International Journal of u- and e- Service, Science and Technology
Vol.8, No. 11 (2015), pp.75-84
http://dx.doi.org/10.14257/ijunesst.2015.8.11.08

Research on Sports Video Analysis Based on Rules

Hui Li[1], Linghua Ran[2] and Yang Wang[1,*]

[1] *Physical Education Department, Harbin Engineering University, Harbin 150000, China*
[2]*Physical Education College, Harbin Normal University, Harbin 150001, China*
aoled@126.com

Abstract

Motivated by principles of natural language processing, we talk about rule-based methods for sports video analysis and propose a video parsing system by using grammar. Firstly, we segments video into elementary shots. Then through event detection, shots are annotated with semantic labels to form a token sequence. Finally, the sequence is parsed and validated with grammar to construct a table of content for the video. Different from other systems in the literature, our system is not only able to annotate semantic events, but also able to recognize hierarchical game structures of sports videos. Thus, it satisfies user's requirements better.

Keywords*: Video Analysis, Video Retrieval, Sports Video, Rule-Based*

1. Introduction

By automatic extraction of highlights in sports videos, it's probable to get initial information, such as splitting video contents to highlight part and not wonderful part. But it's not possible to analyze the structure of sport competition and related events, like dividing diving competition to several rounds to recognize events such as players' appearance, jumps. Here using for reference the principle and idea of natural language processing, we take rule-based method to discuss general sematic events in sports videos as well as their interrelationship.

Most of the existing analysis methods for sport videos are centered on semantic annotations such as shot classification, highlight retrieval, event detection etc. Despite some documents mentioned structural analysis of sport videos, such work refers to largely the breakdown of basic scenes. As indicated in [1-2], soccer videos include ongoing and suspended part; in [3-4], the concern is detecting serving scene of tennis match; [5-6] raised a universal algorithm with no need of domain model [7-8]. It utilized the clustering algorithm based on time constraint to create a three-tier structure of the video: shot, lens and scene. However, in sport videos, different competition is not of similar structure. Such uniform structure is not appropriate for sport videos [9].

In the paper, we propose a grammar-based strategy to analyze sport videos. It combines the function of semantic annotation and structural analysis to automatically generate videos' index and catalog. Unlike other video analyzing systems here, the proposed system is capable to discern hierarchical framework of sports match, easy for users to browse video contents. Moreover, with such structure expressed as grammar, it can be used for error detection and recovery in the video analyzing process [10-11].

Normally, the natural language understanding lexical treated as basic object, including word segmentation, part of speech tagging, parsing and several stage. Similarly, our sports video analysis system includes three parts: shot boundary detection, semantic mark tagging and syntax analysis. Firstly, the original video stream segmentation is lens. A lens is a

* Corresponding Author

ISSN: 2005-4246 IJUNESST

period of uninterrupted sequence of frames taken by the same camera. Since the lens is the smallest unit of expression has the complete contents of the video, we will be shot as the basic unit in sports video analysis, similar to the words of a language. Secondly, semantic annotation for each lens. The objective is from sports video to automatically detect the field related events, and mark the events of lens. Finally, we use Context-Free Grammar (CFG) to describe the structure of sports competition.

Not only can be established tree structure of video content based on grammar for syntax analysis, and can handle the error annotation caused the error. The experimental results show that our video analysis system is effective.

Although in this paper we mainly in diving video as an example for study, but the methods can be extended to other similar in sports video analysis.

2. Sport Video Analysis Based on Grammars

From other videos like news, movies, sport videos have two unique features. The uniqueness represents the domain knowledge of sports video analysis. The first feature is there're domain-related events recurring in those videos. For audiences, such events are often the most important and meaningful fragments there. We can classify them into three types: playback events, state events and target events, as shown in Figure 1. Playback events refer to slow-motion replayed snippets which are alternately played in the sportscast. They're indicative of fabulous moments in which audiences show interest. State events happen when the competition state has changed, *e.g.*, scoring of diving at the end of each round; the start shot of one set in the tennis ace. The detection of state events is very significant to the analysis of video structure. Target events are some rather enjoyable specific movements in sport matches, such as players' jumps in diving competition, goals in soccer game. They're usually manifestations of kinematic relation of objects and among them.

