INTERMODAL COLLABORATION: A STRATEGY FOR SEMANTIC CONTENT ANALYSIS FOR BROADCASTED SPORTS VIDEO

Noboru Babaguchi and *Naoko Nitta*

2-1 Yamadaoka Suita, 565-0871 Japan
Graduate School of Engineering, Osaka University
babaguchi@comm.eng.osaka-u.ac.jp

ABSTRACT

This paper presents intermodal collaboration: a strategy for semantic content analysis for broadcasted sports video. The broadcasted video can be viewed as a set of multimodal streams such as visual, auditory, text (closed caption) and graphics streams. Collaborative analysis for the multimodal streams is achieved based on temporal dependency between their streams, in order to improve the reliability and efficiency for semantic content analysis such as extracting highlight scenes from sports video and automatically generating annotations of specific scenes. A couple of case studies are shown to experimentally confirm the effectiveness of intermodal collaboration.

1. INTRODUCTION

Various *highlight scenes* in sports video, e.g. touchdown in American football, homerun in baseball and 3-point shoot in basketball are actually what we want to retrieve. Video retrieval based on such highlights is a typical example of semantic contents based retrieval[1]. Also, abstracted videos and keyframe layouts can be made of highlight scenes. Thus, extraction of highlight scenes is one of the most important tasks in video content analysis. In addition, *annotations* of each scene about a team, a player, a play, etc. are of much importance because such data may become keywords in retrieval of sports video. Automated annotation tools based on semantic content analysis are strongly required.

A number of researchers, e.g. [2], have proposed methods of extracting events and actions in sports games from visual streams. It is noted that these methods focus on single modality of the video data. However, it is very difficult to achieve reliable and efficient extraction of events and actions through image analysis. If we try it, we need some visual model of the event. What can be the visual model of touchdown or homerun? To define this, we have to assume possible cases of such events, for example, the touchdown by pass, that by running and that after turnover. It is not easy to construct a compact model that covers a number of cases of concern. Instead, we may employ a naive way of collecting a lot of examples of the event from known data. This problem results from the gap between the image signal level and the semantic level. Accordingly we may as well seek for different strategies.

Alternatively, great emphasis has been placed on multimodality[3] of the broadcasted video in recent years. As is well known, the video data is composed of temporally synchronized *multimodal streams* which are visual, auditory, text and graphics streams. The visual stream is a sequence of image frames, and the auditory stream is a mixture of a couple of auditory sources such as speech, music and sound. For broadcasted video, in addition, closed caption (CC) text, which is a transcript of the speech part of the auditory stream, can be viewed as the text stream. The graphics stream is a sequence of overlays or video captions which have rich information about the contents of the video data. Note that these streams are closely related to each other.

In this paper, we discuss the effectiveness of *intermodal collaboration* for broadcasted sports video, investigating a couple of its case studies. Intermodal collaboration means a strategy of collaborative processing taking account of semantical dependency between the multimodal streams[4]. Its aim is to improve the reliability and efficiency in analyzing semantical contents of video. Although the computation for analyzing the visual stream is most costly, the use of other streams may be capable of reducing it.

In what follows, we describe our approaches to collaborative analysis as follows: 1) the visual and text streams[4, 5, 6], 2) the visual, auditory and text streams[7], and 3) the graphics stream and the external metadata[8].

2. RELATED WORK

Let us first discuss the recent work attempting collaborative analysis for semantic video contents. We divide approaches into four classes, considering what kind of stream is used.

The first class is using the auditory and visual streams. Chang et al.[9] used the word spotting technique to limit the search space of the visual stream. Rui et al.[10] developed a method of extracting baseball highlights through classification of auditory sources.

The second class is using the text and visual streams. For sports video, Babaguchi et al.[4] and Nitta et al.[5] tried to extract events, actions and players from American foot-

This work was supported in part by Telecommunications Advancement Organization of Japan.

ball broadcasts, using the CC and visual streams. Lazarescu et al.[11] proposed analyzing a player's action and a team's formation combining natural language processing and video processing.

The third class is using the auditory, text and visual streams. Smith et al. [12] proposed a video summarization method based on features obtained from those streams. Huang et al. [13] and Jasinschi et al. [14] proposed methods for segmenting a news video and semantically classifying each segment combining features of all the streams. Li et al. [15] proposed a method of detecting semantically related scenes based on the similarity between the visual and auditory streams, as well as acquiring the semantic contents from the text stream. Recently, Miyauchi et al.[7] have attempted collaborative analysis for highlight detection.

The fourth class is using addtional streams of overlays or video captions. Satoh et al. [16] have developed a system that identifies faces, by associating the faces extracted from the visual stream and their names from the streams of CC text and video captions. Babaguchi et al.[8] tried exploiting graphical overlays including text to find highlights in the video.

