Story Segmentation of Broadcasted Sports Videos with Intermodal Collaboration

Naoko Nitta, *Member, IEEE* and Noboru Babaguchi, *Member, IEEE*

This work was supported in part by a Grant-in-Aid for the Japan Society for the Promotion of Science Fellows.

N. Nitta is with the Department of Communication Engineering, Osaka University, Osaka 565-0871, Japan (e-mail: naoko@nanase.comm.eng.osaka-u.ac.jp).

N. Babaguchi is with the Department of Communication Engineering, