Automatic Story Segmentation of Closed-Caption Text for Semantic Content Analysis of Broadcasted Sports Video*

Naoko Nitta, Noboru Babaguchi
ISIR, Osaka University
8-1 Mihogaoka Ibaraki Osaka, 567-0047, Japan
E-mail: naoko@am.sanken.osaka-u.ac.jp

Abstract

Sports videos can be characterized as a sequence of recurrent semantic story units. Storing sports videos in this story-unit-based form will lead to develop an intelligent content-based retrieval, browsing, and summarization system. The storage requires segmentation of videos and semantic understanding of each segment. Since transcribed broadcasted video speech, the *closed-caption text*, can be the useful information source for semantic indexing of each story unit, this paper proposes a method to automatically segment the closed-caption text of sports videos into the semantic units. The proposed method firstly tries to segment the speech transcript into the *scene units*, a set of which composes a *story unit*, in a probabilistic framework based on Bayesian networks. Finding the boundaries of the set of the scene units enables us to generate the story units in the close-caption. In this paper, we discuss some experimental results and the potentiality for utilizing them for indexing of the video and speech summarization.

1 Introduction

Continuous increase in the amount of multimedia data has strongly required efficient browsing, retrieval and summarization systems. And video structure analysis and indexing are important sub-problems as the basis of further detailed processing. Video structure analysis requires segmentation of the video, and segmenting videos into shots is often the first step in video analysis. However, the *semantic scenes* which are often composed of the several shots play a more important role for semantic understanding of videos. To solve this problem, some works have restricted the target videos to a specific domain and segment the video semantically applying some domain-specific rules to image stream analysis. For indexing, visual features also have been analyzed. While these features are useful in retrieving scenes based on visual similarity, they do not necessarily contain much information at the semantic level. Intuitively, the speech in the video contains more detailed semantic information, and analysis of the speech transcript is easier than image analysis. Moreover, the availability of the closed-caption text facilitates the acquisition of the speech transcript without speech recognition process. As a step toward the extraction of the semantic information from the speech transcript, its semantic segmentation according to the domain-specific structure is desirable. In this paper, we address the semantic structure of sports videos and propose a method to automatically segment the closed-caption text of sports videos into the *story units*, the results of which can be a good foundation for semantic indexing.

*This work was supported in part by a Grant-in-Aid for scientific research from the Japan Society for the Promotion of Science and also by the Telecommunications Advancement Organization of Japan.

Videos have some typical structures which depend on their genres. For instance, a news video can be considered as a sequence of units each of which starts with an image frame presenting an anchor person followed by a variety of news. A drama video can be considered as an assembly of the semantically interrelated units. Such flows of the units construct every video and give it the semantic meaning or the story. There has been some prior work analyzing semantic structure of sports videos [1, 2, 3]. Zhong et al. [1] focused on the content structure of a sports video and tried to extract some patterned event boundaries from the image stream. Li et al. [2] also tried to extract the patterned event boundaries from the image stream and summarize the video by assembling them. Xu et al. [3] also tried to segment a soccer video into play/break scenes with frame-based image analysis. Most of them have exploited image analysis and segmented the video stream into some story units. However, they did not consider the detailed semantic content of each unit and the potential indexing of each unit is simply restricted to "play" or "break".

Moreover, as previous work on more detailed indexing or annotation of sports videos, Lazarescu et al. [4] tried to make annotation about the movement of the players by searching keywords from the text stream and analyzing the image stream. Chang et al. [5] tried to detect important events by integrating the audio and the image streams. Babaguchi et al. [6] also proposed event scene detection by integrating the text and the image streams. Since they did not take the structure of sports videos into consideration, the indexing can not always be done for each unit. However, based on these researches, other information streams than the image stream such as the audio and the text stream seem good candidates as information sources for indexing or annotation.

