AUTOMATIC PARSING OF AMERICAN FOOTBALL VIDEOS BY INTERMODAL COLLABORATION BASED ON TRANSITION RULES

Naoko Nitta and Noboru Babaguchi

Department of Communication Engineering, Osaka University,
2-1 Yamada-Oka, Suita, Osaka 565-0871, Japan
E-mail: {naoko,babaguchi}@comm.eng.osaka-u.ac.jp

ABSTRACT

This paper proposes an automatic American football video parsing method based on transition rules of an American football game. Combining the results of live scene extraction and superimposed text detection based on image features enables us to segment the video into play units of a game. Temporally associating the segmented play units with the detected superimposed texts and the closed-caption text attaches possible semantic content information to the play units. Finallly, selecting only the play units which comform to transition rules of the sports game from the obtained play unit sequence, while discarding or complementing unnecessary or insufficient play units and attached semantic content information, realizes the semantic video parsing.

1. INTRODUCTION

Effective handling of videos such as browsing, retreiving, and editing requires semantic understanding of the videos. Semantic video content analysis is quite a challenging problem due to a large variety of video content. However, there are some similarities among certain types of videos, which can be cues to solve the problem. For instance, a news video can be considered as a sequence of video segments which start with an introductory anchor talk followed by details of news [3]; a cooking video as a sequence of steps in step-by-step instructions of recipes [2]; and a sports video as a repetition of play and break scenes [4, 5, 6]. As stated above, a video is often considered as a sequence of video segments, each of which can be considered as a unit to understand the semantic content or the story of the video. Therefore, structuring videos according to their semantic compositions, while understanding the semantic role of each video segment, is a step toward semantic understanding of the videos.

There has been some prior work aiming at video content analysis focusing on video structures. Miura et al. [2] tried to segment a cooking video with image analysis and associate video segments with preparation steps by keyword matching. Marlino et al. [3] also tried to segment a news video into story segments with Finite State Automata with time transitions based on cues in the image, audio, and text streams. Zhong et al. [4] focused on the content structure of a sports video and tried to extract some patterned event boundaries from the image stream. As you can see, videos are often analyzed using several multimodal information streams. We call this method intermodal collaboration [1]. However, semantic synchronization between the video content and the content information obtainable from these streams is often hard to accomplish, especially for sports videos which have similar content throughout a whole video. In this paper, we focus on American football videos as a case study of sports videos and propose a method to segment a video into story units and structure the story unit sequence based on rules how the video should be composed. We call this **Video Parsing**.

Here, we use three different information streams. The image stream, the graphics stream called the **superimposed text**, which generally appears in sports videos to display situations of a game, and the text stream called the **closed-caption text**, which is a transcript of the audio stream including announcers' commentaries. Our proposed method first segments the video stream into **play units** by combining the results of live scene extraction and superimposed text detection by template matching using image color features. Next, temporaly associating semantic information extracted from the detected superimposed texts and the closed-caption text to the obtained play units provides a **play unit sequence** with possible semantic information. Parsing the play unit sequence based on transition rules of an American football game keeps only the play units with correct information by discarding or complementing unnecessary or insufficient play units and semantic information associated with play units.

2. TRANSITION RULES FOR AMERICAN FOOTBALL GAME

A TV sports game program can be regarded as a sequence of **play units**, each of which starts with a **live scene** and ends with the beginning of the next live scene. A live scene is defined as the time interval which begins with the start of

0-7803-9332-5/05/$20.00 ©2005 IEEE

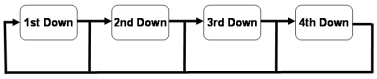

Fig. 1. Game State Transition Rule

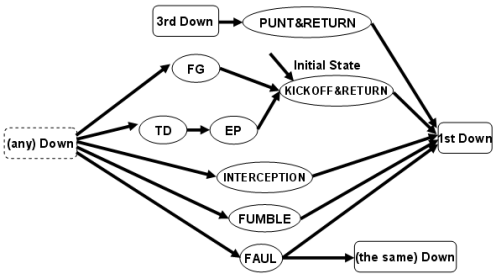

Fig. 2. Event Transition Rules

a play and ends with the end of the play (e.g. the score, the outbound of the ball, etc.). Note that the live scenes usually start with some characteristic images and the beginning of the live scenes is considered as the boundary of play units.

