TennisMatchViz: A Tennis Match Visualization System

Xi He and Ying Zhu
Department of Computer Science
Georgia State University
Atlanta - 30303, USA
Email: xhe8@student.gsu.edu, yzhu@gsu.edu

Abstract

Sports data visualization can be a useful tool for analyzing or presenting sports data. In this paper, we present a new technique for visualizing tennis match data. It is designed as a supplement to online live streaming or live blogging of tennis matches. It can retrieve data directly from a tennis match live blogging web site and display 2D interactive view of match statistics. Therefore, it can be easily integrated with the current live blogging platforms used by many news organizations. The visualization addresses the limitations of the current live coverage of tennis matches by providing a quick overview and also a great amount of details on demand. The visualization is designed primarily for general public, but serious fans or tennis experts can also use this visualization for analyzing match statistics. We demonstrate this visualization technique with the example of 2015 French Open final match.

Introduction

Data analysis has been an important part of many sports, especially at the elite level. For example, statistical analysis has been long been used in baseball and basketball. Advanced sensing and imaging technologies can track balls and athletes' on-court movements, providing a rich set of data for viewing and analyzing sports games in different ways. Sports data visualization is an emerging field that explores visualization techniques to effectively present sports data. Sports data visualization can be used for different purposes. Some sports data visualizations are designed for athletes, coaches, and experts to conduct performance analysis. Some sports data data visualizations are designed for news media and fans. Because different sports have different rules, sports data are often different from sport to sport. Therefore, sports data visualizations are often domain specific.

In this paper, we present a new technique for visualizing tennis match data. It is designed to supplement live coverage of tennis matches. Currently, important tennis matches are streamed online or broadcasted on TV. Many news media also provide online live blogging on important matches (see [1]). There are two limitations to this kind of live coverage. First, they provide a linear, one dimensional narrative of the match. One knows what is happening now but it is difficult to find out quickly what has happened before. For online streaming or TV broadcasting, it's difficult to go back in time to watch earlier points. In live blogging, readers can go back and read comments on a previous game, but they have to page through many comments in between. One cannot quickly jump to a particularly interesting point. Second, current live coverage is not interactive. Users cannot get a quick answer to questions such as "How many Aces has Roger Federer served?" or "How many backhand winners has Stan Wawrinka hit?"

Our visualization technique addresses these issues by presenting tennis match data in a 2D interactive view. This Web based visualization provides a quick overview of match progress, while allowing users to highlight different technical aspects of the game or read comments by the broadcasting journalists or experts. Its concise form is particularly suitable for mobile devices. The visualization can retrieve data directly from a tennis match live blogging web site. Therefore it does not require extra data feeding mechanism and can be easily integrated with the current live blogging platform used by many news media.

Designed as "visualization for the masses", this visualization is concise and easy to understand and yet can provide a great amount of details on demand. Because it is a 2D view, users can quickly jump to an interesting point without paging or dragging slider bars. Users can interact with the visualization by typing a question and see the technical data being highlighted in the visualization. The visualization is designed primarily for general public, but serious fans or tennis experts can also use this visualization for analyzing match statistics.

The rest of the paper is organized as follows: Section briefly reviews background and related tennis match visualization techniques. Then we overview our system in Section , and describe the functionality and important techniques of its three major components in Section , and . We also introduce implementation details in Section . Lastly we present the case studies in Section and conclude the work in Section .

Background and Related work

We first give a general overview of sports data visualization [2] and then focus on tennis data visualization.

Sports data visualizations can be classified based on multiple parameters: data, visualization techniques, target audience, and domain.

Sports data can be roughly divided into three categories: on-court performance data, game statistics, and off-court statistics. On-court performance data are often collected by sensors or advanced imaging techniques. For example, the Hawk-Eye systems can provide tennis player's location, speed, ball speed, ball trajectory, etc. Special sensors put on racquets can record racquet speed. News media sometimes visualize performance data (e.g. player position and movement) and superimpose them on live images. In general, on-court performance data are not freely available.

Game statistics are compiled by human and often freely available [3]. They include scores, number of errors, number of aces, number of double faults, number of forehand or backhand

winners, etc. Off-court statistics [4] may including player's ranking, age, height, weight, number of titles won, salary, prize money earned, etc.

