Multi-Modal Summarization of Key Events and Top Players in Sports Tournament Videos

Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko and Cher Han Lau
Faculty of Science and Technology, Queensland University of Technology
126 Margaret Street, Brisbane QLD 4000, Australia
{dian, x.tao, w.sasongko, andy.lau}@qut.edu.au

Abstract

To detect and annotate the key events of live sports videos, we need to tackle the semantic gaps of audio-visual information. Previous work has successfully extracted semantic from the time-stamped web match reports, which are synchronized with the video contents. However, web and social media articles with no time-stamps have not been fully leveraged, despite they are increasingly used to complement the coverage of major sporting tournaments. This paper aims to address this limitation using a novel multimodal summarization framework that is based on sentiment analysis and players' popularity. It uses audiovisual contents, web articles, blogs, and commentators' speech to automatically annotate and visualize the key events and key players in a sports tournament coverage. The experimental results demonstrate that the automatically generated video summaries are aligned with the events identified from the official website match reports.

1. Introduction

Major sports tournaments are widely covered by multimodal information, including broadcast TV and web articles, while worldwide fans like to discuss interesting topics via blogs, and share exciting video clips online. With this information torrent, sports fans have a lot of media varieties to follow a tournament and develop new social connections, thus reducing their time in watching a full match. An automatic system that can summarize and visualize the key events and key players of a sports tournament will improve users experience.

Sports highlights can be generally identified by detecting slow motion replay scenes [3], while specific sports events, such as a soccer goal, can be detected automatically based on machine-learning analysis of patterns in specific audiovisual features [10]. Play segments have been widely accepted as the basic semantic event unit that contains a particular self-contained story, thus video key segments can be obtained based on the sematic importance of each play segment [7]. However, to extract the semantic content of a video segment in a descriptive manner, we need to analyze the audio and visual signals into textual contents and apply some text analysis techniques to recognize semantically important keywords. This can be achieved via speech recognition, or optical characters recognition.

The semantic gaps in audiovisual features have been well documented [8], motivating the use of external contents [4]. Sports fans and journalists generally share many common interests, as such trends in social insights would represent events of global users' interests. Hence, web social media, such as match reports, editorial articles, and blogs, can be used to help automatically annotate sports videos and complement the existing solutions for detecting video key events based on audio-visual analysis. Some recent attempts have been made to analyze time-stamped match reports, and use them in identifying semantic events of a video. The web and video contents are synchronized based on the extraction of time from web articles, and super-imposed scoreboards in broadcasted sports videos [9]. However, they have not used web articles with no time-stamps, such as blogs and twitters, whereas these types of social media are becoming more prominent to reveal users' interests.

Most of the state-of-the-art approaches in using text analysis for video events detection have used domain-specific keywords. These techniques need a tailored dictionary for each sports genre, and may not leverage generic descriptive words that often can reveal the significance and excitement of the current play segment. For example, when the commentator makes a statement like ”what a magnificent play by Federer”, the segment should contain a highlight, even it contains no sports-related keyword.

Based on the discussion so far, we have identified three important challenges to extend the current achievements:

1. How do we use non time-stamped web contents, in-

cluding social media, and incorporate them into identifying important contents in a sports match?

2. How do we develop a domain-independent approach in analyzing the important keywords?

3. How do we develop a multi-modal analysis framework to automatically detect, annotate, and visualize sports summaries in match and tournament levels?

This paper aims to address these challenges, while making contributions to the current achievements in video events detection, text analysis, and multi-modal summarization and visualization. For the experiment, tennis tournament broadcasts are used as a case study due to its world-wide popularity, thus maximizing the impact, and to test the non-time-stamped game structure. Furthermore, it is generally more difficult to detect highlights in tennis games as we cannot simply detect specific keyword-based highlights such as “goal”, “foul”, etc. This work is also one of the few attempts to summarize sports videos in the level of a tournament, as most current research work usually focuses on match-level summarization.

The paper is organized as follows. Section 2 outlines the overall system framework, and describes the first two modules: video/text alignment and social media classification, while Section 3 focuses on the text analysis module. Section 4 discusses the summarization process. Section 5 analyzes the experimental results, and Section 6 provides conclusion and directions for future work.

2. System Framework

A multi-modal sports video summarization system is proposed. This domain-independent system aims to summarize a sports event at different levels of matches and tournament by analyzing textual information extracted from multiple resources, and identify important contents in a sports video.

Figure 1 depicts the conceptual framework of the proposed system, which consists of three key modules:

Video/Text Alignment Module identifies the boundaries of play segments as the basic/smallest semantic units From each play segment, the speech transcript is extracted to reveal what the commentators said during the game;

Social Media Classification Module performs Web and social media crawling, and classify the social media into the corresponding matches using the features extracted from the match-specific articles.