(a) Playback Events (b) State Events (c) Target Events

Figure 1. Sports Video Event

Video contents are of a variety. It's so hard to develop a universal event detection method to remove barriers between low-level features and high-level semantics. Hence we choose to talk about the application of domain knowledge, including competition-related knowledge and video production knowledge. From sports videos, we observe that:

(a)In most sport videos, in order to remind viewers, there are specific shot cuts as seen in the above picture, before and after playing back events;

(b)State events are captioned to show states of the competition;

(c)In target events, object and camera movements are noticeable; meanwhile, there are audiences' cheers and other related sounds like jumping sound in the diving and hitting sound in baseball match.

Another characteristic of sports video has a tree structure. For example, a game of tennis can be divided into a plurality of set, a set also includes a number of games. Each game hits several ball back and forth. It shown in Figure 2. In order to facilitate browsing of video content, original video data should be analytical and organized into hierarchical directory according to these structures.

Based on the above analysis, we think of sports video processing can be similar to the language processing based on dictionary and grammar. In sports video, dictionary is our set of events, while the grammar is the tree structure of the set of rules. In diving video as an example, we design a sports video analysis system.

The goal of the system analyzed sports video content to establish the index and tree based on event. Through these indexes and directories, users can easily find what they really want and sense of video content of interest. Similar to the language processing, our system consists of three stages: shot detection, semantic tagging and syntax analysis.

2.1. Semantic Annotation

Semantic annotation, in its essence, is to classify shots according to predefined event model. Events have three types: playback events, state events and target events. Next we'll introduce algorithms for identifying those events in diving competition videos.

2.1.1. Detection of Playback Events. Special marks appear in shot cuts before and after playback events. As per the characteristic, we propose a single playback event detection algorithm. It has the following steps:

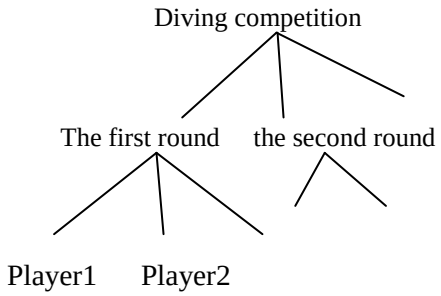


Figure 2. The Tree Structure of Diving Competition

In the region R where tokens appear, calculate the distance d between the frame image f in the shot boundary and sample image $\bar{f}$ in the mark template;

$$d = \sum_{R} | f(x, y) - \bar{f}(x, y) | \quad (1)$$

If the gap between two shot boundaries which both have a playback token is shorter than the maximal time span, it means having detected playback events. All shot cuts in the two boundaries should be marked with the same token. By equation (1), we calculate the distance between the frame image and the chosen image, selecting the minimal distance as the result for return.

2.1.2. Detection of State Events. State events are often accompanied with subtitles to show states of the match. In diving competition, there are three state events: "Take your mark", "Scoring" and "End of each round". "Take your mark" refers to events of making

preparations on the diving platform or board. Right at the moment, there're captions regarding diver's name, current ranking and movement difficulty. And then, the diver jumps into the water. When the diver swims out of the pool, captions show again regarding referees' grades. This is "Scoring" event. "End of each round" involves events when it's over. Captions are showing a table of integrals, which includes the diver's current scores and placement.

We note that in different state events, subtitles show in different places and key words are changing. Based on that, we put forward the method for recognizing state events through subtitle match. It firstly detects out character blocks in all video frames, then according to those blocks it computes the similarity degree between video frames and sample pictures in the template of state events.