Visualization techniques

The visualization techniques are generally chosen based on data types. On-court performance data are often displayed as markers, heatmaps [5–7], or trajectory lines superimposed on a court image [8]. Game statistics and off-court statics are displayed in a wide variety of visualizations [9–13].

Target audience

Sports data visualizations often target three types of audiences: experts, serious fans, and general public. For experts, the goal of data visualization is to help their data analysis. Therefore the visualizations tend to display more technical details gathered from on-court performance data, with high data density and interactions. For serious fans, the data visualizations often serve dual purposes. On one hand, they are used for casual analysis. On the other hand, they are treated as a type of visual art. For example, many data visualizations are designed to be posters or infographic [14]. This type of visualizations often feature colorful, complicated, and creative visual forms. For general public, the goal of sports data visualization design is to provide quick overview as well as easy interaction. The visual forms should be simple and clean, with details on demand.

The visualization techniques proposed in this paper is based on game statistics. The visualization features a series of mini-timelines arranged on horizontal bars. The target audiences are general public and serious fans.

Because different sports have different rules and different data properties, it is often difficult to compare visualizations across different sports. In this review, we focus on related work in tennis data visualization. Polk, et al. [15] developed a tennis match data visualization system called TenniVis. Both TenniVis and our visualization focus on presenting game statistics. But there are two main differences. First, the data in TenniVis needed to be collected by human, while the data in our visualization are retrieved automatically from a tennis match live blogging Web site. This means TenniVis is suitable for post-match analysis but not for live coverage. Second, TenniVis is designed for expert analysis. With glyph based graphs, it displays many technical details in a complex visualization design. Our visualization is designed for general public. It features a simpler design of basic match statistics. In addition, our visualization is web based program, while TenniVis is a standalone toolkit.

Saunder's tennis visualizations [8] mostly deal with on-court performance data. With access to Hawk-Eye data, he superimposes data visualizations on a tennis court image in both 2D and 3D. He has also designed a game tree visualization of Nadal's game statistics in 2013. Saunder's main target audiences are experts. But his work is also featured in a documentary film and news media. In the latter case, the data visualizations are used as infographics. Our work is quite different from Saunder's work because we focus on game statistics rather than on-court performance data.

In summary, existing tennis data visualizations focus on post-match analysis for expert users. The main contribution of our work is its focus on automatic data collection and live match data visualization.

Overview

TennisMatchViz is a tennis match visualization system that takes as the input a web page containing live commentary of a tennis match, parses it and presents the tennis match visually to audiences. As shown in Figure 1, TennisMatchViz is designed as a Web application and can be divided into two parts: server-side system and browser-side system. The server-side system establishes a tennis associated knowledge base by collecting information from sports websites. This knowledge base also includes history tennis commentaries which are crucial to the parsing of live tennis commentary. After this setup, the commentary parse is ready to accept a tennis web page, extract match associated information, parse comments, understand the match progress, and repackaging and transferring data objects describing the match to the visualization engine in the browser-side system. The visualization engine is responsible for displaying the tennis match to users. Its interaction with user is achieved with help of the interaction component. The interaction component accepts questions from users, have them parsed by the question parser in the server-side system which then issues commands to instruct the visualization engine to update the visualization accordingly.

The following sections will describe data processing and data visualization in more details.

Data Processing

Before we can process data, we need to obtain them from the sports websites. Nowadays, many sport webpages use AJAX technology to fetch updated data from their back-end servers. The AJAX technology can avoid the refreshing of the whole web page and thus improve user experience, but it also make it more challenging for our program to obtain data with the traditional web crawling technology. In such case, we have to scrape the web page, examine and identify the URL for fetching data.

The basic task for data processing is to transform the streaming tennis data from a sports website to hierarchical data objects, and prepare data for visualization. As every sports website has its unique data format and commentary style, it is rarely possible to develop a general data processing program for all the sport websites. Our goal is to design and develop a customized program to process data from a given website.