Text Analysis Module serves as the core of the proposed summarization system, which evaluates the interestingness of the sports at different levels of segmented clips, including points, games, matches, or the entire tournament. The interestingness value is evaluated by analyzing textual data extracted from the videos and crawled from the Web by utilizing natural language processing and sentiment analysis techniques;

Based on the outputs of all modules, the system identifies the key players and key events during a sports match and tournament. The details of the first two modules will be described here, while the text analysis module will be discussed in the next section.

2.1. Video/Text Alignment Module

Super-imposed texts on sports video are used by broadcasters to describe some important information about the current match to allow viewers keep track on what is currently happening and being reminded of what has previously been happening [6]. In tennis videos, the score-line can keep track of the match progress and identify the boundary of all play segments.

A play segment contains the scenes where the ball is in play. In tennis, every time a play is completed, it marks the completion of a *point* (i.e. 15, 30, 40, Advantage), and depending on the current score, the boundaries of a *game* (i.e. 1 to 6, up to 7) and *set* (1 to 5) can be identified. A *match* boundary is determined by player change.

However, we found that a score change cannot always accurately detect the start and end of a play, as the scoreboard can be displayed in between two play segments. As such, a tennis point segment needs be corrected to cover the *serve* (i.e. start-of-play), the *score change* (i.e. end-of-play), and the slow motion replay scene. Figure 2 illustrates the process of correcting play boundaries to take this structure into account.

By observing the audio energy levels, each of the start and end positions of a segment is adjusted to the time where a relatively silent subsegment occurs. This strategy uses the fact that in tennis custom, the umpire always requires spectators to be quiet before a serve and until the ball is out of play. This observation is consistent with the result of another study showing significant correlation between ”silence” and ”ball hits” classification in tennis audio [2].

For our implementation, optical characters recognition was performed using Tesseract OCR (http://code.google.com/p/tesseract-ocr/) to extract the on-screen scoreboards into characters, which can be analyzed into numbers or letters and then used for revealing the current score-line after applying a simple post-processing to correct the characters.

After video segmentation, the system correlates the (play) segments’ time-stamps with the subtitle time-stamps, and builds a database of each play segment and its commentator’s speech contents such as the one shown in Table 1.