In fact, many researchers have used the automatically transcribed or the closed-caption text stream for topic segmentation and categorization of news videos [7, 8, 9]. Takao et al. [7] tried to find the topic boundary of the news speech and summarize them using TF-IDF with the speech transcript. Hauptmann et al. [8] also tried to segment the news video integrating speech recognition, text, and image streams. Greiff et al. [9] used Hidden Markov Model associated with the parameters which reflect the occurrence of words for segmentation of news videos. They all used the characteristics of word occurrence for each topic or the topic boundaries, however, due to the relative uniformity of the topics for the sports videos, few succeeded in semantic segmentation of the closed-caption text of sports videos.

In this paper, we also focus on the *closed-caption text* which is the speech transcript of the announcers as the important information source for indexing and present a method of segmenting it according to the semantic structure of sports videos. Our method exploits the superficial features avoiding the use of many keywords or key phrases which will complicate the versatility of the system, and segment the closed-caption text into the *semantic units of the sports program*, parts of which have useful information to grasp the story, on the basis of the probabilistic Bayesian Network framework. Since a set of these units construct the *unit of the sports game*, finding the boundary of the sets leads us to segment the closed-caption into *story units*, that is, the semantic units of the sports game.

2 Story Based Structure of Sports Video

Sports videos have two kinds of structures seen from different points of view: sports TV programs and sports games. Here, we summarize these different structures and address the semantic units which should be useful for the subsequent handling.

A sports game can be expressed as a tree. For example, an American Football game is divided first into two halves, then into four quarters, several team offenses, and several downs each of which corresponds to a play, and a soccer game contains two halves, each of which is divided into several plays. *A play* is defined as an interval which starts with the formation view taken by the cameras at the fixed locations and, when some events (the score event, the foul, etc.) occur, ends with a scene transition to the player's closeup, spectators view, etc. We define the play interval as the *story unit*.

On the other hand, a TV program of a sports game can be regarded as a sequence of the several scenes. We define the "Live" scene as the time interval during which a play continues, the "Replay" scenes as the one in which they make the detailed explanation about the "Live" scenes. Other scenes such as "Report", "Studio", "Commercial Message (CM)",

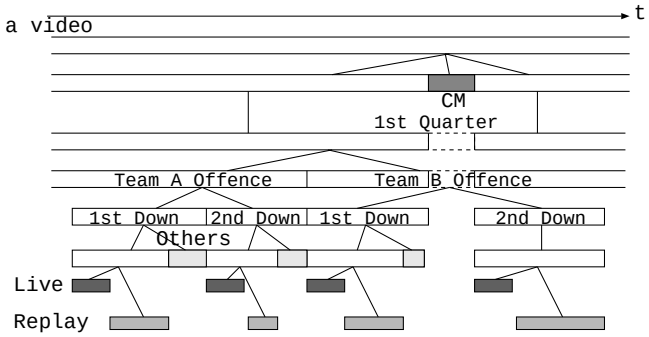

Figure 1: Overall Structure of Sports Video

etc. are considered unrelated to the game. We define each scene as the *scene unit*, and the time interval between a "Live" scene unit and the next "Live" scene unit can be regarded as a *story unit* of the game structure. As a result of summarizing these two structures, a sports video can be structurized as in Figure1.

3 Scene and Story Segmentation

Previous work [1, 2, 3, 10] also focused on the structure of the sports TV programs, as discussed in Section 2. However, most of them have neglected the structure of the sports game, which is also important semantic contents for sports videos. Therefore, focusing attention on the sports speech transcript called the closed-caption(CC) as the useful information source for the extraction of semantic content, we try to segment it into both the *scene units* and the *story units*.

The CC text is embedded in the video signals as the character code and is easy to be reproduced in the text form. We add the "Timed-Stamp", which represents the image frame in which the first character of each line appears, to the original CC text. The "Timed-Stamp" increases by 1/30 seconds. The CC text contains several markers such as "$>>$" to indicate the change of the speaker, "NAME:" to indicate the speaker, etc., however, for sports videos, most of them only indicate the speaker changes which reflect neither the scene changes nor the story changes.