A game of Ameican football is basically a repetition of 4 types of sequences of downs from *1st Down* to *4th Down*. These 4 types of sequences are (1), (1,2), (1,2,3), and (1,2,3,4) as shown in Fig.1, with each number representing each down. Here, we call these downs *game states*, the sequence of the 4 types of sequences *game state sequence*, and the transition rule **Game State Transition Rule**.

Plays occur during these game states, and some plays called *events* change the basic transitions of the game states. We consider 8 types of event here: *Touch Down(TD)*, *Extra Point(EP)*, *Field Goal(FG)*, *KickOff(KO)&Return*, *Punt&Return*, *Fumble*, *Interception*, and *Foul*. Some events can occur by themselves , while others can only occur in a specific order, composing *event sequence* such as (*Fumble*) and (*TD*, *EP*, *KO&Return*). These *event sequences* can occur at any game state and the game state will return to *1st Down* after the last event. As exceptions, *Punt&Return* can only occur on *3rd Down*, and after a *Foul*, the game state can return to either *1st Down* or its immediately preceding game state. The rules for composing the *event sequences* and how they transit from and to game states, hereafter called **Event Transition Rules**, are shown in Fig.2.

Note that each game state or event in these rules corresponds to a play unit. Combining these two types of rules, an American football game can be considered as a sequence of play units constructed as shown in Fig.3. We call these rules **Play Unit Transition Rules**. These rules show that a video of an American football game can now be considered as a repetition of **offence units**, each of which is (*game state sequence*, *event sequence*). Note that every event framed in a dotted line in Fig.3 occurs during its immediately preceding game state, and therefore, corresponds to a play unit together with the preceding game state.

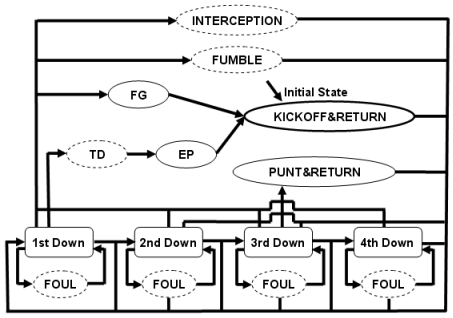

Fig. 3. Play Unit Transition Rules

3. VIDEO PARSING BASED ON TRANSITION RULES

Our proposed method segments an American football video into play units and parse the play unit sequence according to the transition rules described in Section 2. The video stream is segmented into play units by detecting the **live scenes** and **superimposed texts**. The superimposed text is a kind of graphical text which is overlayed on the video in the production process to provide important information about the game. The information related to game states and events is extracted respectively from the detected superimposed texts and a transcript of the audio stream called **Closed-Caption(CC) Text**. The play unit sequence associated with the game state and event information is parsed according to the transition rules. We firstly parse the play unit sequence according to Event Transition Rules to find the final states of offence units. Once the *event sequences* are confirmed, the offence units are parsed according to the Game State Transition Rule.

3.1. Image Analysis

For American football, we found three kinds of characteristic images which indicate the beginning of live scenes based on the analysis of actual broadcasted videos. Here, we find the beginning of live scenes with template matching of these example images and the initial few frames of each **shot**, which is a video segment recorded by a single camera. The details of the method can be found in [7].

3.2. Superimposed Text Analysis

Information about the game states(*1st Down–4th Down*) and *Fouls* is often displayed in the superimposed texts in American football videos. Fig.4 shows examples of the superimposed texts. Different production companies use different types of superimposed texts(Fig.4-(a),(c)); however, they almost always appear at specific places in an image(Fig.4-(a),(b)). Fig.4-(d) is an example of *Fouls*. These superimposed texts do not necessarily appear for all live scenes; however, when they do, they appear sometime between the middle of the preceding live scene and the end of the current live scene. They can disappear and reappear more than

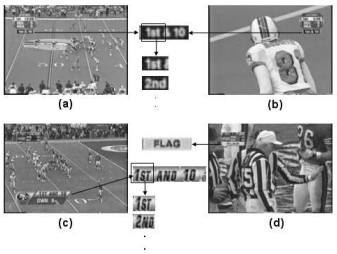

Fig. 4. Examples of Superimposed Texts

Table 1. Keywords for Events

Event	Keywords
TD	TOUCHDOWN, TOUCH DOWN
EP	EXTRAPOINT, EXTRA POINT
FG	FIELDGOAL, FIELD GOAL
KO&RETURN	KICKOFF, KICK OFF, RETURN
PUNT&RETURN	PUNT, RETURN
INTERCEPTION	INTERCEPT, INTERCEPTED, INTERCEPTION
FUMBLE	FUMBLE

once during this time interval. Note that these superimposed texts never appear at the event node framed in a solid line in Fig.3. We extract game state information by recognizing those superimposed texts by template histogram matching with superimposed text models.