3. COLLABORATION BETWEEN TEXT AND VISUAL STREAMS

This section presents three methods for semantic content analysis with collaboration between text and visual streams. The text stream we use here is called closed-caption(CC) text, which is a transcript of the speech and the sound. The CC text also includes the information about speakers and topic change. Because the text stream is inherently a sequence of words, its fragment can be a good semantic index. Moreover the text stream analysis is less costly than any other stream analysis in the video data. Therefore, the CC text can become a promising clue for content analysis.

3.1. Highlight extraction

The first method[4] is developed for extracting *highlight scenes* such as touchdown and homerun scenes from the video data. We here focus on temporal correspondence between the visual and text streams. The proposed method attempts to seek for time intervals in which events are likely to take place through extraction of keywords from the text stream. To avoid error extraction, keyword chains including not only a single word corresponding to the highlight, e.g. "touchdown," but also its related words, which are predifined based on our heuristics, are taken into account. After the keyword extraction, the visual stream located during the given time interval is examined. Partitioning it into some shots, we consider color distribution of each shot and match with an example image sequence which is a temporal image model of the highlight scene.

We tested our method for American football TV videos. There were 40 highlight scenes containing touchdown, extra point and field goal scenes in three different videos, each of which was about three hours in length. About 40 keywords were determined beforehand. The average rates of precision and recall for highlight extraction were 74% and 81%, respectively. The ratio with respect to the actual processing time between the text and visual stream analyses was approximately 1:1500. This implies that intermodal collaboration enables us to efficiently search video data by its contents.

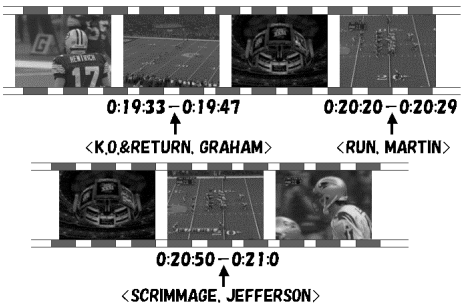

Fig. 1. Example of descriptions about plays and players.

3.2. Extraction of information about plays and players

The second method[5] attempts to attach textual indexes about *plays* and its related *players* to each *live scene*, in which the plays are actually going on. Announcers in sports broadcasts usually talk about the current situation of a game in the live scenes. Therefore, we try to extract CC segments corresponding to the live scenes from the CC text, and then acquire the information about plays and players.

Since players tend to take their routine stance at the beginning of each live scene, we can often see stationary images which are captured by a camera positioned at a fixed location at that time. Considering a shot starting with these stationary images as the live scene, we can find it in the visual stream by means of matching between these characteristic stationary images and the initial frames of each shot. Finally, temporally associating the information extracted from the CC text with the live scene realizes semantic indexing of the video.

We carried out experiments with two American football videos. Correct descriptions to the live scenes were obtained with the recall rate (= ♯ of correct results / ♯ of actual live scenes) of 86% (76/88) and the precision rate (= ♯ of correct results / ♯ of total obtained results) of 95% (76/80) on average. Examples of the attached descriptions are shown in Fig.1. In this figure, each image represents the first image of each shot, and the descriptions are shown in a form of <Play, Player>. The time in the video was estimated from the first and the last image frame numbers of each extracted live scene.

3.3. Story segmentation

The third method[6] is to segment the CC text into meaningful units, called *CC story units* and to link them with *video*

Table 1. Examples of obtained keywords from each CC story unit. The annotation is a fragment of CC text. The common words between them are underlined.

Obtained Keywords	Actual Annotations
LEWIS IZELL REESE	punt by Barry Cantrell Lewis's return taken down by Izell Reese
QUDRY ISMAIL DILFER PRIEST HOLMES	Dilfer's pass complete to Priest Homes
JAMAL LEWIS RYAN McNEIL FLAG FIVE-YARD PENALTY	Jamal Lewis runs flag for Ryan McNeil five-yard penalty
OGDEN BRANDON NOBLE	Lewis came from behind Ogden stopped by Brandon Noble

story units. First, with a Bayesian network, we try to learn characteristic patterns of each CC segment to be classified as 'live', 'replay', 'commercial', and 'others.' For its applicability to other kinds of sports, we use domain-independent features such as speaker's name and sentence length. Using the learned network, we probabilistically classify each CC segment into the four classes. A temporal interval from a 'live' CC segment to its next 'live' CC segment is determined as the CC story unit. Next, the video story unit is formed by consecutive shots which begin with specific image frames characterizing the 'live' scene. Finally, both units are associated with each other. The keywords in the CC story unit are viewed as descriptions of a scene.

For ten American football videos, we were able to successfully associate CC story units with video story ones. Given were the recall rate of 92% (191/207) and the precision rate of 89% (191/214) on average. Table1 indicates examples of obtained keywords, which inform us of the situation of each scene. The associated CC story units themselves can be used as descriptions to be searched against a query by keywords, working as a rich information source for a retrieval system.

