Using Minute-by-Minute Match Report for Semantic Event Annotation in Soccer Video

Zengkai Wang, Junqing Yu*
School of Computer Science and Technology
Huazhong University of Science and Technology
430074, Wuhan, China
zengkai.w@gmail.com, *Corresponding: yjqing@hust.edu.cn

ABSTRACT
In this work, we propose a soccer video annotation approach based on semantic matching with coarse time constraint, where video event and external text information - match report are synchronized by their semantic correspondence along the temporal sequences. Different from the state of the art soccer video analysis methods which assume that the time of event occurrence is given precisely in second, this work solves the problem that how to annotate the soccer video using the match report with coarse gained time information. Compared with previous approaches, the contributions of our approach include the following. 1) The approach synchronizes the video content and text description by their high-level semantics with coarse time constraint instead of the exact timestamp. In fact, most of the text descriptions from the famous sport websites provide the coarse time information in minutes rather than seconds. Therefore, we argue that our approach is more generalized. 2) We propose an attack-defense transition analysis (ADTA) based soccer video event boundary detection method. The previous methods give coarse boundaries which could be refined, or simply give the clips with fixed duration which may cause larger bias. The results of our method are more in line with the development process of soccer events. 3) Different with the existing audio features analysis based whistle detection method, we propose a novel Hough transformation based whistle detection algorithm from the perspective of image processing, which facilitates the game start time detection combing with the ellipse detection algorithm, and further helps the synchronization of video and text events. The experimental results conducted on large amount of soccer videos validated the effectiveness of our proposed approach.

Categories and Subject Descriptors
I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms
Algorithms, Experimentation, Verification.

Keywords
Video analysis, event annotation, event boundary detection, attack-defense transition analysis, matches report.

HuEvent'14, November 7, 2014, Orlando, Florida, USA.

ACM 978-1-4503-3120-3/14/11…$15.00.
http://dx.doi.org/10.1145/2660505.2660511

1. INTRODUCTION
The widespread popularity and tremendous commercial potentials of soccer game make automatic event detection and annotation essential for querying the semantic content of soccer videos. Many works try to annotate video content by the intrinsic audio-visual features and complicate learning models. However, the gap between low-level features and high-level semantics dictates the difficulty to infer detailed event semantics. The published results [1] could give simple labels for objects such as jersey number of player, field zone, goalmouth, ball etc., and for events like goal, shoot, foul, but could not tell us who and how scored. External information such as web-casting text [2-4], which contains particular description that audio-visual signals lack, could be utilized to solve the problem. However, the main issue of this method is how to match the text description with video content. The most existing works of web-casting text based method assume that the text description has accurate timestamp (in seconds) [2]. In fact, this assumption is not always valid. Many text descriptions from the live broadcast pages only provide inexact timestamp in minutes rather than seconds, especially in soccer game. So, how to synchronize the text description with video content still face challenges.

Besides, events are video clips, which should be the time intervals include semantic start and end boundaries, rather than the exact time points. But little studies pay attention on how to determine the semantic boundaries of events accurately. Ref. [5] limits the event boundary to a shot. The work in [8] first detects the event timestamp from web-casting text, then the duration of the aligned event clip is fixed to one minute in the time of timestamp before and after, which is somewhat arbitrary, apparently. References [2, 6, 10] simply take the play-break (PB) segments as the coarse boundaries of events. Ref. [9] determines the boundaries of event clips on affection arousal curve combined with video production knowledge — event temporal transition pattern (ETTP), by which the start point of a far view shot is regarded as the start boundary of an event. However, the far view shot, which usually lasts long time, contains many tedious (non-exciting) contents that could be further filtered. However, with regard to soccer video event summarization and advanced applications like resource constrained video content transportation, it is very important to provide reasonable start and end boundaries for video events. If the duration of detected event clip is longer, it may be high redundancy, and needs expensive transmission cost. Otherwise, viewers could not enjoy the whole exciting moment. So, it is necessary to accurately determine the event boundary, which not only covers the whole exciting moment, but also has lower redundancy.