Transforming flat data into hierarchically data objects

A live tennis commentary on sports websites usually consists of a set of records. Each record contains three parts: a comment which expresses the opinion of the commentator towards a specific point, a current scoreboard (sets, games, points) and a timestamp. There can be multiple records for one single point. These records are easy for human beings to read and understand, and hard for computers to process. In order to effectively manage, understand and visualize this tennis match information, we need to transform the flat data into hierarchical data objects. Firstly, according to the tennis scoring system, a set of hierarchical data objects (match, set, game, point) are constructed. Then for each record, we parse the score board information it contains, determine which (set, game, point) this record is made for, and populate the data objects with the record. Below we will discuss some

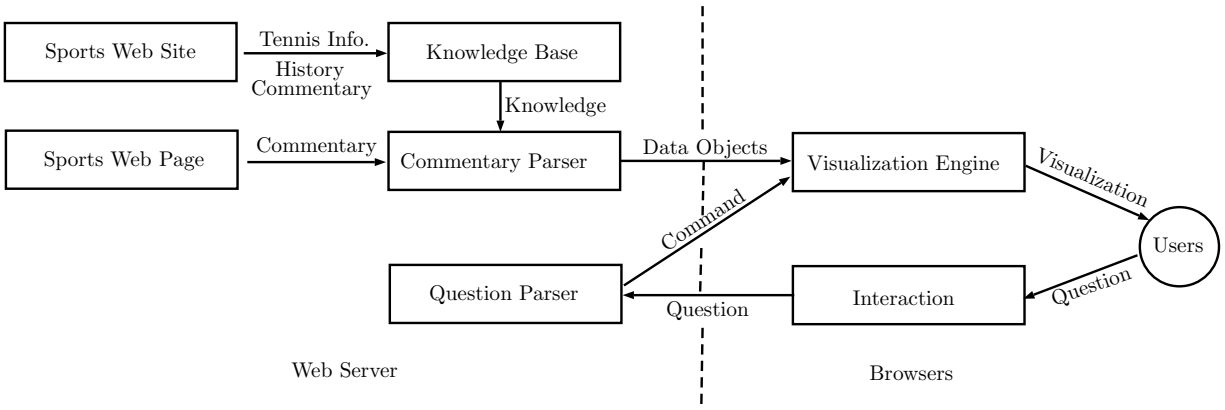

Figure 1: The Architecture of TennisMatchViz

important detail for implementing this transformation.

- Not all the records contain useful information. Some of them contains commercial ads or other pictorial information. We need to preprocess these records, and remove unrelated records.
- The scoreboard changes only when a score is made. For the record that tracks the score, we can easily parse its scoreboard information and process it. But for other records made for the same point, their scoreboard information does not include the score, and is not correct. Our algorithm needs to identify these records, retrieve the comments in these records and store them in the proper data objects.
- Records may have different scoring rules and format for deuce points and points in tiebreak game.
- As records are input impromptu by commentators, typos are possible. The strategy we use to catch these typos and correct them if possible is based on the assumption that the order of records is in accordance with the progress of the match. If there is an abrupt change between the scoreboards of two successive records, it is likely that a typo is in the records. We can fix some of the typos, but lacks enough information for other typos. This is the limitation of our system.

Parsing the tennis comments

The full understanding of arbitrary text is the ongoing research in the field of natural language understanding or computational semantics, and is traditionally considered to be a difficult problem. The current solution replies heavily on human-annotated text and sophisticated machine learning methods, but still can not achieve a satisfactory result. In our case, however, tennis comments text is a small subset of human language in terms of vocabulary and syntactic structure. With some tricks, we can grasp the key meaning out of them and provide data for tennis match visualization.

One important observation on the tennis comments is that many of sentences share the same structures. For example,

```
Murray fires a forehand over the baseline.
Federer fires a forehand over the baseline.
```

The above two sentences describe two different players making the same actions. These sentences are almost the same except that the subjects of the sentences are different. We have seen plenty of similar patterns in the tennis commentary and we believe that some sports websites use templates for their tennis live commentary. Based on this observation, we come up with an algorithm for training and understanding the tennis comments. The basic idea is to identify these patterns in the templates by computing common strings in the history comments and annotate them, and then use these annotated patterns to parse the new comments. The outlines of the algorithm is as follows:

1. Collect a certain number of tennis commentaries from the same sport website. Extract the comments from records in the commentaries.
2. Replace player names with the same string.
3. For each pair of comments, find the longest common string.
4. Count the frequency of longest common strings.
5. Sort these longest common strings according to their frequency.
6. Annotate these common strings with high frequency.
7. For a new comment, parse it by checking if it contains any annotated strings.