4. COLLABORATION AMONG TEXT, AUDITORY AND VISUAL STREAMS

For more reliable extraction of highlights, we extend the method which was described in Section 3.1, augmenting *auditory stream analysis*[7]. The audience of a football stadium or a ball park may cheer or applaud when meeting highlight scenes. Therefore, the sound in sports video can help us detect highlights. We here focus on the short time energy (STE)[17] as an auditory feature. In addition, to avoid employing the heuristics in our method, example based learning and nearest neighbor classification are introduced. The outline of this method is illustrated in Fig.2. First, temporal intervals when highlights are likely to happen are detected based on appearance patterns of words, called indexed terms, in the CC text. Unlike the previous

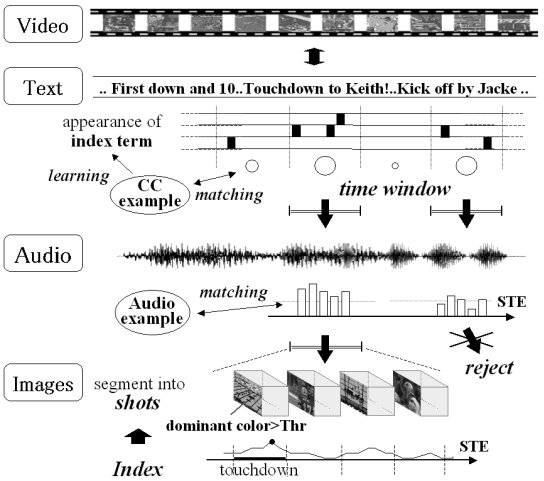

Fig. 2. Outline of collaborative analysis among the text, auditory (audio) and visual (images) streams.

method[4, 5], the indexed terms are automatically derived from domain vocabulary. Second, these intervals are verified by the short time energy of the audio signal. Finally, the highlight shot is determined by dominant color and audio energy from all the shots in the interval.

For six actual video streams of broadcasted American football programs, we compared the two cases: case 1) analysis of the text stream and case 2) collaborative analysis of the text and auditory streams. For case 1), the recall and precision rates for detection of highlights such as touchdown and field goal were 81% (38/47) and 66% (38/58), respectively. For case 2), the recall and precision rates were 77% (36/47) and 84% (36/43), respectively. While the recall decreased by 4 %, the precision increased by 18 %. This means that the auditory stream analysis is effective for false detection. Further, the accuracy of shot indexing by highlights was 75 % for the first candidate and 97 % for up to the second candidate. The stepwise analysis for text, auditory and visual streams allows extremely efficient processing. In particular, the text stream analysis restrict the search space to 5 % of the whole stream. This demonstrates the advantage of intermodal collaboration. In addition, this method can be successfully applied to the baseball games.

5. COLLABORATION BETWEEN GRAPHICS STREAM AND EXTERNAL METADATA

As described in the preceding sections, though intermodal collaboration is useful, its accuracy is insufficient from practical point of view. For example, consider making an abstracted video from highlight scenes. The viewer will not be satisfied with the abstracted video if the most significant scene is lacking in it.

We exploit a completely original strategy: the use of *external metadata*. For professional sports games, we can obtain *gamestats* at some websites. The gamestats possess valuable information about the game. For example, "1st Q,

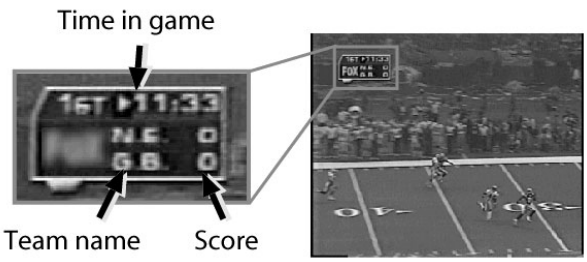

Fig. 3. Example of an overlay.

3:32, GB, Rison 54 pass from Favre (Jacke kick)." is a description in the gamestats of an American football game. Taking advantage of the time in a game (game time) in this description helps us find highlight scenes in the video data. In this case, we have to link the description with the video segments. To achieve this, we consider a characteristic specific to the broadcasted sports video. It is appearance of overlays including text. A series of overlays can be regarded as the graphics stream. In actual TV programs, there are several kinds of overlays. Fig. 3 shows an example of the overlay. Among them, we deal with the overlay indicating the game time when an event happened, called event time.