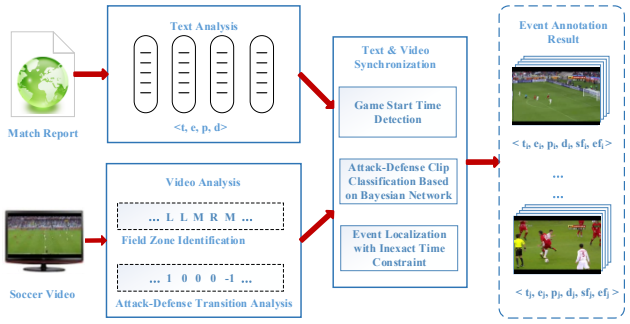


Figure 1. The proposed soccer video annotation framework.

In this paper, we propose a robust soccer video annotation approach based on semantic matching with coarse time constraint, where video content and external text information - match report are synchronized by their semantic correspondence along the temporal sequences. The proposed soccer video event annotation framework is shown in Fig. 1. The match report provides semantic-rich description about the interested/highlight events in soccer game, while it has inexact time information in minutes. We combine text event detection, attack-defense transition analysis, synchronization of attack-defense clip and text event to annotate the video content and accurately determine the event boundaries, which are very important for video summarization, transmission and retrieval. Compared with previous works in the related fields, the main contributions of our approach are summarized as follows:

1) The approach synchronizes the video content and text description by their high-level semantics with coarse time constraint instead of the exact timestamp. In fact, most of the text descriptions from the famous sport websites provide the coarse time information in minutes rather than seconds. Therefore, we argue that our approach is more generalized.

2) We propose to detect event boundary according to the attack-defense transition analysis of soccer game. The previous methods give coarse boundaries which could be refined, or simply give the clips with fixed duration which may cause larger bias. The results of our method are more in line with the development process of soccer events.

3) Different with the existing audio features analysis based whistle detection method, we propose a novel Hough transformation based whistle detection algorithm from the perspective of image processing, which facilitates the game start time detection combing with the ellipse detection algorithm, and further helps the synchronization of video and text events.

2. TEXT EVENT DETECTION

Different with the previous works [3, 4] which use web-casting text with precise timestamp in seconds, we use match report shown in Fig. 2 (both in English and Chinese) in minutes to annotate the soccer video. For pages limited, we do not give the examples of web-casting text here, readers could find the web-casting text in [3, 4] or the famous sports websites such as FIFA[1] or ESPNFC[2]. The characteristics of web-casting text and match report are summarized in Table 1.

[1] http://www.fifa.com/clubworldcup/matches/round=259715/match=300260480/playbyplay.html

[2] http://www.espnfc.com/commentary/395752/commentary.html

Table 1. Comparison of characteristics of web-casting text and match report.

	Web-casting texts	Match reports
Coverage of soccer matches	Important or famous matches	Almost all the matches
Purity of text descriptions	Almost all the events	Focus on interested/highlight events
Semantic-richness for single event	Short on details	In details
Presentation style	Well-defined structure	Freestyle
Time granularity	Second/minute	Minute

Comparing with web-casting text [3, 4], the match report has three advantages. Firstly, the coverage of match report is more extensively than web-casting text. Only the very important or famous soccer matches have web-casting text, while almost all the matches have match reports. Secondly, the description of match report is more pure than web-casting text. The match report mainly focuses on the interesting events, while the web-casting text comprises almost all the events and actions, whatever they are important or not. In fact, the soccer video structure is loose, and the significant events of the game are sparse. Many descriptions, such as sb.'s passing force is too large or sb. failed to pick up the ball, which are not very interesting descriptions are meaningless for soccer video event annotation. The third is that, the description of match report is more detailed for single event than web-casting text. Given the above, it is more generic to use match report to annotate video content, and the results will have richer semantics.

However, there are two challenges should be solved if using match report to annotate the video content. Firstly, different with the well-defined structure of web-casting text which is easy to analyze, the match report has freestyle structure which is not easy to analyze. As claimed in ref. [4], the descriptions of same type of events in the web-casting text have the similar sentence structure and word usage. So, the predefined or learned keywords could be used to definitely recognize the events of web-casting text. But, the event description of match report often comprises more actions and players, which may make misjudgments using keywords matching method. Secondly, the time information of match report is not as precise as some web-casting texts, which make us cannot employ the timestamps of both video and text to synchronize the video content and text description.