For completeness, we briefly discuss the algorithm for computing longest common string of two strings. Let us suppose we have two strings X and Y with a length of m and n. For any suffixes of X and Y, their longest common string has the following optimal substructure property:

$$LCS(i,j) = \begin{cases} 0 & X[i] != Y[j] \\ LCS(i+1,j+1)+1 & X[i] = Y[j] \end{cases}$$

Where i and j are the starting index of the suffixes in X and Y, respectively.

Therefore, we can create a $m*n$ table and use dynamic programming to compute the common string of any pair of suffixes of X and Y, and select the longest one.

Data Visualization

After the data processing step, processed tennis match data will be delivered to the browser and presented visually to audiences. The main goal of our visualization engine is to develop a general front-end framework for tennis match data visualization.

Figure 2: The Original Web Page of the Tennis Match Between Novak Djokovic and Stan Wawrinka (in Chinese)

We primarily classify the tennis match data into two categories: basic data and technical detail data. Basic data include the number of sets in the match, the number of games in each set, the scores in each games and the duration for each game and each set. Technical detail data contain information describing the tennis match from different perspectives. The following listing shows a subset of technical detail data that audiences are interested in.

- ACE, double fault
- Backhand, forehand
- Approach shot
- Volley, half volley
- Foot fault
- Inside out, inside in
- Lob

Basic data are visualized with a "non-traditional" table with one row denoting a set of tennis games. Each row is equipped with a Cartesian coordinate system, whose horizontal and vertical axis stand for the time of the games and the scores, respectively. A chart consisting of circles and lines is used to represent a game. Each player and his associated data are assigned a color. The colored circles in the chart represent the scores of players at every point. When the pointer hover over these circles, the associated comments will be popped up and displayed to audiences. Circles with same color are connected to form lines to show the trend of the game. These lines can provide a straightforward view to audiences on how the games are carried out. For example, in close games these lines will alternately grow. Charts are assigned background colors which are similar to that of players who win the games. The width of the chart is proportional to the duration of the game, and offers another aspect of game information to audiences.

Interaction

In comparison, technical detail data is harder to be visualized and shown to audiences. First of all, we can not show all the technical detail data to audiences in one visualization. They contain too much information and can easily confuse audiences. Secondly, in many cases, the audiences do not wish to understand the whole data set. They may want to filter these data, and retrieve what interest them most. For example, they may just want to see the double faults that one player performs in a certain set. Thirdly, it is more meaningful that the selected technical detail data can be visualized alone with the basic data.

We come up with the concept of "visualization on demand". The idea is similar to that of Question-Answering in the field of Natural Language Processing. Nowadays, many famous products like Apple Siri, Google Now, Microsoft Cortana are based on Question-Answering technology, and are able to communicate with human being in a more user-friendly way. They take a user question as input, parse it, carry out the semantic search based on the parsed query and output results to users. Visualization on demand operates in the similar way, but instead of conducing semantic search, it produces a visualization that meets users' need.

