SEMANTIC EVENT EXTRACTION FROM BASKETBALL GAMES USING MULTI-MODAL ANALYSIS

Yifan Zhang[1], Changsheng Xu[2], Yong Rui[3], Jinqiao Wang[1], Hanqing Lu[1]

[1]National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100080, China
{yfzhang, jqwang, luhq}@nlpr.ia.ac.cn,
[2]Institute for Infocomm Research, Singapore 119613
xucs@i2r.a-star.edu.sg
[3]Microsoft China R&D Group, Beijing, 100080, China
yongrui@microsoft.com

ABSTRACT

In this paper, we present a novel multi-modal framework for semantic event extraction from basketball games based on web-casting text and broadcast video. We propose novel approaches to text analysis for event detection and semantics extraction, video analysis for event structure modeling and event moment detection, and text/video alignment for event boundary detection in the video. Compared with existing approaches to event detection in sports video which rely heavily on low-level features directly extracted from video itself, our approach aims to bridge the semantic gap between low-level features and high-level events and facilitates personalization of the sports video. Promising results are reported on real-world video clips by using text analysis, video analysis and text/video alignment.

1. INTRODUCTION

With the fast development of multimedia techniques, an explosive proliferation of sports video is made available on broadcast and Internet. The wide variety of services via new media channels such as network TV and mobile devices has shown tremendous commercial potential of sports video, and brought huge needs of personalized sports video according to consumers' preferences. The traditional one-to-many broadcast mode cannot meet different audiences' demands. For example, while one of Yao Ming's fans is interested in seeing the whole game, another one may only want to watch his slam dunks and blocks. In order to generate personalized summary for sports video, we need to extract event semantics (e.g. event type, how the event is developed, and players involved in the event, etc.) and use them to annotate and index events in the video. However, most existing approaches on event detection rely heavily on low-level audio/visual features directly extracted from the video itself. Due to the semantic gap between low-level features and high-level events as well as dynamic structures of different sports games, it is a challenging task to detect events and extract event semantics with high accuracy.

In this paper, we exploit a new external resource, web-casting texts, to help detect events and extract semantics from basketball games. In our previous work, we developed a live event detection system for soccer games based on web-casting text and broadcast video [1]. In this paper, we significantly improve the approach to text analysis, video analysis and text/video alignment, and broaden the application to new application domain: basketball games. Compared with soccer games, basketball games are very different. First, basketball games have more events with faster transition of shots and scenes, which may lead to false positive and bias in event detection. Second, the time-stopping and unpredictable game break in basketball games make it more challenging for alignment between time-tags in web-casting text and event moment in broadcast video. Specifically, the improvements in this paper include:

1. propose an unsupervised clustering based method instead of pre-defined keywords to automatically detect event from web-casting text;
2. improve the game time recognition algorithm to recognize the break time in basketball games;
3. propose a statistical approach instead of finite state machine as in [1] to detect event boundary in the video.

The framework of our proposed solution contains four major parts: (1) web-casting text analysis, (2) broadcast video analysis, (3) text/video alignment, and (4) semantic annotation and indexing for personalized retrieval. We first use text clustering and tagging to generate semantic keywords automatically. Then we detect and formulate text events by these keywords. In the video analysis module, automatic shot boundary detection and shot classification are conducted and applied for event structure modeling; game time is further detected by recognizing the clock digit overlaid on the video. A statistical method based on Hidden Markov Model (HMM) is used to extract event boundary in the alignment between video and text. Finally, event segments are annotated and indexed by the keyword descriptors.

2. RELATED WORK

A number of approaches have been proposed for basketball video analysis, including shot classification, scene recognition and event detection. Tan et al. [2] used low-level information available directly from MPEG compressed video, combined with domain knowledge of basketball, to identify certain events and classify video into wide-angle and close-up shots. Nepal et al. [3] presented temporal models with audio/visual features to detect goal segments. Xu et al. [4] proposed an approach to generate audio keywords for basketball video based on low-level audio features and applied audio keywords together with heuristic rules to event detection. All

these approaches were based on low-level features extracted from the video itself, which may be difficult to extract semantic meanings from the video sources.