The CC text and the image stream usually have the time-lag between them. Moreover, the announcers do not necessarily talk about the present scene. Therefore, CC segmentation should be done individually without the interference of other information streams so that the CC segments correspond to the video scenes semantically.

We firstly consider the change of the speaker and the interval of the talks as the boundary of the CC text and define the segment between the boundaries as a *CC segment*. Each segment is supposed to belong to one of the scene categories. Based on the structure of sports TV programs, the scene categories we try to categorize are "Live" , "Replay", "Others" ("Report", "Studio", etc.), and "CM" scenes.

Table1 shows the characteristics of each scene category. The "Speakers" column shows the speakers who usually talk in each scene. The "Length of Sentences" and the "$\sharp$ of Sentences" respectively shows the general length and the number of the sentences in each scene. For example, in live scenes, since they usually make simple comments about the on-going play, the length of the sentences tends to be short and they usually use a few sentences. The "Situational Phrases" shows likeliness of the appearance of the *situational phrases*, which we define as the phrases expressing the situation of each story unit, such as "First and 10" for American Football, "One ball and two strikes" for baseball, and "15-0" for tennis.

Considering these characteristics, we use 6 features for each CC segment: **the name of the announcers**, **the number of the sentences**, **the length of the sentences**, **the number of the players' names**, **the situational phrases**, and **the numbers** (which possibly represent the score, yards, etc.).

Moreover, we assume from the structure of the sports TV program that the scenes have some rules in how they line

Table 1: Characteristics of each scene in CC text

	Live	Replay	Others	CM
Speakers	Announcer	Announcer, Commentator	Announcer, Commentator, Reporter, etc.	Others
Length of Sentences	Short	Long	cannot be determined	cannot be determined
♯ of Sentences	A few	Many	cannot be determined	cannot be determined
Situational Phrases	highly likely	less likely	probably	rarely

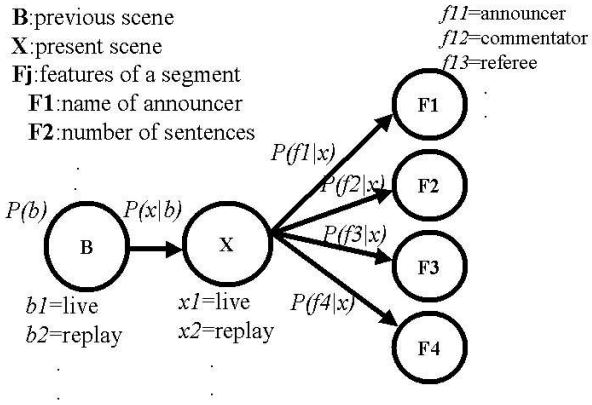

Figure 2: Bayesian Network

up. Therefore, the scene category of a CC segment depends on the scene category of the previous CC segment as well as its own features. In this paper, to tackle the uncertainty of the information, we use a probabilistic framework which can handle such information, namely Bayesian Network(BN), to categorize each CC segment.

The BN in Figure2 shows the relationship between the present scene category and other factors. Node *B* represents the category of the previous scene and is the parent of the node X, which represents the category of the present scene. Nodes F_j, which are the jth children nodes of X, represent the features of the present segment. $P(x|b)$ and $P(fi|x)$ represent the probability of the transitions, where *x* represents the values of variables on nodes X, and *b* and f_i the values of each state of the corresponding nodes. For example, $P(live|live)$ represents the probability that a live scene follows a live scene, and $P(announcer|live)$ represents the probability that the live scene has the "announcer" as its speaker.

Based on this BN, we can calculate the probability of each value *x* for the present scene categories as

$$P(x|e) = [\sum_{all\ b} P(x|b)P(b)] \prod_{j=1}^{|F|} P(f_j|x), \quad (1)$$

where *e* represents the values of variables on other nodes except X. For example, $P(live|e)$, the probability that the present scene category is "live" with the value $e(b, f_1 = announcer, f_2 = \cdots)$, is calculated as

$$P(live|e) = [P(live|live)P(live) + P(live|replay)P(replay) + \cdots] \times P(announcer|live) \times \cdots. \quad (2)$$

Bearing those discussed above in mind, we categorize each CC segment into the scene categories as follows.