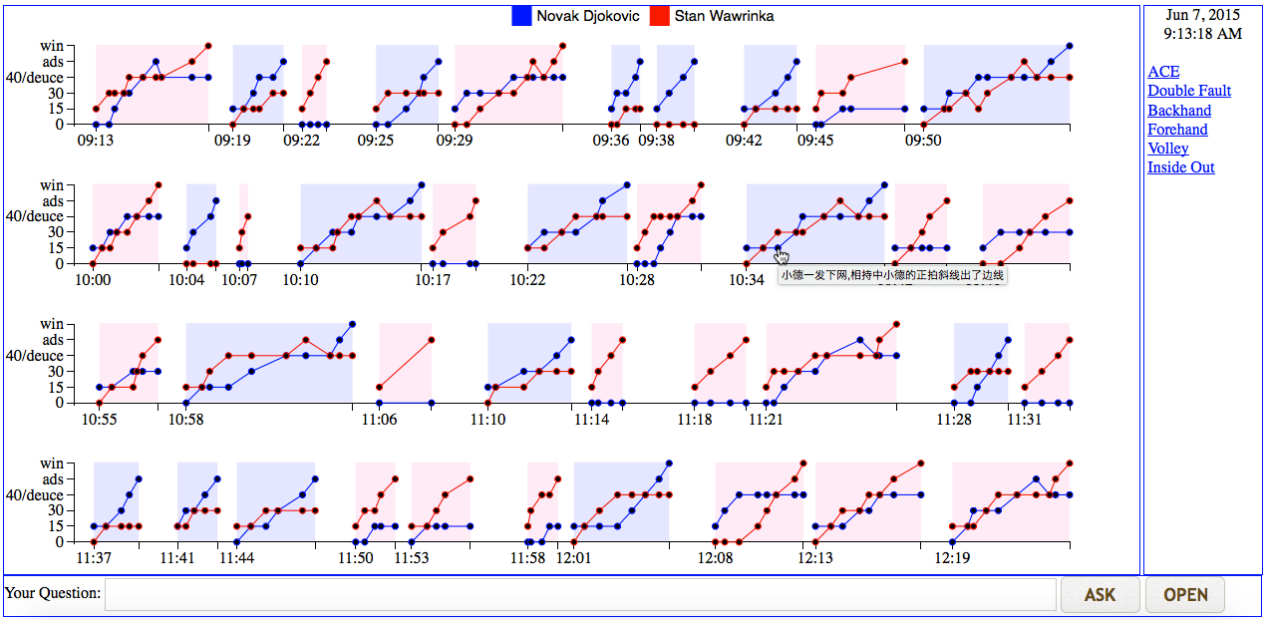

Figure 3: The TennisMatchViz Visualization for the Tennis Match Between Novak Djokovic and Stan Wawrinka

The visualization on demand is implemented with the collaboration of the visualization engine, the interaction component and question parser. The interaction component can accept questions from users. These questions are parsed, and commands are produced and issued to the visualization engine. The visualization engine reads the commands and selects the proper technical detail data to update the visualization.

Implementation

TennisMatchViz is based on the browser-server architecture and can be divided into a sever-side system and a browser-side system. The server-side system is implemented with Java Servlet and Google Json library and is running in the Tomcat Web server. The browser-side system is implemented with Javascript and Javascript library such as D3.js, JQuery and JQuery UI.

Case Study

To demonstrate the use of our system, we will visualize the man single finals between Novak Djokovic and Stan Wawrinka in ATP France Open 2015. The live commentary is from a famous Chinese sports website sports.sina.com.cn. One reason for collecting data from Sina Sports is that it has been providing tennis match live commentary service for a long time, and have hundreds of tennis matches data available to the public. This is convenient and useful for building our knowledge base. Other reason is its comments follow a fixed pattern, making it possible for our program to retrieve information from these comments. Although these comments are written in Chinese, we can translate them into English in the human-annotation stage.

Figure 2 shows the original Web page for the match. The lower-left corner is the commentary area where the latest comments are popped up regularly. If users go to the interface of our system which is a web page, they can see a "OPEN" button at the lower-right corner of the page. Click it, type in the address of the tennis web site in the pop-up window, and submit it, a visualization page for the match will be presented to the users (See Figure 3). The lower panel of the page contains a search box which can enable users' interaction with TennisVis. For example, if a user submit a question like "How many Aces have been served in the first two sets?", Figure 4 is the returning visualization which shows the user the ACEs in the first two set of the match. Figure 5 is another visualization in response to the question "How many Volleys have been hit?".

Conclusion

We have presented a new visualization technique for presenting tennis match statistics and live comments. Retrieving data automatically from tennis match live blogging Web sites, this visualization is designed as a supplement to live tennis match coverage. It provides a quick overview of a tennis match statistics but can also provide many details on demand. While existing tennis data visualizations focus on post-match analysis, our visualization is the first to feature automatic data retrieval and live visual broadcasting. All the basic match statistics are presented in a series of mini-timelines. The concise form is particularly suitable for mobile devices. Users can interact with the visualization by typing a question. In the future, we plan to enhance our program with better natural language processing capabilities as well as the ability to integrate data from multiple sources.

Author Biography

Xi He is a PhD candidate of Computer Science at Georgia State University. His research interests focus on text visualization and storyline visualization.