3. WEB-CASTING TEXT ANALYSIS

Web-casting text, which is the description of the game progress, is available on many sports websites [5][6]. Exciting or important events during the game are reported in the web-casting text with all key information such as time, players, teams, actions, etc. In our previous work [1], we used pre-defined keywords to match related words in the texts to detect event, which is less general to different sports domains. Based on our observations, the descriptions of the same events in the web-casting text have similar sentence structures and word usage (See Fig.1). We can therefore employ an unsupervised approach to firstly cluster the descriptions into different groups corresponding to certain events and then to extract keywords from the descriptions in each group for event detection.

09:23 Dwyane Wade makes 18-foot jumper.
05:10 Richard Hamilton makes 9-foot jumper.

10:43 Rasheed Wallace shooting foul (Shaquille O'Neal draws the foul)
03:54 Antoine Walker offensive foul (Tayshaun Prince draws the foul)

FIG.1 Examples of web-casting texts from ESPN

In this paper, we use Latent Semantic Analysis (LSA) [7] to cluster events into different groups. LSA is a technique widely used in natural language processing. It is assumed that there are underlying or latent structures in word usage corresponding to semantic meanings of documents. A term-document matrix is built to describe the occurrences of terms in different documents. This matrix is then analyzed by singular value decomposition (SVD) to derive the particular latent semantic structure model. Finally, the cosine distance between vectors is computed to measure the document similarity in the reduced dimensional space. Before we apply LSA, we first filter out the names of players and teams in the text by Name Entity Recognition (NER) due to the consideration of their affection to clustering result. For example, "*Rasheed Wallace shooting foul (Shaquille O'Neal draws the foul)*" is processed into "*shooting foul (draws the foul)*". Then we build the term-document matrix $A_{t\times d}$ by regarding each description as one document. The SVD projection is computed by decomposing the matrix $A_{t\times d}$ into the product of three matrices:

$$A_{t\times d} = T_{t\times n} \times S_{n\times n} \times D^T_{d\times n} \quad (1)$$

where *t* is the number of terms, *d* is the number of descriptions, *n* is the rank of A, *T* and *D* have orthonormal columns and *S* is a diagonal matrix. Only the first largest k (k < n) singular values and the corresponding columns from the *T* and *D* matrices are used to give the estimate matrix $\tilde{A}_{t\times k}$.

$$\tilde{A}_{t\times k} = T_{t\times k} \times S_{k\times k} \times D^T_{d\times k} \quad (2)$$

Each column in the reduced model $\tilde{A}_{t\times k}$ is the textual feature vector of each description in the text. Cosine distance of each column is applied to cluster them into different groups corresponding to different semantic events. In each semantic event group, the rule-based part of speech tagger is utilized to tag the words of descriptions and recognize nouns and verbs. Finally, we rank the nouns and verbs by their $tf \times idf$ weights, where tf is the term occurrence frequency in each group and idf is the inverse term occurrence frequency in the entire document corpus. Several top rank terms are set as the keywords of each group. Table 1 shows the ranking result of the terms after unsupervised clustering. K-means approach is employed to cluster all the descriptions into 9 groups, which have the smallest within-cluster distance sum. Here we list the top 4 rank terms in each group. The last column is the semantic events in basketball games corresponding to each group.

Table 1 Keywords generated from web-casting text

Group	Rank1	Rank2	Rank3	Rank4	Event
1	shot	point	foot	tip	Shot
2	jumper	foot	makes	assists	Jumper
3	layup	assists	makes	misses	Layup
4	dunk	slam	assists	makes	Dunk
5	blocks	jumper	shot	assists	Block
6	rebound	assists	ball	draws	Rebound
7	foul	draws	ball	loose	Foul
8	throw	makes	misses	ball	Free throw
9	enters	game	foul	ball	Substitution

It can be seen that these 9 groups contain most semantic events in the basketball game. Assist event does not have keywords because it is not an independent event but always occurs with other events (e.g. shot, jumper, dunk). Some events can be further classified into more specific sub-events, for example, jumper and shot can be classified into three-points and two-points. The rank1 words can be selected as keywords for each group to detect text events.