For the first challenge, we treat it as a probability classification problem, and the LSI (latent semantic indexing) technology [11] is employed to detect text event. Large corpus of texts for each type of event in match report are manually collected and labeled to train the LSI model. In the classification phase, we first segment the match report into sentences according to the punctuations including full stop or exclamatory marks which mean the end of a sentence, and filter out the irrelevant words including the time-word, stop-word, and player or team names. Then each sentence is regarded as a query document and mapped into latent semantic space to compute the similarities with each type of event. The sentence (event) is classified into the type of event with the max similarity. In addition, the time information and involved players of the query event could be easily extracted by the regular expressions and name matching methods, respectively. Finally, the text event semantics are represented as the four-tuple: $<t, e, p, d, >$, where t is the time, e is the event type, p is the involved player(s), d is the text description.

The South American champions needed only two minutes to get on the scoresheet when Marcos Rocha crossed an early ball to find Diego Tardelli on top of the six-yard box who volleyed into the net.

Guangzhou responded just seven minutes later after Elkeson won the ball in midfield before shooting an initial shot off the crossbar when Lin Gao reacted quickest to pass back for Muriqui to tap in from short distance.

The chances came thick and fast as Atletico's Fernandinho missed a shot just over the crossbar in the 11th minute before Jo had a volley saved by Guangzhou goalkeeper Shuai Li one minute later.

In the 15th minute Guangzhou sprung in front of their Brazilian opponents after Atletico defender Lucas pulled down Lin Gao inside the penalty box, and Dario Conca, who was playing his last match for the AFC Champions, converted the spot kick.

Guangzhou almost took a two-goal lead in the 19th minute when Xiang Sun crossed from the left wing for Elkeson who headed the ball onto Victor's crossbar only to be ruled offside.

Two minutes later Tardelli attempted a 40-metre shot after a long solo run but 'keeper Li saved easily.

After Conca had a dangerous free-kick parried away by Victor, the match settled down with neither side creating significant chances.

Ronaldinho nearly scored another trademark free-kick in the 37th minute, but Li dove to his right to save and keep his side in the lead.

In the 41st minute Muriqui controlled a deflected ball inside the box, shook off a defender only to be denied by Victor who rushed off his line to keep Guangzhou at bay.

(a)

第3分钟皮克铲倒格里兹曼，格里兹曼中路30米外左脚任意球射门被巴尔德斯扑住。第10分钟内马尔被踢伤右脚踝，巴西人起身后一瘸一拐表情痛苦。18分钟巴萨首次射门，伊涅斯塔左路突破后和佩德罗两次做出撞墙式配合，佩德罗大禁区中路左脚推射偏出远门柱。21分钟布斯克茨中场铲倒祖鲁图萨被黄牌警告，皮克防定位球时禁区内无意手球，主裁波尔巴兰拒绝判罚点球，1分钟后何塞-安赫尔前场战术犯规拉倒伊涅斯塔也吃到黄牌。

28分钟祖鲁图萨左路45度传中，卡洛斯-贝拉禁区中路胸部停球后，左脚抽射偏出近门柱。1分钟后卡纳莱斯回传，卡洛斯-贝拉右侧禁区前左脚弧线球推射，巴尔德斯飞身将球扑出。**31分钟卡纳莱斯右路45度战术角球传中，埃鲁斯通多小禁区外前点头球射门，皮球击中亚历山大-宋胸膛后弹入网窝。0比1，巴萨客场落后！35分钟蒙托亚右侧横传，布斯克茨前点巧妙漏球，梅西大禁区中路左脚低射，皮球贴地飞入球门右下死角。1比1，巴萨扳平比分，梅西西甲生涯打进229球，超越劳尔排在历史第三位！**

36分钟萨尔杜亚右路45度传中，迈克尔-冈萨雷斯前点头球射门顶高。38分钟巴尔特拉中场拉拽卡纳莱斯被黄牌警告，1分钟后皮克停球失误把球停在了卡洛斯-贝拉面前，卡洛斯-贝拉大禁区中路左脚推射高出横梁。41分钟皮克头球解围不远，卡洛斯-贝拉点球点后左脚倒钩射门踢偏。44分钟伊涅斯塔战术角球回敲，阿德里亚诺左侧禁区前右脚劲射被后卫挡出底线。

(b)

Figure 2. Examples of match report in minute. (a) English, comes from "www.fifa.com"; (b) Chinese, comes from "sports.sohu.com".