Set $F = (f_1, ..., f_n)$ and $G = (g_1, ..., g_n)$ respectively represents a collection of text block of video frames and sample picture. $|f|, |g|$ said the number of pixels in the block of text. Similarity calculation is given by:

$$s(F, G) = \frac{\sum_{f \in F} \sum_{g \in G} \tau(f, g) |f \cap g|}{\max(\sum_{f \in F} |f|, \sum_{g \in G} |g|} \quad (2)$$

2.2. Grammatical Analysis

Introducing grammatical analysis to sports video analysis will bring about three benefits:

(1) We can build effectively tree view architecture of videos with compiler technology;
(2) Describing competition-related domain knowledge to grammars is helpful to separate domain knowledge from specific parsing algorithm, making the system with better generalization;
(3) By using grammatical analysis, we introduce error processing mechanism, which reports errors occurring in the automatic analysis course, enhancing the system's friendliness and usability.

Table 1. Mark in the Diving Video

Mark	Category	Semantic
r	Playback events	Replay of fragment
b	State events	The athletes ready
s	State events	Score
e	State events	The end of round
d	target events	Diving
u	Un-definition	Other

After event detection and semantic annotation, sport video stream is converted to token sequence, as seen in the Table1. After that, it's needed to parse grammatically the token sequence in order to fetch its structure and construct user-friendly hierarchical table. The method used here follows the same process as the compiler builds parsing tree from input entries according to grammars. We make use of context-free grammars to describe internal rules of sports competition.

We take this as example. Figure 2 shows tree structure of one diving competition, which can be expressed in the following grammatical form:

$$\begin{aligned} S &\rightarrow R \mid RS \\ R &\rightarrow Pe \mid PR \\ P &\rightarrow bdrs \end{aligned} \tag{3}$$

Where, S is grammar symbol, R represents consists of several P, P represents each player game.

3. Experimental Analysis and Results

In order to validate analysis method based on grammar video, we implemented a diving video content analysis system. The system used Java language and Java Media Framework API in Win7, running on the Pentium IV 3.0GHz computer. Figure 3 is scene of the system operating.

When you open a diving video and analyze it with automatic analysis tools, you will get index and table on the left side of the screen. The table is of hierarchical structure, where each node is a shot with a token and start frame. Entries in the index are tokens and start frame images of every single event. You can search interesting videos or preview the whole video by using key frame images and tokens. By clicking key frames or tokens, you'll locate the video fragment, instead of pressing "Fast Forward" or "Fast Backward" for the same function. Besides, the system provides a great interactive environment. After automatic analysis, you can modify relevant analytical structure based on reported errors.

To validate the system performance, we used a great number of video data to test it. The video lasts totally over four hours. Table2 lists the data set for testing. All videos are MPEG-1 format, 352×288 DPI, FPS at 25 frames per second. These videos were extracted from different diving competitions in different venues, referring to (A) Men's Diving Synchronized 3m Springboard; (B) Women's Diving Synchronized 10m Platform; (C) Women's Diving 3m Springboard; (D) Men's Diving 10m Springboard. All related events in in the data set were marked by professionals before the test as for the real reference standard.

Table 2. The Test Data Set

Event type	Length of time	Replay events	State events	Target events	Total
A	0:45:56	40	24	40	164
B	0:44:16	40	25	40	165
C	1:34:24	60	125	50	245
D	1:25:44	72	150	62	294
Total	4:15:18	212	444	212	868

The system we developed has three modules: shot segmentation, semantic annotation and grammatical analysis. In specific implementation, we evaluate mainly algorithms used for respectively semantic annotation and grammatical analysis. The experiment includes two parts:

(1) Assessing the algorithm for semantic annotation, i.e. performance of event detection;

(2) Verifying the system performance in the grammatical analysis period, including recognition rate of high-level structural units and effectiveness of error reporting.