3.3. CC Text Analysis

The CC text is first segmented into **CC segments** based on speaker changes or the interval of their talks. The CC segments are then classified into 4 scene categories including live scenes, and based on the results, the CC text is segmented into **CC play units**. The details of the segmentation method are found in [8]. Next, keywords for events shown in Table 1 are searched in the CC text. Since "RETURN" applies to both *KO & RETURN* and *PUNT&RETURN*, the event is only detected as *RETURN* here and will be determined at the parsing stage.

3.4. Video Parsing

Fig.5 shows how we parse a video and each step of this procedure is explained below.

1) **Play Unit Segmentation and Game State Extraction**
   The detected superimposed texts are associated with the shots where the superimposed texts disappear. The video stream is segmented into play units at the initial frames of both the shots which are associated with the superimposed texts and the live scenes detected in Section 3.1.
2) **Event Extraction**
   The CC play units with event keywords are associated with the closest play units. If the events are one of those framed in a solid line in Fig.3, the closest play unit without a superimposed text is selected.
3) **Parsing based on Event Transition Rules**
   For each play unit with event information, we check

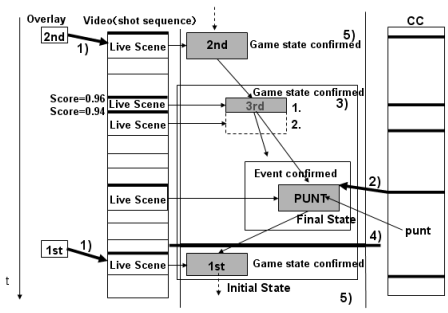

Fig. 5. How to parse a video

if it, together with its adjacent play units, can conform to any rule in Event Transition Rules. If it can, these play units are confirmed as an *event sequence* with their event/game state information being determined according to the corresponding rule; otherwise, the event information is discarded as unnecessary information in the CC text.

4) **Offence Unit Segmentation**
   The whole play unit sequence is segmented at the beginning of the last event play unit in the confirmed *event sequences*. Each segmented sequence corresponds to an offence unit.
5) **Parsing based on Game State Transition Rule**
   Each offence unit is parsed by selecting the play units which conform to the Game State Transition Rule. The play units which have been associated with the superimposed texts are selected prior to others as long as they conform to the transition rule and the unselected play units will be discarded as errors falsely detected at Step 1).

The following two strategies are used for Step 3) and 5).

1. Those play units without any semantic information can be used as either game state or event play unit as long as they conform to transition rules. Their semantic information is complemented according to their corresponding rules.
2. If there are more than one play unit which can be used according to the rules, the one with the highest similarity to the example images at the live scene detection stage is selected.

4. EXPERIMENTAL RESULTS

We tested our method with 6 broadcasted American football videos, each 20 minutes long. Table 2 shows the details of the videos, such as their production companies, production years, and participating teams. Moreover, "♯ of Superimposed Texts" represents the number of play units where the superimposed texts appear at least once, "♯ of Game States" the number of play units which correspond to game states, "♯ of Events" the number of play units which correspond to events, and "♯ of Play Units" the total number of play

Table 2. American Football Videos

Video	Company	Year	Teams	♯ of Superimposed Texts (ideal ♯ of Superimposed Texts)	♯ of Game States	♯ of Events (-♯ of *Fouls*)	♯ of Play Units
Video1	abc	1999	Denver v.s. San Diego	13(17)	16	4(3)	18
Video2	CBS	1999	Miami v.s. Seattle	14(+2)(16)	16	5(5)	20
Video3	FOX	1998	Green Bay v.s. San Francisco	22(24)	21	6(3)	24
Video4	FOX	2000	New York v.s. Cleveland	17(+6)(17)	16	3(2)	18
Video5	FOX	2000	Philadelphia v.s. Pittsburgh	18(21)	19	6(4)	23
Video6	FOX	2000	Dallas v.s. Baltimore	13(16)	15	5(4)	19