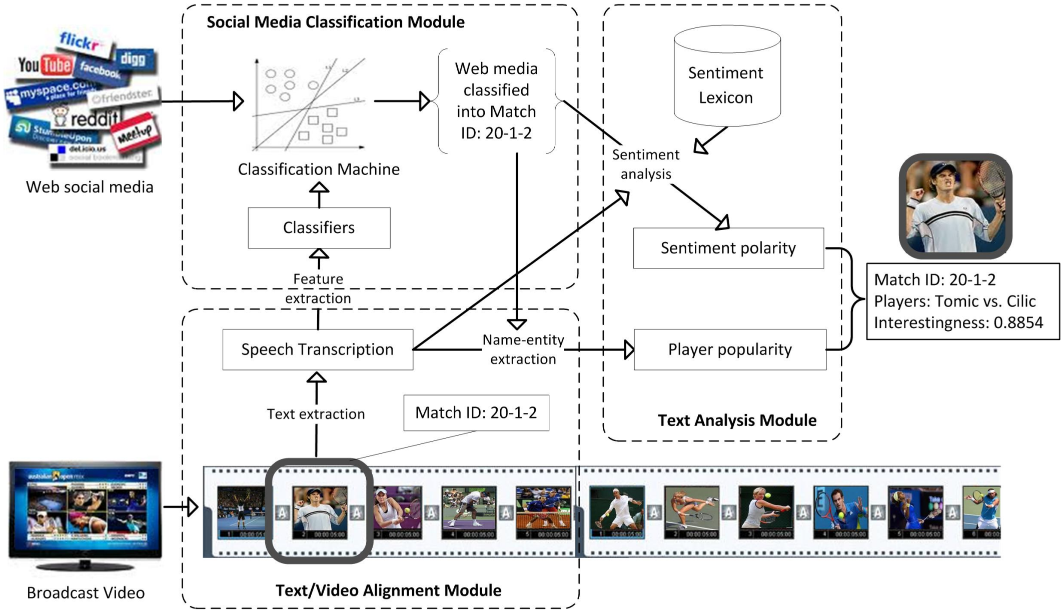


Figure 1. System Framework

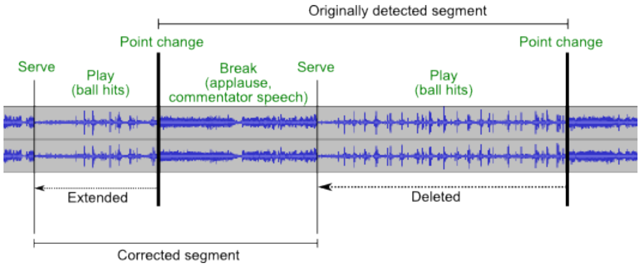


Figure 2. Audio Structure of a Tennis Point and the Segment Correction Process

Match	Time	Int	Speech Transcription
20-1-2	10665	2	I can't get over how quick these games are going
20-1-2	10755	1	And that is not a fast serve either
20-1-2	10779	0	And that is the consistency we were talking about
20-1-2	10805	3	Three consecutive break of serve so she would be looking for a hold to break that trend
			(a)
20-1-2	10855	4	That's the advantage that all the top players get, because they play basically all their matches on either Rod Laver Arena or Hisense Arena
20-1-2	11025	0	Is it just me or have the grunts got louder again
			(b)

Table 1. A Sample of Video and Text Alignment with Time Tags and Interestingness Values

2.2. Social Media Classification Module

Web news and social media can be used to reveal the interestingness of sports games. An interesting game usually charges public's attentions. The discussions related to such an interesting game are usually alive. In contrast, a less interesting game will die quickly in the sight of public because they have nothing interesting to talk about. The aim of social media classification in the proposed system is to utilise the rich textual information from (less structured) web blogs and combine it with web news to identify important events and players during a sport tournaments.

To extract the features of a document, we applied some text preprocessing, including noise removal, stopwords removal, word stemming, and term grouping. The weight w of a term is defined by the multiplication result of $tf \times idf$, a common method for feature extraction from text.

The next step is to classify the articles from web news and social media to their respective (tennis) match based on the content similarity. We employ cosine distance [1] as the similarity metric. Compared to other measurement, cosine distance measures the similarity of two articles by the angle between the two documents vector, the similarity score is between [0,1] with 0 being the most dissimilar and 1 being the most similar. As some articles and blogs can discuss about multiple matches, this method can assign one article to multiple videos as long as they are similar. When the similarity measure is below an empirically set threshold value, an article will not be assigned with the match video.

3. Text Analysis Module

Using the video/text alignment database, the system determines if a play segment is interesting by analyzing the aligned text. If the text contains many interesting elements, the play is interesting. In the same way, we can also determine the interestingness of a match by counting how many interesting play segments it contains and how strong these interesting clips are.

Figure 3 shows the text analysis ladder, stepping-up from the sentence level to the aligned play segment level, the

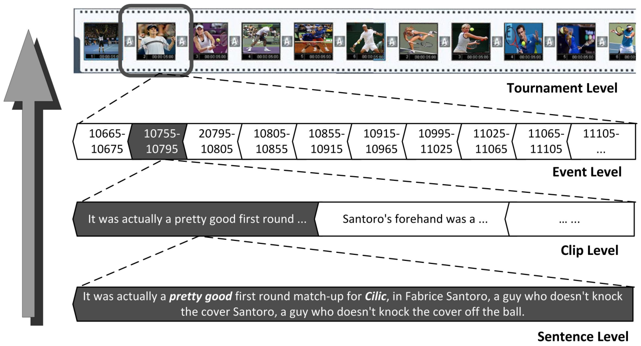

Tournament Level
10665-
10675
10755-
10795
20795-
10805
10805-
10855
10855-
10915
10915-
10965
10995-
11025
11025-
11065
11065-
11105
11105-
...
Event Level
It was actually a pretty good first round ...
Santoro's forehand was a ...
... ...
Clip Level
It was actually a pretty good first round match-up for Cilic, in Fabrice Santoro, a guy who doesn't knock the cover Santoro, a guy who doesn't knock the cover off the ball.
Sentence Level

Figure 3. Text Analysis Ladder for Summarization

match level, and eventually the tournament level. However, this text analysis ladder only applies for the text extracted from videos, not from Web and social media. There is no such a precise alignment existing between the contents of social media (such as blogs) and video segments. Thus, the primary (prioritized) source of information in our text analysis are web articles that describe a match. The following sub-sections will discuss in details how interestingness measures are calculated.