After proper keywords are identified, the text events can be detected by finding the item of description which contains the keywords and analyzing context information in the description. The time-tag of each detected event together with some semantic information (e.g. event type, player, team, etc.) are then recorded for the text/video alignment and event indexing and annotation tasks.

4. BROADCAST VIDEO ANALYSIS

With the result of text analysis, we can map the text event time-tag into the game video stream to detect event moment and event boundary in the video. Hence we need to analyze the video structure and recognize the game time.

4.1. Shot classification

During the sports game broadcasting, when an event occurs, the broadcast director always uses a temporal shot pattern to highlight the event. This is the so-called broadcast production rule, which is common to a variety of sports game broadcasting. For example, when an exciting event occurs, there is firstly a far-view shot followed by a close-up shot of players who are involved in the event, and then a replay to review the event from different camera angles, finally a far-view shot to resume the game. Hence our video analysis is based on semantic shot analysis. To limit the computing complexity and enhance the robustness, we classify the shot view type into three classes: (1) far-view, (2) close-up, (3) replay.

We use the mean absolute differences algorithm of consecutive frames for shot boundary detection [1]. With the obtained shot boundary, shot classification is conducted using a majority voting of frame view types. Due to the variety of the color and patterns of different basketball fields, the dominant color detection which is used in soccer games does not work robustly for basketball games. Here we calculate the edge pixel number of each frame to decide the frame view type in a basketball video. As shown in Fig.2, the number of edge pixels in a far-view frame is much larger than that

in a close-up frame, thus we can empirically set a threshold to divide the frames into two classes.

(a) (b) (c) (d)

FIG.2 Video frames and corresponding edge pixels (a) far-view frame (b) edge pixels of far-view frame (c) close-up frame (d) edge pixels of close-up frame

Replay detection is relying on the flying logo template matching technique in R, G, B channels [1]. The detected replay/non-replay state of each shot is denoted by value 1 and 0, respectively and collected as a shot sequence.

4.2. Game time recognition

Since the text event has the exact time tag when the event occurs, the intuitive way to find the counterpart in a video is to find the same time in the video and use this time as an event moment. However, in a broadcast sports video, the starting points of game time and video time are not synchronized due to non-game scenes such as player introduction, ceremony, half-time break. Hence, we need to detect the exact game time in the video and then detect the exact event boundary in the video based on the event moment. We proposed an approach [1] to first detect a clock overlaid on the video and then recognize the digits from the clock to detect the game time. This approach works well for soccer videos due to the non-stopping clock time in soccer videos. Once the game time is recognized at certain video frame, the time corresponding to other frames can be inferred based on recognized game time and frame rate. However, the clock in a basketball video may stop or disappear at any time of the game, which makes the game time recognition more challenging than a soccer video. Due to unpredictable clock stopping, game time recognition in each frame is necessary in a basketball video. As the clock region in the video may disappear during the clock stopping time, we employ a detection -verification-redetection mechanism, which is an improvement to our previous approach [1], to recognize game time in a basketball video.

We first use the static region detection to segment the static overlaid region. Then the temporal neighboring pattern similarity measure is utilized to locate the clock digit position because the pixels of clock digit region are changing periodically. The technique detail of digit region location and digit number template capturing is referred to [1]. Once we get the template and clock digit position, we need to verify the matching result by the following formula:

$$M(i) = max_i\{\sum_{(x,y)\in R} D(x,y) \odot T_i(x,y)\} \qquad (3)$$

$$i = 0,1,2\ldots,9,10$$

where $T_i(x,y)$ is the image pixel value in position *(x,y)* for the template of digit number *i*, $D(x,y)$ is the image pixel value in position *(x,y)* for the digit number to be recognized, *R* is the region of digit number, $\odot$ is EQV operator, i =10 corresponds to the region without any digit. $M(i)$ is calculated in the SECOND digit of the clock position in each frame. If $M(i)$ is smaller than a threshold which is empirically set during a certain time (2 seconds in our work), a failure of verification will occur and lead to the redetection in the ROI. After the successful verification, the game time will be recorded as a time sequence corresponding to the frame number, which is called time-frame index. During the clock stopping time with the verification failure, frames are also denoted as the label of "- - : - -" in time-frame index. Since some specific events (e.g. free throw and substitution) only happen during the clock-stopping time, the clock-on and clock-stopping information is also a semantic feature which will be used in the text/video alignment task.

