds were fixed at 3.0 for $event_{\mathrm{thr}}$ and 10.0/3 for $series_{\mathrm{thr}}$ (the similarity between two 3-rallies is at most $3 \times 5.0 = 15.0$; and the whole sequence satisfies a stricter threshold condition than its three individual events combined), while $min\_support$ was in each case chosen (at 197.3, 311.0 and 71.0, respectively, largely determined by

the size of the matches) such that precisely 20 patterns of length 3 were generated for each match — in other words, the 20 most frequent patterns were selected in each case. These subrallies are referred to as *interesting patterns*. Note that the nature of the patterns found can be influenced by the $min\_support$ parameter, but we expect the interesting patterns to be more or less stable among different matches.

It turned out that for match II the interesting patterns could be divided into 5 groups, each group consisting of patterns that get most of their support from the others within the same group. The two leftmost partial rallies from Figure 2 show two interesting patterns from the same group for this match. For match I we distinguished 2 or 4 groups (depending on the level where splits are made), and for match III 4 groups. Also worth mentioning is that some of the original (full) rallies gave rise to several interesting patterns, in particular for the smaller matches I and III. Some patterns even arise from overlapping subrallies; this might indicate the occurrence of a longer pattern, or a complicated intertwining of shorter ones.

The three rightmost patterns from Figure 2 show clearly that the player is returning the serve in exactly the same direction as the incoming ball. Changing the ball direction is difficult and riskier, so this shows a tactical pattern if repeated many times.

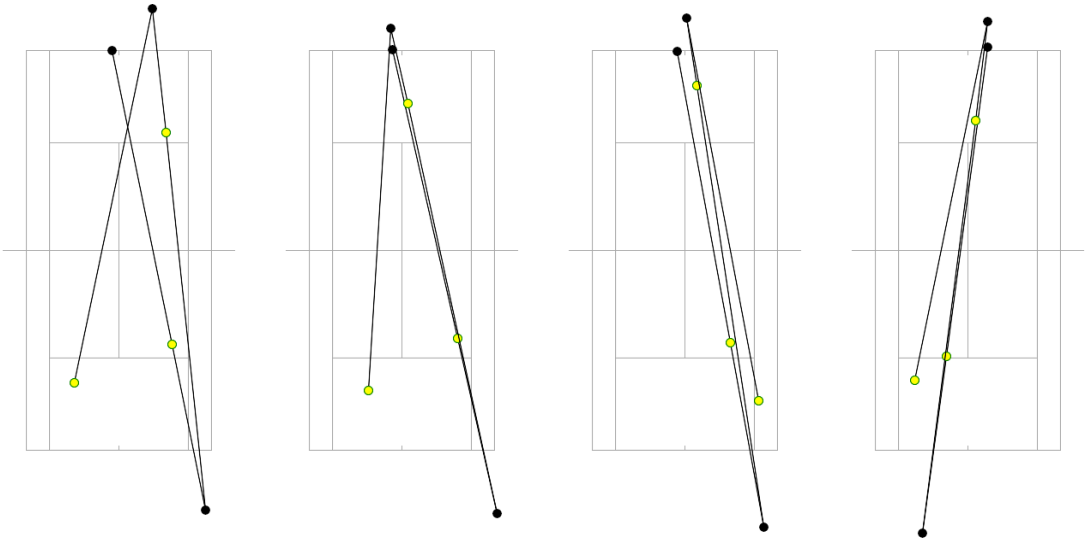

Figure 2: Four subrallies from match II. Black dots denote player positions. All these partial rallies start with a serve. The two leftmost subrallies are very similar, and so are the middle two. The two rightmost subrallies are somewhat similar to the leftmost one, but are not similar to one another.

4.3 Patterns for Matches

We broadened the experiment by also addressing MINING PROBLEM 2 for the three matches. Only adjusting $min\_support$ (to 371.0, 371.0 and 91.0, respectively) we again generated 20 patterns for each match, again referred to as *interesting*.

For match I, and for match III too, 16 out of these patterns were also present in the list for MINING PROBLEM 1, while for match II 12 patterns were already found earlier. The two leftmost patterns in Figure 2 show two of these, the two rightmost ones are "new" interesting patterns that belong to the same group as that mentioned in the previous section. These are interesting if one allows for rotation, i.e., *both* players have subrallies similar to these. These subrallies might be candidates for being more general patterns, but might also be just exemplary for this match. Some of the other subrallies were not interesting anymore: they did not meet the higher support threshold, and are clearly too player specific.

The relative number of frequent patterns that start with a serve is much higher in match II than in the other two matches. In fact, 4 out of the 20 interesting patterns for match II start with a serve (making use of the rotated support), versus none of the interesting patterns for matches I and III.

In match I application of the rotation substantially increases the support (justifying the high $min\_support$ needed for this match), contrary to the situation for the other two matches. Apparently, the two players in

match I do act more like one another than the players in matches II and III.

4.4 General Patterns

As a last experiment we examined MINING PROBLEM 2 for the union of the rallies from all three matches. Note that MINING PROBLEM 1 makes less sense in this situation. We generated the 30 subrallies with the highest support. From these, 13 patterns also occur as interesting pattern for the individual matches. None of the 30 subrallies originated from match III (although rallies from this match do provide support for interesting patterns).