3.1. Sentiment Analysis for Interestingness Evaluation

From a large sample of broadcast sports videos, we observed that the interestingness level of a game can be revealed from the terms used by the commentator(s). For example, Table 1 (a) shows that when a game was more interesting or exciting, many emotional adverbs and adjectives were used more strongly and frequently, such as "how quick ... these games", "fast serve", and "three consecutive breaks"; whereas, less emotional terms were used when the game was less interesting. Table 1 (b) shows how commentators talked outside of the scope of what actually happened in the video and offered their personal opinion on certain topics, such as the arena, and the grunts during the tournament, when the play is less interesting. As such, the first factor of interestingness is the polarity at the sentence level, which can be evaluated by counting the number of emotional terms appearing in a sentence:

$$pol(s) = \sum_{t_i \in \vec{s}} v(t_i) \tag{1}$$

where $pol(s)$ denotes the polarity value of the sentence s, t is a term in the vector of $\vec{s}$, and $v(t)$ is a function evaluating the polarity of t. $v(t)$ returns 2 if t is a strong adverb or adjective, returns 1 if a weak adverb or adjective, or 0 if not an adverb or adjective. As a result, the line with time stamp 10665 in Table 1 (a) has a polarity value of 2, line with 10755 only has a value of 1, and line with 10855 has a large polarity value of 4 because more strong emotional terms appears in it.

By aggregating the polarity values of all sentences $s \in \tau(p)$ (where $\tau(p)$ is a function returning the text captured from the speech given by the commentator(s) during a play-segment p), the polarity of a play-segment $pol(p)$ can be evaluated. By laddering up with the same mechanism, we aggregate the $pol(p)$ for all p in a match m to evaluate the match polarity $pol(m)$.

In our implementation of the proposed model, the subjectivity lexicon generated by OpinionFinder (see http://www.cs.pitt.edu/mpqa/) was employed to measure the sentiments and their polarities. The OpinionFinder subjectivity lexicon is a large repository including 8221 lexicons with weak or strong polarity specifications. This proposed method is domain-independent because no expert knowledge is applied to mark sports specific keywords.

3.2. Name-Entity Extraction for Player Popularity

Entity extraction techniques can identify person, location, organization, and other semantically important concepts [5]. In sports context, players can be used as the centre point of attention as users' interest usually revolve around a particular player, especially when the sports game is mainly individual (i.e. one versus another) like tennis. Based on a comprehensive observation, we found that when a game was interesting, the commentators' speech and web social media were often focused on the key players who were playing (in the current match). At the time a game was less interesting, the frequency of the key players' names was considerably lower, comparing with that of other names. Instead, the commentator(s) usually use other stories to attract audiences to keep their attention. Based on this observation, it is assumed that the frequency of the key players' names appearing in the sports context is a factor revealing the interestingness of the game, as well as the distribution of different names' frequencies in the context.

Based on this assumption, the system measures the popularity *pop* of an entity e (a player in this case) in a sentence s by simply counting its frequency: $pop(e, s) = freq(e, s)$. By laddering up with the same mechanism, the popularity of e at the play-segment level $pop(e, p)$ can be measured by aggregating the $pop(e, s)$ for all s in the text of p. Finally, we can evaluate $pop(e, m)$, the popularity of player e in a match m, by aggregating the $pop(e, p)$ for play segments e played. The normalized form of $pop(e, m)$ can then be described as:

$$npop(e, m) = \frac{pop(e, m)}{\sum_{e_j \in E} freq(e_j, \tau(m))} \tag{2}$$

where E denotes the set of all players in the tournament.

Theoretically the value of $npop(e, m)$ is between [0,1]. If a player is not mentioned in the context, his (her) popularity is 0. If only one player is mentioned in the context of a match, the player receives the highest popularity value of 1. In the real-world, the highest value of popularity should not happen because no sport games play with just one player. Thus, an interesting game, especially an individual game having only two players against each other, has their players sharing the full popularity, and having the popularity values close to 0.5.

To evaluate the interestingness of an individual match, the top two popularity values are chosen for the key players. For team-based matches like soccer, the number of popularity values considered depends on the number of players in a team.

4. Multi-level Summarization of Sports Videos

Summarization of sports videos can be achieved using different viewpoints. When summarizing a tournament, users may want to just watch *the most interesting (or most boring) matches*, or *the most popular players*, etc. When summarizing a specific match, users are usually interested with the *the most interesting moments in the match*. Using the outputs of the proposed method to evaluate the sentiment polarity and player popularity, the system can achieve summarization of sports games at match and tournament levels.

4.1. Summarizing a Match

The summarization at the sentence level should be made first before summarizing at the play segment, a match, or a whole tournament level, according to the mechanism depicted in Fig. 3. Based on the justifications discussed in Section 3, the interestingness int of a sentence can be evaluated, giving consideration of both sentiment polarity and the key players' popularity:

$$int(s) = npol(s) + \frac{\sum_{e_k \in \rho(m)} npop(e_k, s)}{|\rho(m)|} \tag{3}$$

where $\rho(m)$ is a function returning all the players performing in match m.