3. EVENT BOUNDARY DETECTION

In soccer game, when the ball moves from one half side of the field to another half side through the mid center zone, it always means the start of a new attack-defense phase. Fig. 3 shows an ordinary play-break (PB) segment, in which the left half zone, mid center zone and right half zone of soccer field are denoted as 1, 0 and -1, respectively. The '*' marks the moment when the attack-defense state changes. The last attack-defense change point (ADCP) of the PB segment is regarded as the start point of the effective attack-defense clip (ADC), whose end boundary is the same with the end boundary of the PB segment.

By this method, the video event boundary could be further refined according to the feature of soccer game. The field zones could be detected by the slope of field lines [7]. Denote the refined event clip as $< sf, ef >$, here, sf and ef are the start and end frame, respectively.

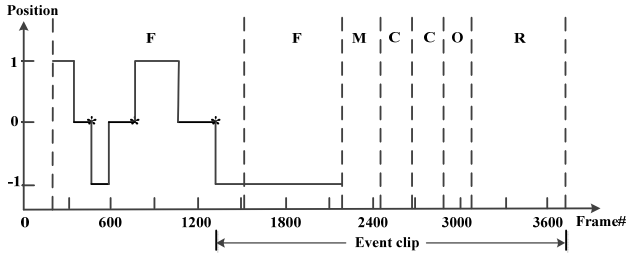


Figure 3. Illustration of event boundary detection (F denotes far view shot, M denotes medium shot, C denotes close-up shot, O denotes out-of-field shot, R denotes replay shot).

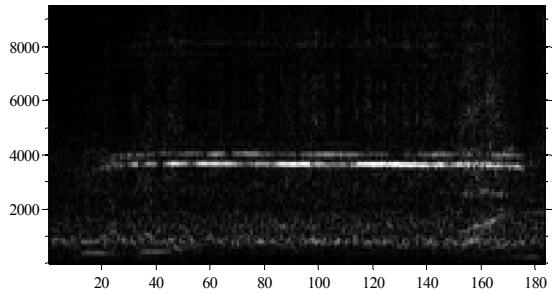
4. SYNCHRONIZATION OF ATTACK-DEFENSE CLIP AND TEXT EVENT

Given the extracted text events and the possible video events with definite start and end boundaries, the next task is certainly to synchronize the text event description with corresponding video event to achieve semantic-rich annotations for soccer video content. If the exact time (in second) of the text event is known, it is easy to locate the corresponding video clip. The event clip which contains the exact time of the text event could be considered as the corresponding video clip. This idea is adopted by the works in refs. [3, 4], which employ the web-casting text with exact timestamp. However, there are many external texts including web-casting text and match report only has minute-by-minute time information. Next, we introduce how to synchronize the video clips and text events with coarse grained time information.