A procedure to identify an image frame on which the target overlay appears is as follows. Presence analysis of the overlays is first performed. Pattern matching between the overlay model and the image frame informs us whether the target overlay is present on the frame. Next, we design a digit recognizer to identify an event frame which is a frame at the event time. Employing a template corresponding to each digit representing the event time to be searched, we find a good matching position in the overlay region. Matching is made by sliding the template horizontally and vertically for each digit. If the event time described in the gamestats is found with the recognizer, we determine that the image frame with the matched overlay should be the event frame. In this way, detection of the significant events can be reduced to searching the text indicating the event time.

For 45 events in five different video streams, the accuracy of event detection was 96%. Only two events were missed because no overlay was present at the event time. As a result, we detected all the event frames when the overlays showing the event time were really present. The above procedure is introduced to highlight based video abstraction[8]. This approach seems somewhat limited, but is vital for the broadcasted sports video with the graphics stream if the external metadata is available. We think that the approach is a kind of intermodal collaboration using distributed heterogeneous information sources and is a novel idea in video contents processing.

6. CONCLUSIONS

We have discussed intermodal collaboration from a couple of case studies for broadcasted sports video. As indicated in this paper, intermodal collaboration is a powerful and promising strategy for diverse applications about semantic video contents: highlight extraction, annotation generation, and so on. However, the approaches to intermodal collaboration have just got launched, and there exist a lot of open problems. On one hand, deeper level analysis of sentences in CC text should be required. On the other hand, more complicated image analysis such as motion analysis should be considered. Flexible integration of each analysis should be further explored.

7. REFERENCES

[1] P. Aigrain, H. J. Zhang, and D. Petkovic, "Content-based Representation and Retrieval of Visual Media: a State-of-the Art Review," *Multimedia Tools and Applications,* vol.3, pp.179-202, 1996.

[2] S.S. Intille and A.F. Bobick, "Recognizing Planned, Multi-person Action," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 414-445, March 2001.

[3] C.G.M.Snoek and M.Worring, "A Review on Multimodal Video Indexing," in *Proc. ICME2002*, vol.2, pp.21-24, 2002.

[4] N. Babaguchi, Y. Kawai and T. Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration," *IEEE Trans. Multimedia*, vol.4, no.1, pp.68-75, March 2002.

[5] N. Nitta, N. Babaguchi and T. Kitahashi, "Generating Semantic Descriptions of Broadcasted Sports Video Based on Structure of Sports Game," *Multimedia Tools and Applications* (to be published)

[6] N. Nitta and N. Babaguchi, "Automatic Story Segmentation of Closed-Caption Text for Semantic Content Analysis of Broadcasted Sports Video," in *Proc. MIS2002*, pp.110-116, 2002.

[7] S. Miyauchi, A. Hirano, N. Babaguchi and T. Kitahashi, "Collaborative Multimedia Analysis for Detecting Semantical Events from Broadcasted Sports Video," in *Proc. 16th ICPR*, pp.1009-1012, 2002.

[8] N. Babaguchi, Y. Kawai, T. Ogura and T. Kitahashi, "Personalized Abstraction of Broadcasted American Football Video by Highlight Selection," *IEEE Trans. Multimedia* (to be published)

[9] Y.Chang, W.Zeng, I.Kamel and R.Alonso, "Integrated Image and Speech Analysis for Content-based Video Indexing," in *Proc. IEEE ICMCS'96*, pp.306–313, 1996.

[10] Y. Rui, A. Gupta and A. Acero, "Automatically Extracting Highlights for TV Baseball Programs," in *Proc. ACM Multimedia 2000*, pp.105-115, 2000.

[11] M. Lazarescu, S. Venkatesh, G. West and T. Caelli, "On the Automated Interpretation and Indexing of American Football," in *Proc. IEEE ICMCS'99*, vol.1, pp.802-806, 1999.

[12] M. A. Smith and T. Kanade, "Video Skimming and Characterization Through the Combination of Image and Language Understanding Techniques," in *Proc. CVPR97*, pp.775-781, 1997.

[13] Q.Huang, Z.Liu, A.Rosenberg, D.Gibbon and B.Shahraray, "Automated Generation of News Content Hierarchy by Integrating Audio, Video, and Text Information," in *Proc. IEEE ICASSP'99*, pp.3025-3028, 1999.

[14] R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, and J. Zimmerman, "Video Scouting: an Architecture and System for the Integration of Multimedia Information in Personal TV Applications," in *Proc. IEEE ICASSP'01*, 2001.

[15] Y.Li, W.Ming, and C-C.J.Kuo, "Semantic Video Content Abstraction Based on Multiple Cues," in *Proc. IEEE ICME'01*, 2001.

[16] S.Satoh, Y.Nakamura and T.Kanade, "Name-It: Naming and Detecting Faces in News Videos," *IEEE Multimedia*, pp.22–35, 1999.

[17] T. Zhang and C. J. Kuo, "Heuristic Approach for Generic Audio Data Segmentation and Annotation," in *Proc. ACM Multimedia 1999*, pp.67-76, 1999.