3.1. Experimental Results of Semantic Annotation

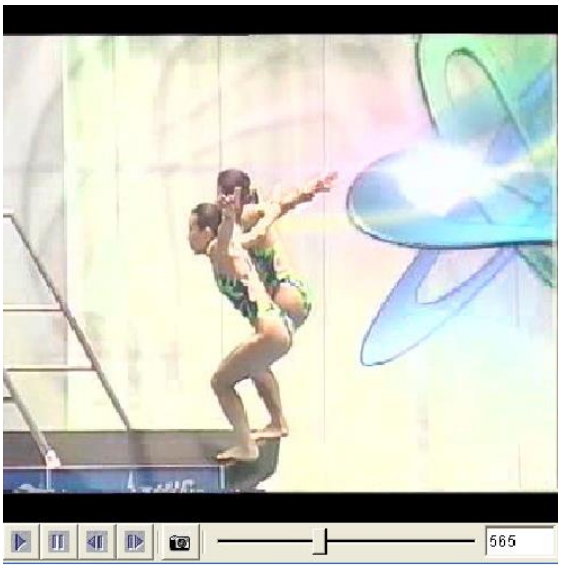
Table 3-5 presents recall ratio and precision rate of detection of various events. Recall ratio and precision rate can be defined in the following two equations:

$$\text{Precision} = \frac{\text{The event number of correct detection}}{\text{Detection of the total number of event}} \quad (4)$$

$$\text{Recall} = \frac{\text{The event number of correct detection event}}{\text{The event total number of data concentration}} \quad (5)$$

In it, one event being detected correctly means over a half shots in the event are marked accurately. From Table 3-5, we see the system performs well for the detection of playback events and state events, but unsatisfactory detection of target events. The reason we think rests with too changeable movements in shots, which had impacts on the detection of target events. For improvement, it's necessity to develop better feature extraction technologies and more powerful statistic recognition model. That will be our next research objective.

(a) Content Directory Includes all Events and Hierarchy



(b) The Content Index includes only Diving Events and Replay

Figure 3. Diving Video Content Analysis System

Table 3. Replay Events Experimental Results of Semantic Annotation

Event type	Precision	Recall
A	40/40=100%	40/40=100%
B	40/40=100%	40/40=100%
C	60/60=100%	60/60=100%
D	71/73=97%	71/72=99%
Total	99%	100%

Table 4. State Events Experimental Results of Semantic Annotation

Event type	Precision	Recall
A	78/78=100%	78/84=93%
B	78/78=100%	78/85=92%
C	114/116=99%	114/125=91%
D	118/119=99%	118/150=79%
Total	99%	87%

Table 5. Target Events Experimental Results of Semantic Annotation

Event type	Precision	Recall
A	30/43=70%	30/40=75%
B	25/34=74%	25/40=63%
C	58/71=82%	58/60=97%
D	58/84=69%	58/72=81%
Total	74%	81%

3.2. Experimental Results of Grammatical Analysis

The grammatical analysis module is, on one hand to analyze hierarchical content structure of videos as per token sequence; on the other hand to report timely errors found during the analysis probably because of wrong tokens in the sequence. For the two functions, we have made evaluations.

Table6 gives the evaluation about error processing in grammatical analysis. The adopted immediate recovery method detected errors effectively. But when the method finds errors, it needs to wait till the next synchronous token appears before recovering the analysis. In that case, some errors detected in the period may be overlooked. In our current system, 62% of errors can be automatically reported. Giving that the method here is quite simple, the experimental results are encouraging.

Table 6. Error Handling in Syntax Analysis

Event type	Total number of errors	Report number of errors
A	29	22
B	28	18
C	28	22
D	80	40
Total	165	102

In diving video experiment, high-rise structure unit includes players unit and each round unit. It shown in Figure 2. Players unit defined as Video clip from the player prepare to score event. Each round of the competition unit defined as video clip of the end of each round.

The performance of recognition of high-level unit based on the user's subjective evaluation algorithm in [12], this makes the experimental results are not objective and convincing. In contrast, the paper is proposed the high-level unit definition and evaluation are clear and objective, the experimental results are more accurate.