Ying Zhu is an Associate Professor of Computer Science at Georgia State University (GSU), where he leads the Graphics and Visualization Group. His research areas are 3D computer graphics, data visualization, and human computer interactions. He is also an Associate Member of the Neuroscience Institute. His research projects have been supported by the National Sci-

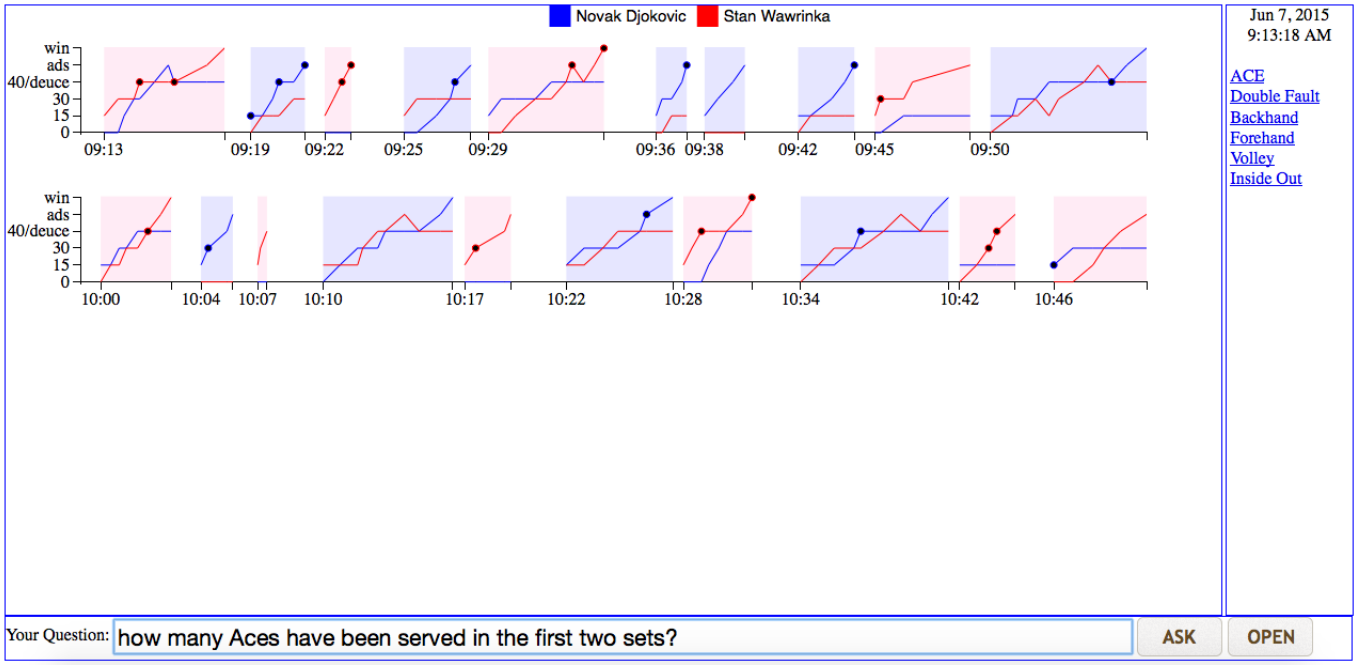

Figure 4: The TennisMatchViz Visualization in Response to A User Question about Aces

ence Foundation (NSF) and National Institute of Health (NIH). He teaches a comprehensive set of courses on computer graphics, game design, and data visualization.

References

[1] "Us open men's final: Djokovic beats federer," http://www.bbc.com/sport/live/tennis/34067072, accessed: 2015-09.

[2] "The 1st workshop on sports data visualization," http://workshop.sportvis.com/, accessed: 2013-10.

[3] "Atp scores & stats," http://www.atpworldtour.com/en/scores/current/us-open/560/results?, accessed: 2015-08.

[4] "Roger federer," http://www.atpworldtour.com/en/scores/current/us-open/560/results?, accessed: 2015-09.

[5] H. Pileggi, C. D. Stolper, J. M. Boyle, and J. T. Stasko, "Snapshot: Visualization to propel ice hockey analytics," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2819–2828, 2012.

[6] F. Beck, M. Burch, and D. Weiskopf, "Visual comparison of time-varying athletes' performance," in *The 1st Workshop on Sports Data Visualization*. IEEE, 2013.

[7] B. Moon and R. Brath, "Bloomberg sports visualization for pitch analysis," in *The 1st Workshop on Sports Data Visualization*. IEEE, 2013.