$Npol(s)$ is the normalized form of $pol(s)$ and is calculated by:

$$npol(s) = \frac{pol(s)}{max_{\{s,s_i\} \subseteq \tau(m)}(pol(s_i))} \tag{4}$$

$max_{s_i \in m}(pol(s_i))$ is a function returning the max polarity value of sentences occurring in $\tau(m)$ that contains s as well.

$npop(p, s)$ is the normalized form of $pop(p, s)$ and is calculated by:

$$npop(p, s) = \frac{pop(p, s)}{max_{\{s,s_i\} \subseteq \tau(m)}(pop(p, s_i))} \tag{5}$$

$max_{s_i \in m}(pop(p, s_i))$ is a function returning the max popularity value of sentences occurring in $\tau(m)$ that contains s as well.

The interestingness of a play segment can then be evaluated by aggregating the values of its containing sentences:

$$int(p) = \sum_{s \in \tau(p)} int(s) \tag{6}$$

With the evaluation of play segment's interestingness and the text/video alignment table displayed in Table 1, a match can be summarized by a chart with a curve plotting to describe the interestingness flow of the match. Figure 5 illustrates a sample summarization of a match, whereby the top graph indicates the most interesting moments in the game based on the temporal flow of the play segments' interestingness score. Next to the graph, each of the highlight segments (i.e. top N play segments) are described by the interestingness value, the key frames and the contents of commentators' speech. The links to the related web and social articles are also displayed to enrich the story of the match.

4.2. Summarizing a Tournament

A tournament can be summarized by expanding the view from the match level to the tournament level, as depicted in Fig. 3. With the interestingness values of play segments are measured, one single value of interestingness can also be calculated and assigned to a match:

$$int(m) = int_\sigma(m) + \sum_{p \in m} int(p) \tag{7}$$

where σ indicates the evaluation of social media and

$$int_\sigma(m) = \frac{pol_\sigma(m)}{max_{m_i \in \mathcal{T}} pol_\sigma(m_i)} + \frac{\sum_{p \in \rho(m)} npop_\sigma(e, m)}{|\rho(m)|}$$

where

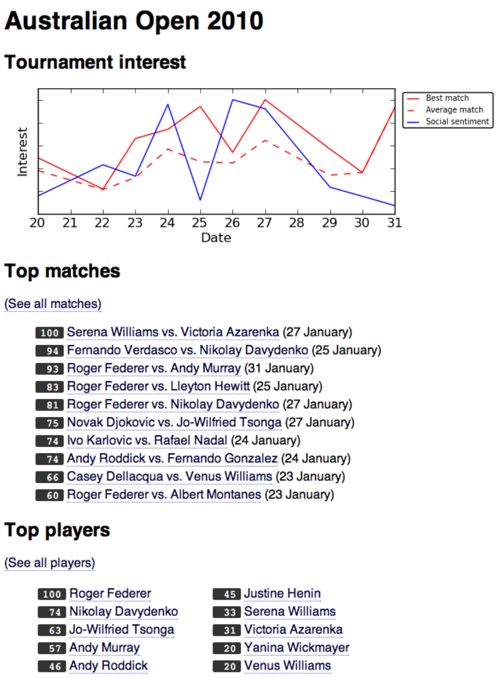
$$npop_\sigma(e, m) = \frac{pop_\sigma(e, m)}{max_{e_k \in \rho(m)}} \tag{8}$$

A tournament can then be summarized, after evaluating the interestingness of all containing matches. Figure 4 depicts the summarization of Australian Open 2010 tournament by plotting the interestingness flow of all matches, showing the most interesting matches, and the top (i.e. most popular) players.

4.3. Summarizing the Top Players

The performance of a player in the whole tournament can also be summarized, based on the player's popularity values calculated by Eq. (5) and (8):

$$npop(e, m) = npop_\sigma(e, m) + npop(e, m) \tag{9}$$



Australian Open 2010

Tournament interest

Top matches

(See all matches)