The interesting patterns could be divided into two main groups; the two leftmost patterns in Figure 3, taken from match I and match II, respectively, are much alike, and so are the two rightmost subrallies (after rotation), both taken from match I. It is possible to further subdivide the two groups into smaller components.

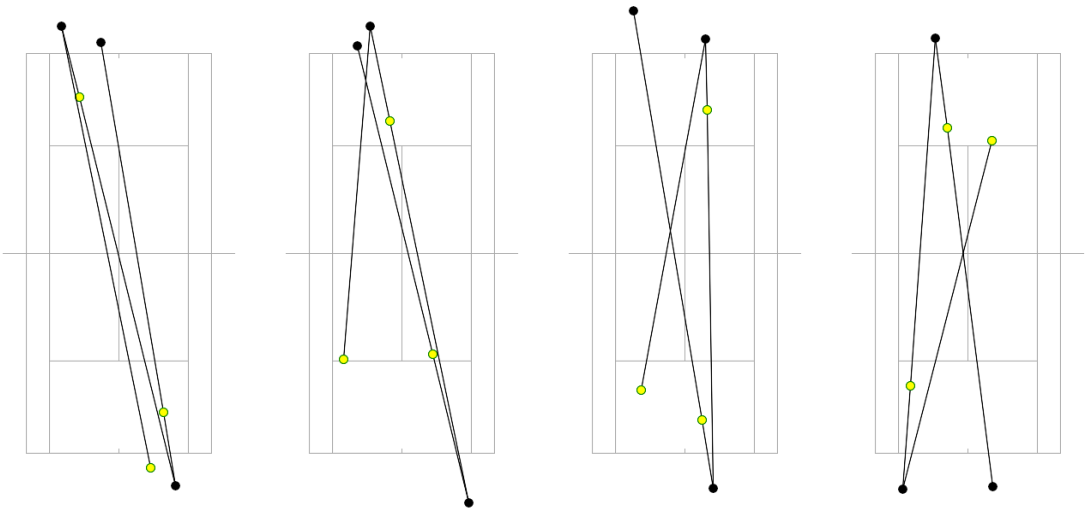

Figure 3: Four interesting subrallies from the three matches seen as a whole. The two leftmost subrallies are very similar, and so are the rightmost two (after rotation).

Although the first pattern from Figure 2 and the second from Figure 3 somewhat look like one another when inspected visually, their similarity is not that high since those from Figure 2 all start with a serve. Also, the third pattern from Figure 2 and the first from Figure 3 show some visual resemblance.

Figure 3 does not really provide much tactical insight. Patterns like these can be observed many times; there is not much about them to indicate a strategic intention. Indeed, in tennis it often occurs that players exchange strokes during a longer period of time without directly attacking one another, see also [9].

5 Conclusions and Further Research

In this paper we apply the framework of pattern mining to the game of tennis. We define similarity measures, thresholds, support, and the corresponding mining problems. Experiments show that the approach works in practice. Some tactical patterns do show up. However, tennis is a complicated game to analyze, and the results are not at all easy to interpret.

Further research includes the search for more general patterns by constructing so-called pseudo-rallies (merging or generalizing several frequent subrallies) and proper (statistical) analysis of more datasets. For specific players strengths and weaknesses could be explored, and patterns for matches and more general ones should be further examined. Finally, the general structure of the (groups of) interesting patterns needs further analysis.

References

[1] W.-T. Chu and W.-H. Tsai. Modeling spatiotemporal relationships between moving objects for event tactics analysis in tennis videos. *Multimedia Tools and Applications*, 2009.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2008.

[3] International Tennis Federation (ITF), 2010. `http://www.itftennis.com/`, retrieved May 17, 2010.

[4] J.M. Jager, J. Perl, and W.I. Schollhorn. Analysis of players' configurations by means of artificial neural networks. *International Journal of Performance Analysis in Sport*, 7:90–103, 2007.

[5] M. Lames. Modelling the interaction in game sports — Relative phase and moving correlations. *Journal of Sports Science and Medicine*, 5:556–560, 2006.

[6] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *17th Int. Conf. on Data Engineering (ICDE)*, pages 215–224, 2001.

[7] B. Schroeder, F. Hansen, and C. Schommer. A methodology for pattern discovery in tennis rallys using the adaptative framework ANIMA. In *Second International Workshop on Knowledge Discovery from Data Streams (IWKDDS)*, 2005.

[8] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[9] A. Terroba, W.A. Kosters, and J.K. Vis. Tactical analysis modeling through data mining, Pattern discovery in racket sports. In *International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010)*, 2010.

[10] J.R. Wang and N. Parameswaran. Analyzing tennis tactics from broadcast tennis video clips. In *11th Int. Multimedia Modelling Conf.*, pages 102–106, 2005.

[11] P. Wang, R. Cai, and S.-Q. Yang. A tennis video indexing approach through pattern discovery in interactive process. In *Advances in Multimedia Information Processing (PCM)*, pages 49–56, 2005. LNCS 3331.

[12] Wikipedia — Tennis, 2010. `http://en.wikipedia.org/wiki/Tennis`, retrieved May 17, 2010.

[13] G. Zhu, Q. Huang, C. Xu, Y. Yui, S. Jiang, W. Gao, and H. Yao. Trajectory based event tactics analysis in broadcast sports video. In *15th Int. Conf. on Multimedia*, pages 58–67, 2007.