[Procedure to categorize CC segment]

1) Calculate the conditional probability distributions(CP table) for every transition from the sample CC text.
2) Input the features of a CC segment and the scene category of the previous CC segment, then calculate the probability of each scene category for the segment based on the generated CP table and determine its scene category as the one which has the maximum value.
3) Since the present scene category is probably the previous scene category in the next step, update the $P(b)$ with the calculated values in step 2), and repeat 2) and 3) for the rest of the segments.

After categorizing all CC segments, the *story unit* can be detected by identifying sequences between a live segment and the next live segment. Note that the live scenes can sometimes occur successively without any other in-between scenes. When there are consecutive live segments, we consider them as the consecutive separate live scenes if they have an interval more than a threshold, Th seconds, between themselves. Otherwise, we determine them as included in the same live scene.

4 Experimental Results

We experimented our method using 10 broadcasted American football videos with 7 different announcers and 8 different commentators. They also differs in the production company and the created year: FOX (1997×1,1998×1, and 2000×3), abc(1998×1 and 1999×1), and CBS(1999×3).

We used every one of these streams as a test stream learning from the other 9 streams and evaluated the results with the accuracy defined as below.

$$Accuracy = \frac{\sharp\, of\ correctly\ categorized\ segments}{\sharp\, of\ all\ the\ test\ data\ segments}$$

Moreover, the results of categorization to each category are evaluated in terms of the CC category recall rate and the CC category precision rate defined as below.

$$CC\ Category\ Recall = \frac{\sharp\, of\ correctly\ categorized\ segments}{\sharp\, of\ the\ actual\ segments\ of\ the\ corresponding\ category}$$

$$CC\ Category\ Precision = \frac{\sharp\, of\ correctly\ categorized\ segments}{\sharp\, of\ segments\ categorized\ as\ the\ corresponding\ category}$$

These experiments demonstrated that

- The accuracy was about 60% on the average. The performance difference among the video streams seems mostly due to the errors in the CC text regardless of the production company or the speakers. Especially, the confusion about “Others” and “CM” were caused by the most common errors in the CC text, which is the missing information about the speaker change. Speaker recognition from the audio stream will help us to improve the results.
- They tend to make less errors in the “Live” segments because of the simpleness of the sentences in the scene, and moreover, the live scene seemed to have the most distinctive characteristics of the four kinds of scenes. Therefore, among the scene categories, “Live” were most successfully categorized with the recall rate 78% and the precision rate 69% on the average. Since “Live” and “Replay” are the significant segments for the semantic information, our segmentation method is considered effective for that matter.

We also detected the story segments from the CC segmentation results with $Th = 5$. We evaluated the results with the story unit recall and the story unit precision rate which are calculated as:

$$Story\ Unit\ Recall = \frac{\sharp\ of\ correctly\ extracted\ segments}{\sharp\ of\ segments\ to\ be\ extracted}$$

$$Story\ Unit\ Precision = \frac{\sharp\ of\ correctly\ extracted segments}{\sharp\ of\ all\ the\ extracted\ segments}.$$

The segmentation results sometimes shifted slightly from the actual story breaks. However, most of these gaps between the segmented break and the actual one were within a few segments, which seems allowable for the purpose of the extraction of the semantic content, and we evaluated the results accepting the errors within 1 segment. Since the "Live" segments play an important role in the story segmentation, the story segmentation worked relatively better than the scene categorization with the recall rate 87% and the precision rate 76% on the average.

In addition, we have done the preliminary experiments of integrating our CC segmentation method and the live video segment extraction method with the image analysis proposed in [10]. For details of the association method, see [11]. The results demonstrated that each result was mutually useful in improving the other result, i.e. supplement of the insufficient CC story units and elimination of both the excessive CC and video story units, and we were successfully able to attach the segmented CC story units to the segmented video story units with the recall rate 89% and the precision rate 92% on the average.