5. TEXT-VIDEO ALIGNMENT

Mapping the text events detected in Section 3 into the game video stream is an important module in our framework, as it needs to decide the exact event boundary for the event segment generation. Since we have parsed the game video into basic shot units with the label of view types, we only need to detect the start and end shots of the events. Our previous method [1] for event boundary detection used a FSM to model event structure, which is less general to various sports domains. Here we use a hidden Markov model (HMM) to model the temporal event structure and detect the event boundaries. We first train two HMMs using labeled shot sequence to obtain the parameters for each HMM of clock-on event and clock-stopping event. The features used to train HMMs are described in section 4. After training, the HMMs can be used to detect the boundaries of different event types.

During the detection process, the shot containing the detected event moment is used as the reference shot to obtain a search range by a rule for event boundary detection. For clock-on events, the search range is empirically set to start from the first far-view shot before the reference shot, and end at the first far-view shot after the reference shot. For clock-stopping events, the time-tag in the text is denoted just before the moment when the clock stops, while the specific event will occur on any time during the clock-stopping time. Hence, the search range is set to all the clock-stopping shots after the time-tag point (See Fig.3).

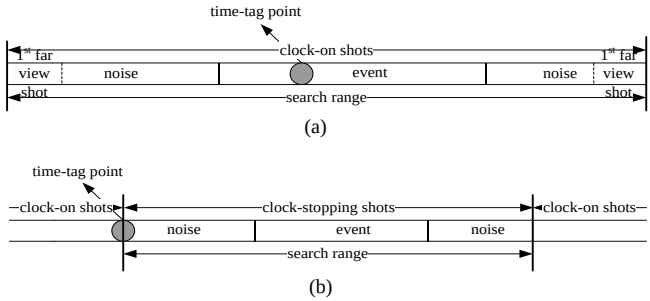

FIG.3 Search range for (a) clock-on event (b) clock-stopping event

Within the search range, the shots which are not much relevant to the event are regarded as noise and cannot be included in the event. We set all the possible partitions of shot sequences in the search range as candidates and send them to the trained HMM to calculate the probability scores. The partition which generates the highest probability score is selected as the detected event segment (Eq.(4)) and the event boundaries can be obtained from the boundaries of the first and last shot in this event.

$$P(S|Hmm) = max_i\, P(S_i|Hmm) \qquad (4)$$

where S_i is the possible partition of shot sequence in the search range. Fig.4 shows the diagram of our event boundary detection.

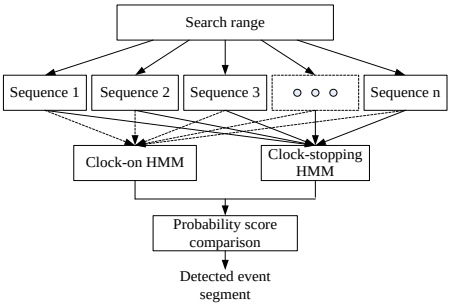

FIG.4 The diagram of our event boundary detection

6. SEMANTIC ANNOTATION AND INDEXING

Once the event segments from the original video stream are generated, they each are annotated with the corresponding descriptions in the web-casting text. Since each item of the event description in the web-casting text is tagged by the rule-based part of speech tagger, we choose the nouns and verbs to annotate the event segments. At the same time, a log file attached to the segment is also to log the item of the event description in the web-casting text which contains the information of time, event type, player and team. After that, the indexing of event segments is built in the database for personalized retrieval. In our approach, the personalized retrieval can be conducted based on the time, the event type, the player and the team according to the user's interests.