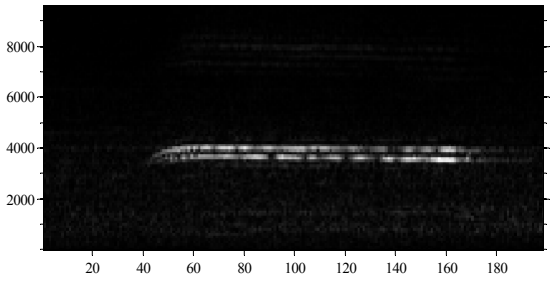
Most of the video events usually last twenty seconds to one minute or so, that is to say, one minute may contain at most 3 events. Because of the minute level bias of match report, it is difficult to automatically find the correct matching relationship between text events and video events. Ref. [8] suggest to perform a manually operation first to find a reference frame, and then use the length-fixed (one minute) event segment around the reference frame to determine the start and end boundary of event. This method could find the correct matching relationship, but is labor consuming and the boundary is not accurate. We try to first determine a location range according to the coarse-grained time information of match report, and then find the correct matching relationship based on the type of text and video event. Obviously, this method has a premise that the start of the soccer video should be the start of the game. According to the game rules, the match always starts up with the referee's whistle and players' kickoff in center circle of the playfield. We can detect the game start time of soccer video using the two salient features, namely whistle and kickoff circle. If the two features appeared in the same time of the beginning of the soccer video and last for certain duration, then we think the game begins.

Ref. [12] computed the zero crossing rates (ZCR) of audio signal, detected whistle by support vector machine (SVM) classifier and an acceptable result was achieved. An improvement was made by typical left-right hidden Markov Model (HMM) with five hidden states in [13], which utilized the classic audio features including Mel-Frequency Cepstral Coefficients (MFCC) and Energy. The above mentioned methods detect whistle from the perspective of machine learning, which need adequate train samples to establish robust models to adapt to complicated audio environments. Here, we propose to detect whistle from the perspective of image processing, which is different with the

conventional audio features based method. The frequency of whistle in soccer video is usually in relative fixed range. Figure 4 shows the time-frequency images of two whistle audio clips in soccer video. There are clear white lines in the frequency range from 3500Hz to 4500Hz in two figures. Based on this observation, the Hough transform (HT) is used to detect line between 3500Hz and 4500Hz, if the duration of the line is greater than 0.1 second, we think that there is a whistle in the audio clip.



(a)



(b)

Figure 4. Two examples of grayscale time-frequency image of whistle audio clips in soccer video. The vertical and horizontal axes denote the frequency and the number of audio frames of whistle audio clip.

Based on the analysis before, a PB segment contains at most one event. As shown in Fig. 5, given the *k-th* text event $TE_k(t_k, c_k, p_k, d_k)$, the location range starts from three PB segments before t_k and ends with three PB segments after t_k. In this interval, there are five candidate attack-defense clips, we should decide which ADC is the most matching event clip with the *k-th* text event. Usually, the type of detected text event is creditable. So, we can choose the video clip with the same event type as the synchronized video clip with *k-th* text event. The work in ref. [2] designs a vote mechanism to choose the PB clip with highest vote score as the synchronized video clip. However, if there are two or more PB clips in the search range are classified into the same type of event, this mechanism does not work.

We use the Naive Bayesian classifier to get the most matching event. The features considered include the following: (1) replay duration (RPD). The duration of replay shot in ADC (in seconds). Usually, the replay duration of goal event is longer than shoot, while the foul event has the shortest replay duration. The optical flow matching algorithm based logo detection algorithm [14] is used to detect replays. (2) Excitement (EXC). The max excitement level of ADC. The goal and shoot event usually has higher excitement than foul event. The excitement value is calculated by affection arousal model [9]. (3) Far view ratio (FVR). Duration of far view frames divided by the duration of ADC. Usually, the foul event has the lowest FVR. (4) Goalmouth ratio (GMR). Duration of goalmouth frames divided by the duration of ADC. The goalmouth usually showed in goal and shoot events, while seldom appeared in the foul events. (5) Whistle (WHS). It is well know that whistle is a strong hint of foul or set-piece kick event in soccer game. (6) Caption (CAP). The caption text is usually appeared when there is a goal event or card event. The discrete cosines transform (DCT) based text detection method [15] is used to determine the start and end frames of each caption, and only the caption appears at the bottom-center of the frame and with sufficient duration is considered.