5 Discussion

In this section, we discuss the generality and the potentiality of our method. Our CC segmentation method is based on the general structure of a sports TV program and a sports game. That is, the sports TV programs generally consist of the four kinds of the scenes discussed in Sections 2 and 3: live, replay, others, and CM, and since a play is defined as an event which occurs in a live scene, the time interval between a live scene and the next live scene can be defined as a story unit which corresponds to a play, and each story unit is generally the element of a sports game. Moreover, the characteristics of these units also discussed in Sections 2 and 3 are considered general to many kinds of sports. Therefore, although we have only tested our method with American Football videos, it seems possible to cope with several sports which can be considered as a sequence of the story units, which themselves are constructed with the scene units.

Nitta et al. [10] also tried to extract the text "Live" segments using some keywords for every kind of plays. However, they had to predefine the kind of plays and the keywords for each kind of sports. However, this segmentation method uses only the players' names and situational phrases as the characteristics of the CC story units, and therefore, it will be easier to apply it to other sports.

Our method not only segments the CC text into the corresponding story units but also segments the story units into the scene units, part of which – "Live" and "Replay" scenes – have the significant information about the story of the game. Though our segmentation method only tried to structurize the video as a repetition of story units, the results could support more practical use. For example, keyword extraction from those important scene units will help us get the information about each story unit such as the player's name, the game situation, and so on. Furthermore, since our method classifies the CC segments on the basis of the importance for the story, selecting the segments according to their importance will enable us to make the sports speech summarization. Note that, as we implied in the end of Section 4, the segmented CC story units can be properly applied to the video segmentation results with image analysis. Thus, the extracted information from these CC story units also can be easily attached to the video story units as the indexing.

6 Conclusion

This paper addressed the general structure of sports videos and proposed a method to segment the speech transcript called the closed-caption text into the semantic units which can be the effective source for extraction of the semantic content of the video. As a result of the experiments, we were able to segment the closed-caption text into the story units with the recall rate 87% and the precision rate 76% on the average. Since the preliminary experiments of the association with the segmented video story units attached the segmented CC story units to the corresponding video segments with good success, we consider the indexing of the video should be possible simply by extracting the semantic information from each generated CC story unit. Though we experimented our method only with American Football video, we consider the method is easy to be applied to other sports videos, since it does not need much domain-specific information. For future work, we will apply our method to other sports videos and also try to extract the semantic content of each story unit from the text stream for the automatic indexing of sports videos.

References

[1] D.Zhong, and S.F.Chang, "Structure Analysis of Sports Video Using Domain Models", *Proc. IEEE ICME'01*, pp.920-923, Aug. 2001.

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

[3] P.Xu, L.Xie, S.F.Chang, A.Divakaran, A.Vetro, and H.Sun, "Algorithms and System for Segmentation and Structure analysis in Soccer Video", *Proc. IEEE ICME'01*, pp.928-931, Aug.2001.

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

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

[6] 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.

[7] S.Takao, T.Haru, and Y.Ariki, "Summarization of News Speech with Unknown Topic Boundary", *CD-ROM Proc. ICME2001*, Aug. 2001.

[8] W.Zhu, C.Toklu, and S-P.Liou, "Automatic News Video Segmentation and Categorization Based on Closed-Captioned Text", ISIS technical report series, Vol 2001-20, Dec. 2001.

[9] W.Greiff, A.Morgan, R.Fish, M.Richards, and A.Kundu, "Fine-Grained Hidden Markov Modeling for Broadcast-News Story Segmentation", *Proc. ACM Multimedia'01*, 2001.

[10] 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.

[11] N.Nitta, N.Babaguchi, and T.Kitahashi, "Story Based Representation for Broadcasted Sports Video and Automatic Story Segmentation", *Proc. ICME'02*, pp.813-816, 2002.

[12] R.O.Duda, P.E.Hart, and D.G.Stork,"Pattern Classification", A Wiley-Interscience Publication.