Table 3. Experimental Results

	Precision	Recall	Discarded Game States	Complemented Game States	Discarded Events	Complemented Events	Discarded Play Units
Video1	100%(18/18)	100%(18/18)	0/0	4/4	2/2	0/0	4/4
Video2	90%(17/19)	85%(17/20)	2/2	0/2	4/4	1/1	2/3
Video3	78%(18/23)	75%(18/24)	0/0	2/2	1/1	0/2	9/12
Video4	86%(18/21)	100%(18/18)	5/6	0/0	1/1	0/0	20/23
Video5	96%(22/23)	96%(22/23)	0/0	2/3	1/1	0/0	5/6
Video6	100%(18/18)	95%(18/19)	0/0	1/2	2/2	1/2	6/6
Total	91%(111/122)	91%(111/122)	7/8	9/13	11/11	2/5	46/54

units. Since the superimposed texts also appear when *Fouls* occur, the number of play units which should have superimposed texts and the number of play units where any event except *Fouls* occur are also shown in the parentheses in the columns of "♯ of Superimposed Texts" and "♯ of Events". Moreover, (+2) and (+6) in the column of "♯ of Superimposed Texts" for Video2 and Video4 represent the number of play units where the superimposed texts appear more than once.

Table 3 shows the experimental results. The results are evaluated with Recall(=P_1/P_2) and Precision(=P_1/P_3), where P_1 represents the number of play units which were correctly parsed and associated with their correct game state or event information, P_2 the number of actual play units, P_3 the number of play units which were associated with game state or event information after parsing. Moreover, in this table, "Discarded Game States" shows (the number of play units whose game state information were discarded / the number of play units whose game state information had been incorrectly detected), "Complemented Game States" (the number of play units whose game state information were complemented / the number of play units whose game state information had been missing), "Discarded Events" and "Complmented Events" show the same numbers for event information, and finally, "Discarded Play Units" shows (the number of discarded play units / the number of play units which had been incorrectly detected).

The table shows that almost all the unnecessary play units with/without game state/event information were correctly discarded. Most of the missing game state information were also complemented, while the events which can occur by themselves were hard to be complemented since the play unit sequence can conform to the Game State Transition Rule without such events. Moreover, our method only falsely discarded 5 play units, complemented 3 game state/event information, and discarded no game state/event information. These results show that using the transition rules enabled us to obtain good results in both precision and recall rates by successfully keeping only the necessary results while discarding or complementing most of the unnecessary or insufficient results often obtained from a simple information stream analysis.

5. CONCLUSION

This paper proposed a method of automatic parsing of American football videos. Our method first obtains a play unit sequence with possible semantic information by combining the results of the image and text analysis, then parse the sequence based on transition rules of an American football game. As the results of the experiments, we successfully obtained good results 91% in both precision and recall rates on average, discarding or complementing most of the unnecessary or insufficient results obtained from simple information stream analysis. The rules and how to obtain the semantic information depend on the types of videos. Experiments with other types of video, which can also be represented with similar transition rules, such as other kinds of sports videos or cooking videos, are necessary to investigate the scalability of our method.

6. REFERENCES

[1] N.Babaguchi, Y.Kawai, and T.Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration", IEEE Transaction on Multimedia, vol.4, no.1,pp.68-75, March 2001.

[2] K.Miura, R.Hamada, I.Ide, S.Sakai, and H.Tanaka, "Associating Cooking Video Segments with Preparation Steps", *Proc. CIVR'03*, pp.174-183, 2003.

[3] A. Merlino, D. Morey, and Mark Maybury, "Broadcast news navigation using story segmentation", *Proc. MULTIMEDIA'97*, pp.381-392, 1998.

[4] D.Zhong and S.F.Chang, "Structure Analysis of Sports Video Using Domain Models", *Proc. IEEE ICME'01*, 2001.

[5] B.Li and M.I.Sezan, "Event Detection and Summarization in Sports Video", *Proc. IEEE CVPR'01*, Demos pp.29-30, 2001.

[6] L.Xie, P.Xu, S-F Chang, A.Divakaran, and H.Sun, "Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models", Pattern Recognition Letters, 25(7):767-775, May 2004.

[7] N.Nitta, N.Babaguchi, and T.Kitahashi, "Extracting Actors, Actions and Events from Sports Video – A Fundamental Approach to Story Tracking –", *Proc. ICPR'00*, pp.718-721, 2000.

[8] N.Nitta and N.Babaguchi, "Automatic Story Segmentation of Closed-Caption Text for Semantic Content Analysis of Broadcasted Sports Video", Proc. 8th International Workshop on Multimedia Information Systems (MIS 2002), pp.110-116, 2002