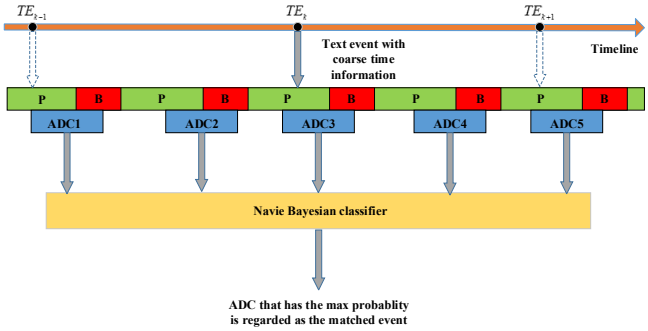


Figure 5. Text and video event synchronization method. (P: play; B: break; ADC: attack-defense clip).

Note that, we only classify the ADC into three event types include goal, shoot and foul. The number of text event types (Table 3) is more than three. It is hard to recognize more event types from the video inherent features by statistical or machine learning methods. Here, other event types share the same BN structure of the three event types. Specifically, the shoot, corner and free-kick events share the same BN structure, and the foul, card, offside and substitute events share the same BN structure. Finally, the annotation result of *k-th* event is represented as $A_k(t_k, sf_k, ef_k, c_k, p_k, d_k)$, where *sf* and *ef* are the start and end frame of the video event, respectively.

Num.	League-team1 vs. team2-[duration:minutes]
1	2012 Olympic Games-Spain vs. Japan-[104]
2	2010 World Cup-Netherlands vs. Spain-[113]
3	2010 World Cup-Portugal vs. Korea-[109]
4	2012 European Cup-Portugal vs. Spain-[108]
5	2010-2011 Serie A-Palermo vs. Rome-[95]
6	2011-2012 English Premier-The second half of Manchester City vs. Manchester United-[54]
7	2011-2012 Bundesliga-Bayern Munich vs. Wolfsburg-[110]
8	2012 UEFA Champions League-Bayern Munich vs. Chelsea-[107]
9	2011-2012 Chinese Super League-Tianjin Teda vs. Beijing Guoan-[104]
10	2012-2013 English Premier League-The second half of Southampton vs. Manchester United-[51]
11	2013 FIFA Club World Cup-Guangzhou Hengda vs. Atletico Mineiro-[101]

Table 2. Soccer video samples.

5. EXPERIMENTAL RESULTS

We conducted our experiment on 1056 minutes of soccer videos (Table 2), which broadly come from various famous soccer matches of the world including the Olympic Games, the World Cup, the European Cup, the UEFA Champions League, the Italian football Serie A, the Chinese Super League, the English Premier League, the German Bundesliga and FIFA Club World Cup. The

match reports for these games are obtained from "sports.people.com.cn", "sports.sohu.com", and "sports.qq.com" which are the most popular sports websites in China.

The LSI is trained on large amount of event corpus, including 201 goal events, 368 shoot events, 70 corner events, 84 free kick events, 112 card events, 50 foul events, 20 offside events and 78 substitute events. The text event detection result is shown in Table 3. As we discussed in previous sections, the free style match report only describes the interested or highlight events, which is different with the well-defined webcasting text. For example, some foul events, which are not serious and do not affect the game process, may not be included in the match report. So, the total number of text events detected from match report is smaller than the number of video events. The number of events in the match report is regarded as the standard number of exciting events.

Table 3. Text event detection result.

Event	Ground truth	Detect Num	Hit	P (%)	R (%)	F (%)
Goal	38	38	38	100	100	100
Shoot	198	203	195	96.1	98.5	97.3
Corner	26	26	26	100	100	100
Free kick	34	34	34	100	100	100
Card	48	52	45	86.5	93.8	90.0
Foul	13	11	10	90.9	76.9	83.3
Offside	6	8	5	62.5	83.3	71.4
Substitute	42	40	38	95.0	90.5	92.7

The semantics of goal, corner and free kick events are captured accurately. The F-measure of the shoot event also achieves a high score 97.3%. However, the precision and recall rate of the other events are lower. There are two main reasons: firstly, the literal expression forms of these events are various in Chinese, it is not easy to detect the semantics very accurately; secondly, different with the well-defined webcasting text which uses the short or concise descriptions to clearly tell the event type, the free style match report gives more details of an event, which may make mistakes. For example, the description that the player shoots on the goal in the possible offside position would make mistakes between shoot and offside events. Although the text event detection results of some events are not very accurate, the main advantage of the match report based video annotation method comparing with the webcasting text based video annotation method is that, the match report is more generic and have richer semantics.