[8] D. Demaj, "Geovisualizing spatio-temporal patterns in tennis: An alternative approach to post-match analysis," 2013. [Online]. Available: http://gamesetmap.com/?page_id=2

[9] "Daily data visualization: Analyzing federer's quarter-final comeback," http://www.si.com/tennis/2014/09/06/daily-data-visualization-analyzing-federers-quarterfinal-comeback, accessed: 2014-09.

[10] "Atp tennis world tour venue map," http://gamesetmap.com/atp2015/, accessed: 2015-09.

[11] R. Cava and C. Dal Sasso Freitas, "Glyphs in matrix representation of graphs for displaying soccer games results," in *The 1st Workshop on Sports Data Visualization*. IEEE, 2013.

[12] S. G. Owens and T. Jankun-Kelly, "Visualizations for exploration of american football season and play data," in *The 1st Workshop on Sports Data Visualization*. IEEE, 2013.

[13] R. Sisneros and M. Van Moer, "Expanding plus-minus for visual and statistical analysis of nba box-score data," in *The 1st Workshop on Sports Data Visualization*. IEEE, 2013.

[14] "Infographic design," http://visual.ly/creative-services/infographics, accessed: 2015-09.

[15] T. Polk, J. Yang, Y. Hu, and Y. Zhao, "Tennivis: Visualization for tennis match analysis," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 2339–2348, 2014.

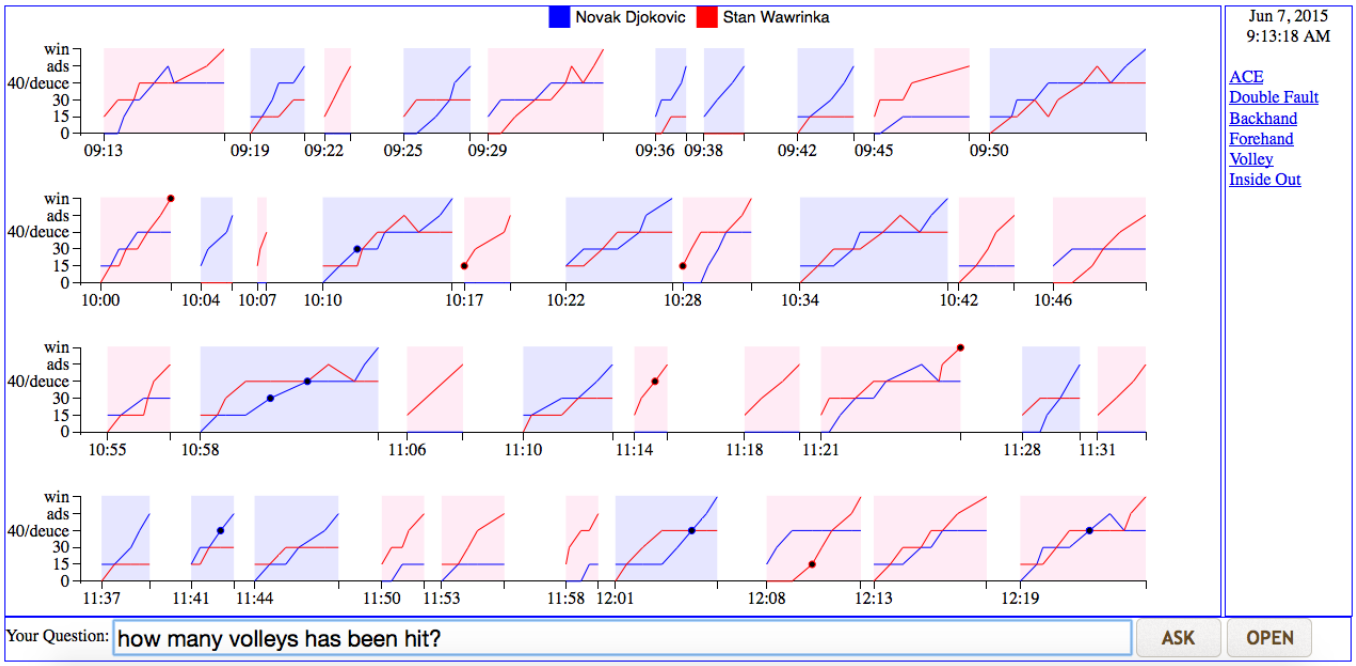

Figure 5: The TennisMatchViz Visualization in Response to A User Question about volleys