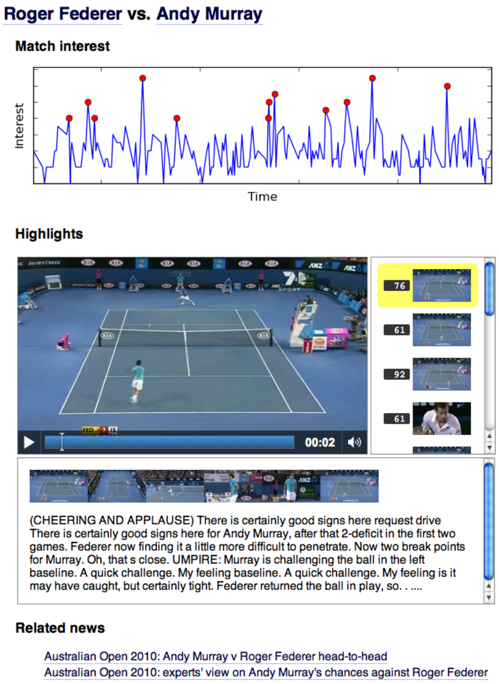
100 Serena Williams vs. Victoria Azarenka (27 January)
94 Fernando Verdasco vs. Nikolay Davydenko (25 January)
93 Roger Federer vs. Andy Murray (31 January)
83 Roger Federer vs. Lleyton Hewitt (25 January)
81 Roger Federer vs. Nikolay Davydenko (27 January)
75 Novak Djokovic vs. Jo-Wilfried Tsonga (27 January)
74 Ivo Karlovic vs. Rafael Nadal (24 January)
74 Andy Roddick vs. Fernando Gonzalez (24 January)
66 Casey Dellacqua vs. Venus Williams (23 January)
60 Roger Federer vs. Albert Montanes (23 January)

Top players

(See all players)

100 Roger Federer
74 Nikolay Davydenko
63 Jo-Wilfried Tsonga
57 Andy Murray
46 Andy Roddick
45 Justine Henin
33 Serena Williams
31 Victoria Azarenka
20 Yanina Wickmayer
20 Venus Williams

Figure 4. Summarization of a Tournament



Roger Federer vs. Andy Murray

Match interest

Highlights

(CHEERING AND APPLAUSE) There is certainly good signs here request drive There is certainly good signs here for Andy Murray, after that 2-deficit in the first two games. Federer now finding it a little more difficult to penetrate. Now two break points for Murray. Oh, that s close. UMPIRE: Murray is challenging the ball in the left baseline. A quick challenge. My feeling baseline. A quick challenge. My feeling is it may have caught, but certainly tight. Federer returned the ball in play, so. . ....

Related news

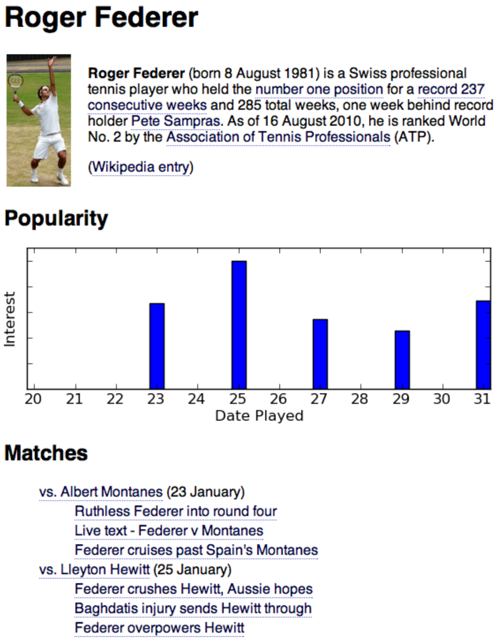
Australian Open 2010: Andy Murray v Roger Federer head-to-head
Australian Open 2010: experts' view on Andy Murray's chances against Roger Federer

Figure 5. Summarization of a Match

where

$$npop(e, m) = \frac{pop(e, m)}{max_{e_i \in \rho(\mathcal{T})}(pop(e_i, m))}$$

$$pop(e, m) = \sum_{s \in \tau(m)} npop(e, s)$$



Roger Federer

Roger Federer (born 8 August 1981) is a Swiss professional tennis player who held the number one position for a record 237 consecutive weeks and 285 total weeks, one week behind record holder Pete Sampras. As of 16 August 2010, he is ranked World No. 2 by the Association of Tennis Professionals (ATP).

(Wikipedia entry)

Popularity

Matches

vs. Albert Montanes (23 January)
  Ruthless Federer into round four
  Live text - Federer v Montanes
  Federer cruises past Spain's Montanes
vs. Lleyton Hewitt (25 January)
  Federer crushes Hewitt, Aussie hopes
  Baghdatis injury sends Hewitt through
  Federer overpowers Hewitt

Figure 6. Summarization of a Player

Figure 6 depicts the player summary of Roger Federer in the tournament of Australian Open 2010. One may see from the summarization that Roger had the most exciting performance on the 25 January, against Leyton Hewitt, who was the tournament host's popular player.

5. Results and Analysis

The proposed system framework has been fully implemented and tested using a large set of tennis videos from the Australian Open 2010 tournament, and the related web and blogs media coverage. The experiment results have demonstrated the effectiveness of our algorithms in detecting and annotating the key events and important players in the tournament. In this section, we will outline the details of our experiment and analyze the findings.