Table 4. Comparison of whistle detection.

Method	Correct	False	Missed	*P*	*R*
HT	54	1	9	98.2%	85.7%
CHMM	49	8	14	86.0%	77.8%

The proposed HT based whistle detection algorithm is compared with typical left-to-right continuous Gauss mixed hidden Markov model (CHMM) based whistle detection algorithm. In the CHMM, The *26*-dimensional feature vectors including *1*-dimensional zero crossing rate, *1*-dimensional short time energy, *12*-dimensional MFCC coefficients, and 12-dimensional MFCC difference coefficients are extracted as the observations of CHMM. The number of hidden state is *5*, and the number of Gauss components is *3* for each hidden state. Both of the proposed HT based whistle detection algorithm and the CHMM based whistle detection algorithm are tested on a *49* minutes soccer video. The results are shown in Table 4. The proposed HT based whistle detection algorithm, which does not need training phase, demonstrated its simpler and more effective performance.

The measurement of boundary detection accuracy BDA [3] is used to evaluate the event boundary. The BDA in this paper is based on the frame level rather than the shot level, and only the correctly detected highlight events are evaluated. As the work in ref. [8] fixed the duration of video clip into one minute, which is evidently unreasonable, we did not compare our method with ref. [8]. The BDA of the proposed approach is compared with the ETTP based approach in ref. [9] and the PB structure based approach in refs. [2, 10]. As shown in Table 5, the proposed approach is better than the other approaches. The average BDA of the proposed approach is 86.8%, while the average BDA of ETTP based approach is 79.2% and PB based approach is 69.2%. This is because the proposed approach further refined the event boundary by ADTA. In contrast, the event boundary by the other two approaches involve more non-exciting clip. However, the boundaries of some foul events could not be well determined by the proposed approach, which is the main reason why the BDA do not reach higher accuracy. In soccer video, the foul events usually happen suddenly, and without area restricted. The shots used to describe foul events are mainly the medium view shot and close-up shot. So, it is not easy to detect the boundary of foul events.

Table 5. Comparison of BDA.

Num	ADTA (%)	ETTP (%)	PB (%)
1	89.0	77.8	70.6
2	86.7	81.6	69.3
3	88.0	79.3	77.7
4	84.9	77.0	78.8
5	85.9	80.5	71.0
6	88.8	85.9	60.3
7	83.1	74.4	70.3
8	88.6	77.0	62.2
9	87.8	76.7	68.1
10	85.1	80.4	62.5
11	86.8	80.3	70.1
Total	**86.8**	**79.2**	**69.2**

The final video annotation result is stored in the XML format, which facilitates the semantic based event retrieval. An Excerpt of the annotation result is shown in Fig. 6. The XML file is organized as a hierarchical structure, each pair of <highlights> tag records the detailed information of an event, including its event type, time, start and end boundaries, involved player(s), players' team and the free text description.

```
<Highlights num="10">
  <EventType>射门</EventType>
  <Time>30</Time>
  <StartFrame>49268</StartFrame>
  <EndFrame>50389</EndFrame>
  <InvolvedPlayer team="米内罗竞技">若苏埃</InvolvedPlayer>
  <FreeAnnotation>第30分钟，若苏埃的远射打高。</FreeAnnotation>
</Highlights>
<Highlights num="11">
  <EventType>黄牌</EventType>
  <Time>32</Time>
  <StartFrame>54540</StartFrame>
  <EndFrame>55447</EndFrame>
  <InvolvedPlayer team="广州恒大">孙祥</InvolvedPlayer>
  <FreeAnnotation>第32分钟，孙祥身后放铲罗查被黄牌警告。</FreeAnnotation>
</Highlights>
<Highlights num="12">
  <EventType>任意球</EventType>
  <Time>37</Time>
  <StartFrame>59080</StartFrame>
  <EndFrame>60427</EndFrame>
  <InvolvedPlayer team="米内罗竞技">罗纳尔迪尼奥</InvolvedPlayer>
  <FreeAnnotation>第37分钟，罗纳尔迪尼奥30米外直接任意球射门，李帅侧扑虽然脱手，但很快将球压下。</FreeAnnotation>
</Highlights>
<Highlights num="13">
  <EventType>射门</EventType>
  <Time>41</Time>
  <StartFrame>66730</StartFrame>
  <EndFrame>67432</EndFrame>
  <InvolvedPlayer team="米内罗竞技">维克托</InvolvedPlayer>
  <InvolvedPlayer team="广州恒大">穆里奇</InvolvedPlayer>
  <FreeAnnotation>第41分钟，穆里奇禁区右侧受到雷维尔干扰后小角度射门被维克托挡出底线。</FreeAnnotation>
</Highlights>
```