Table 7 can show our algorithm rarely formed error identification, but relatively easy to lose some structural unit. So, based on the analysis of grammar, a high-rise structure unit is defined by the relationship of composition of the unit event. Only to identify one of the event cannot judge the events, and meet the conditions of the relationship. In this analysis, there will be some high level unit was lost, it needs to study more powerful error method to solve this problem.

Table 7. Identification Results of High-rise Structure Unit

Event type	Number of lens	Players game unit			Each round of the game unit		
		Correct	Missing	Error	Correct	Missing	Error
A	356	34	6	0	4	0	0
B	448	34	6	0	5	0	0
C	673	49	11	0	5	0	0
D	850	50	22	0	6	0	0
Total	2327	167	45	0	20	0	0

4. Conclusion

In this paper, by Natural Language Processing, we propose a new method for the analysis of sports video based on grammar. It can the comprehensive semantic annotation and analysis of the structure, and automatic generates sports video indexing and directory.

Firstly, the original video stream separated lens, and got shot sequence. Secondly, semantic event related field carried out to test for each lens, and gave semantic mark of lens, and formed semantic indexing. Finally, for mark sequence carried out syntax analysis based on Context-Free Grammar, and Generated the context generation of sports video.

Acknowledgment

This work was supported by The Fundamental Research Funds for the Central Universities. No. HEUCF151601.

References

[1] J. Assfalg, M. Bertini, C. Colombo, and A.D. Bimbo, “Semantic Annotation of Sports Videos,” IEEE Multimedia, vol. 9, no. 2, **(2012)**, pp. 78-82.
[2] Y. F Ma, and H. J. Zhang, “A Model of Motion Attention for Video Skimming”, Proceeding of IEEE International Conference on Image Processing, Rochester, New York, September, **(2008)**.
[3] A. Hanjalic, “Generic Approach to Highlights Extraction from a Sport Video,” Proceedings of IEEE International Conference on Image Processing, **(2013)**.
[4] L.Y. Duan, M. Xu, T.S. Chua, Q. Tian, and C.S. Xu, “A Mid-level Representation Framework for Semantic Sports Video Analysis,”Proceedings of ACM International Conference on Multimedia, November **(2013)**.
[5] N. Babaguchi, Y. Kawai, and T. Kitahashi, “Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration”, IEEE Transactions on Multimedia, vol. 4, no. 1, March **(2012),** pp. 34-38.
[6] L.Y. Duan, M. Xu, X.D. Yu, and Q. Tian, “A Unified Framework for Semantic Shot Classification in Sports Videos”, Proceedings of ACM International Conference on Multimedia, **(2008)**, pp. 419-420.
[7] L. Xie, S.F. Chang, A. Divakaran and H. Sun, “Structure Analysis of Soccer Video with Hidden Markov Models,” Proceedings of International Conference on Acoustic, Speech, and Signal Processing, May **(2012)**.
[8] D. Zhong and S.F. Chang, “Structure Analysis of Sports Video Using Domain Models,” Proceedings of IEEE International Conference on Multimedia and Expo, August **(2009).**

[9] .Y. Rui, T.S. Huang, and S. Mehrotra, "Constructing Table-of-Content for Videos", ACM Journal of Multimedia Systems, vol. 7, no. 5, **1999.**
[10] R. Lienhart, "Comparison of Automatic Shot Boundary Detection Algorithm", Proceedings of SPIE Storage and Retrieval for Image and Video Databases VII, **1999**.
[11] F. Wang, J.T. Li, Y.D. Zhang, and S.X. Lin, "Automatically Extracting Highlights for Diving Video," Proceedings of China National Computer Conference, vol. 1, November, **(2013)** pp. 471-475.
[12] Y. Rui, T.S. Huang, and S. Mehrotra, "Constructing Table-of-Content for Videos", ACM Journal of Multimedia Systems, vol. 7, no. 5, **(1999)**.

Author

Hui Li, he got his B.S degree from Harbin Institute of Physical Education, and got his M.S degree from Harbin Engineering University. He is a lecturer in Harbin Engineering University. His research interests include Sport Economics.