7. EXPERIMENTAL RESULTS

We conducted our experiments on 5 real-world NBA 2005~2006 games. The web-casting texts are collected from the ESPN website [5] and the videos are recorded from TV using Hauppauge PCI-150 TV capture card. For text analysis, all the descriptions of the web-casting text in 5 games are manually divided into 9 groups (shot, jumper, lay-up, dunk, block, rebound, foul, free throw and substitution) and used as the ground truth. We then use the rank1 terms in Table1 as the keywords of each group to detect text events. Table 2 lists the event detection performance from web-casting texts. Shot and jumper events have the relatively low precision because some block events whose descriptions also have the term "shot" and "jumper" are misclassified in these two groups.

Table 2 Text event detection

Event	Precision/ Recall	Event	Precision/ Recall
Shot	87.8%/98.1%	Rebound	98.5%/99.2%
Jumper	89.3%/100%	Foul	100%/98.7%
Lay-up	97.7%/99.0%	Free throw	100%/100%
Dunk	96.7%/100%	Substitution	95.0%/100%
Block	98.5%/99.7%		

A total 90 minutes of the videos from 3 games in our dataset are used for testing shot classification and replay detection in our video analysis module. The result is shown in Table 3.

Table 3 Shot classification

Class	Far view	Close-up	Replay
Precision	91.3%	97.3%	94.7%
Recall	97.1%	89.9%	92.6%

We use the Boundary Detection Accuracy (BDA) [1] to measure the detected event boundary compared with the manually labeled boundary:

$$BDA = \frac{\tau_{db} \cap \tau_{mb}}{\tau_{db} \cup \tau_{mb}} \quad (5)$$

where τ_{db} and τ_{mb} are the automatically detected event boundary and the manually labeled event boundary respectively. 150 minutes from the 5 games are used for training the HMMs, while the rest are for the test. Table 4 lists the BDA scores for the 5 games. It is observed that the BDA scores of foul and substitution are relatively low, because the foul's time-tags from web-casting texts are always not as accurate as others, while the substitution is not a compact event but has loose structures and various temporal transition patterns.

Table 4 Event boundary detection

Event	BDA	Event	BDA
Shot	92.3%	Rebound	90.5%
Jumper	93.5%	Foul	69.8%
Lay-up	91.6%	Free throw	83.0%
Dunk	89.7%	Substitution	63.7%
Block	90.1%		

8. CONCLUSION

In this paper, a novel multi-modal framework for semantic event extraction from broadcast basketball videos has been proposed. Compared with our previous work, we have broadened the application domain and made the approach more general to text analysis, video analysis and text/video alignment. The incorporation of web-casting text into sports video analysis improves the accuracy of semantic event detection. Event segments are automatically annotated and indexed by semantic information to support personalized retrieval. Our future work will investigate more advanced techniques on personalized video presentation to provide more flexible view-ship to meet viewers' various preferences.

9. ACKNOWLEDGEMENT

The research is supported by the 863 Program of China (Grant No. 2006AA01Z315, 2006AA01Z117), NNSF of China (Grant No. 60475010) and NSF of Beijing (Grant No. 4072025).

10. REFERENCES

[1] C. Xu, J. Wang, K.Wan, et al, "Live Sports Event Detection Based on Broadcast Video and Web-casting Text", In Proc. ACM Multimedia, Santa Barbara, USA, pp.221-230, 2006.

[2] Y. P. Tan, D. D. Saur, S. R. Kulkami, et al, "Rapid estimation of camera motion from compressed video with application to video annotation", IEEE Trans. CSVT, vol. CSVT-10, pp.133-146, 2000.

[3] S. Nepal, U. Srinivasan and G. Reynolds, "Automatic detection of goal segments in basketball videos", In Proc. ACM Multimedia, Ottawa, Canada, pp.261-269, 2001.

[4] M. Xu, L. Duan, J. Cai, L. Chia, C. Xu and Q. Tian, "HMM-Based Audio Keyword Generation", In Proc. PCM, Tokyo, Japan, pp.566-574, 2004.

[5] http://sports.espn.go.com/nba/scoreboard

[6] http://www.nba.com

[7] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas and R. Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, vol. 41, iss. 6, pp.391-407. 1990.