Figure 6. An Excerpt of the video event annotation result.

6. CONCLUSIONS

The work in this paper could compensate the existing external information sources based soccer video annotation method for lacking of generality of web-casting text and its limitations in need of accurate text and video time information. However, we only analyze the Chinese match report of soccer video in this work. In the future, we plan to apply our approach to non-Chinese external text information, and extend the approach to more sports videos. Besides, the personalized video event retrieval and summarization are the next work based on the semantic-rich annotations. The theory of social network analysis may benefit for further research.

7. ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61173114, and the Wuhan Application Foundation Research Project under Grant No. 2014010101010027.

8. REFERENCES

[1] Oskouie P, Alipour S, Eftekhari-Moghadam A M. Multimodal feature extraction and fusion for semantic mining of soccer video: a survey. *Artificial Intelligence Review*, 2012: 1-38.

[2] Boyar M, Alan O, Akpinar S, et al. Event boundary detection using audio-visual features and web-casting texts with imprecise time information. In: *proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2010: 578-583.

[3] Xu C, Wang J, Lu H, et al. A novel framework for semantic annotation and personalized retrieval of sports video. *IEEE Transactions on Multimedia*, 2008, 10(3): 421-436.

[4] Xu C, Zhang Y F, Zhu G, et al. Using webcast text for semantic event detection in broadcast sports video. *IEEE Transactions on Multimedia*, 2008, 10(7): 1342-1355.

[5] Kolekar M H. Bayesian belief network based broadcast sports video indexing. *Multimedia Tools and Applications*, 2011, 54(1): 27-54.

[6] D. W. Tjondronegoro and Y. P. P. Chen. Knowledge-discounted event detection in sports video. *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans*, 40(5):1009-1024, 2010.

[7] Aydin S, Karsligil M E. An evaluation of possession information in playfield zones from soccer video using mid-level descriptors. In: *proceedings of IEEE 10th Workshop on Multimedia Signal Processing*, 2008: 680-684.

[8] Halin A A, Rajeswari M, Abbasnejad M. Soccer event detection via collaborative multimodal feature analysis and candidate ranking. *The International Arab Journal of Information Technology*, 2013, 10(5):1-9.

[9] Wang Z, Yu J, He Y, et al. Affection arousal based highlight extraction for soccer video. *Multimedia Tools and Applications*, doi: 10.1007/s11042-013-1619-1.

[10] Qian X, Wang H, Liu G, et al. HMM based soccer video event detection using enhanced mid-level semantic. *Multimedia Tools and Applications*, 2012, 60(1): 233-255.

[11] Landauer T K, Foltz P W, Laham D. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 1998, 25: 259-284.

[12] M. Xu, N. C. Maddage, et al., Creating audio keywords for event detection in soccer vdieo. In *proceedings of International Conference on Multimedia Expo (ICME)*, 281-284, 2003.

[13] M. Xu, C. S. Xu, et al., Audio keywords generation for sports video analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 4(2):11-23, 2008.

[14] Huang Q, Hu J, Hu W, et al. A reliable logo and replay detector for sports video. In: *proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2007: 1695-1698.

[15] David A. S. and Noel E. O. Event detection in field sports video using audio-visual features and a support vector machine. *IEEE Transactions on Circuits and Systems for Video Technology*, 2005, 15(10):1225-1233.