5.1. Data for Experiments

The video, audio, and transcript data used in this experiment were captured from Channel 7 Australia's digital TV broadcast of the Australian Open 2010 tennis tournament. It should be noted that not all matches in the tournament were broadcasted, and not all broadcasted matches were recorded by our system. The final dataset covers 33 matches (from both men's and women's singles) and consists of around 66 hours of recording, spanning from the 20th to the 31st of January, 2010.

The web and social media used in the experiment consists of 1,250 articles collected from 278 sources of web and social media, including the official website, official blogs, national and international news media and blogs which discuss about the Australian Open 2010 tournament from the 14th Jan to the 4st of Feb, 2010 (to cover the discussions before and after the tournament). Some of the most important sources selected in our experiment are: ABC Australia, Australian Open official website, BBC, Breitbart blogs, CNN, Daily Mail, ESPN, SMH, Herald Sun, LA Times, New York Times, NineMSN, Reuters, and Yahoo news.

5.2. Experimental Results

Based on Fig. 4, the system automatically identifies that the most important top matches from the video data set are: 1) Serena Williams vs. Victoria Azarenka, which marked a remarkable comeback from Serena who otherwise had a bad tournament so far); 2) Fernando Verdasco vs. Nikolay Davydenko, which was the most memorable longest match in the tournament, and 3) Roger Federer vs. Andy Murray, which was the final match for men single. To demonstrate the importance of these matches, we identified that the first top match had 1,220,000 results from Google search, compared to the 116,000 search results from the 9th top match (Dellacqua vs. Williams), indicating that the top match is more important and has been widely covered. Similarly, the second and third top matches had 1,120,000 and 4,910,000 results accordingly. The Google searches were performed on 17 Aug 2010 using the search term: ”player name 1 vs. player name 2 (Australian open 2010)”.

To verify the top players, 1) Roger Federer, 2) Nikolay Davydenko, and 3) Jo-Wilfried Songa have 14,800,000; 2,200,000; and 4,320,000 results from Google Search respectively. This is compared to the 1,070,000 results found for the 9th placed player.

It should be noted that the interestingness values had been normalized to reduce the effect of uneven video data set for different players. For example, Federer is top performer not because there were more videos recorded containing his matches, but because the quality of his performance and his popularity throughout the tournament. Moreover, the top popular players are covered more highlight matches. This demonstrates the benefits of combining the players' popularity scores and the play segments' interestingness scores to determine the most exciting matches in the tournament.

To verify the automatically detected key events for each match, we have used the official web site's match report as the base line. To perform the evaluation, we manually segmented the web match report into the paragraphs and sections for each set, and look for the matching key events from the system. Based on the observations, we found that the proposed system was able to find the key events in tennis matches which directly correspond to the events reported in the web match report. This is despite the fact that the system did not directly use match reports to identify the key events as they cannot be synchronized with the video segments due to the absence of time stamps in the report. The followings will show evidences of the direct alignments between our summarization results and the official match report using some examples of the top 3 matches. For each evidence, the first line shows the web match report while the second line shows the commentator's speech from the corresponding highlight segments, which were automatically detected by the system.

For the Federer vs. Murray final men single match, the system has produced 13 highlights and all of them represent the key moments described in the web match report. For example:

Evidence 1 (Set 1) *Murray found himself down a break at 0-2 in the first set before rallying to break back and then level at 2-2.*

> *There is certainly good signs here for Andy Murray, after that 2-deficit in the first two games. Federer now finding it a little more difficult to penetrate. Now two break points for Murray.*

Evidence 2 *He (Federer) produced a superb backhand winner down the line to convert a single break point against Murray's serve in the eighth game to lead 5-3.*

> *Look how fine that one-hander and go down the line for a cold winner. Look how fine he cut that, Federer. He s pretty good ...*

For the Verdasco vs. Davydenko match, the system produced 12 highlights, and they all represent the important parts of the match as reported on the web match report. For example:

Evidence 3 (Set 1) *... Davydenko realised that he had to keep the ball on Verdasco's backhand side or he would be spending the match ducking for cover.*

> *Greater defence from Davydenko. Finally fails at the end of the rally but Verdasco has a bit of a fear factor right now ...*

Evidence 4 (Set 3) *Verdasco, down two sets and with nothing to lose, came out firing in the third set. Big serves and a return to the booming forehands down the line revived the crowd and Verdascos support staff.*

> *There has been so many free points and so many double faults and so many unforced errors he has not been able to get into that rhythm of playing ... and proving how tough he is ...*

For the Williams vs. Azarenka match, the system produced 16 highlights, which are all aligned with the match reports. For example:

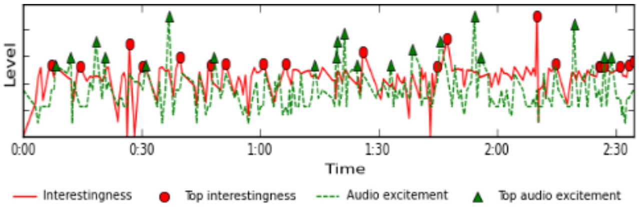


Figure 7. Comparison of audio-based and our interestingness measure

Evidence 5 (Set 1) *Williams was troubled by the 20-year-old from the beginning as she had her serve broken in the opening game of the match.*

> *To think she had not been broken until the first game out here today and the entire tournament and now she's in a deep hole here. A credit to Azarenka ...*

Evidence 6 (Set 2) *But, just as the crowd sensed a second upset following Venus Williams departure at the hands of Li Na Serena struck back.*

> *There s the break, the first of them. She s hanging on ... With Venus out, Li Na going through ...*

During interesting plays in tennis matches, the crowd and commentator usually gets excited, thereby the audio excitement level can measure the interestingness of a game. Figure 7 demonstrates that our system can detect the key events in the Federer vs. Murray match, which are not necessarily accompanied with audio excitement.

6. Conclusions and Future Work

This paper has presented a multi-modal framework to automatically detect and annotate key events and top players in broadcasted sports videos from a match and tournament coverage. The proposed method is novel as it is one of the first attempt to use match reports, web articles, and blogs which have no time-stamps, and thereby cannot be synchronized with the corresponding audiovisual contents. Moreover, our algorithms in extracting key events and key players are based on sentiments and players' popularity, instead of relying on sports-specific keywords that have been mainly used in state-of-the-art approaches, thus ensuring robustness of the proposed approach.

The system has been fully implemented and the experiment was conducted comprehensively using a large video data set (33 matches) from the Australian Open 2010 tournament, and 278 sources of web articles and blogs. The results have shown that the proposed framework is able to successfully detect the important players and key events, and use the various modalities to create an appealing visualization.

For our future work, we will test the robustness of the framework for different sports tournaments, and conduct a comprehensive user tests to determine whether the key events and top players are aligned with their views (especially if the user is a regular follower of the tournament), and whether the generated summary visualizations are appealing for them. We also aim to measure the impact of our proposed multi-modal summary on how it can change the way that users follow sports tournaments using a one-stop service, particularly in mobile settings.

7. Acknowledgment

This research was carried out as part of the activities of, and funded by, the Smart Services Cooperative Research Centre (CRC) through the Australian Government's CRC Programme (Department of Innovation, Industry, Science and Research).

References

[1] Daniel Billsus and Michael J. Pazzani. A hybrid user model for news story classification. In *UM '99: Proceedings of the seventh international conference on User modeling*, pages 99–108, Secaucus, NJ, USA, 1999. Springer-Verlag New York, Inc.

[2] Ewa Kijak, Guillaume Gravier, Lionel Oisel, and Patrick Gros. Audiovisual integration for tennis broadcast structuring. *Multimedia Tools and Applications*, 30(3):289–311, 2006.

[3] M.H. Kolekar, K. Palaniappan, and S. Sengupta. A novel framework for semantic annotation of soccer sports video sequences. In *Visual Media Production (CVMP 2008), 5th European Conference on*, pages 1–9, 2008.

[4] Hua-Yong Liu and Hui Zhang. A sports video browsing and retrieval system based on multimodal analysis: SportsBR. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 8, pages 5077–5081 Vol. 8, 2005.

[5] Dan Roth and Wen-tau Yih. Probabilistic reasoning for entity & relation recognition. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[6] Huang-Chia Shih and Chung-Lin Huang. Content extraction and interpretation of superimposed captions for broadcasted sports videos. *Broadcasting, IEEE Transactions on*, 54(3):333–346, 2008.

[7] Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. The power of play-break for automatic detection and browsing of self-consumable sport video highlights. In *Multimedia Information Retrieval*, pages 267–274, 2004.

[8] J. R. Wang and N. Parameswaran. Survey of sports video analysis: research issues and applications. In *Proceedings of the Pan-Sydney area workshop on Visual information processing*, pages 87–90. Australian Computer Society, Inc. Darlinghurst, Australia, Australia, 2004.

[9] Changsheng Xu, Yi-Fan Zhang, Guangyu Zhu, Yong Rui, Hanqing Lu, and Qingming Huang. Using webcast text for semantic event detection in broadcast sports video. *Multimedia, IEEE Transactions on*, 10(7):1342–1355, 2008.

[10] Huaxin Xu and Tat-Seng Chua. Fusion of AV features and external information sources for event detection in team sports video. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):44–67, 2006.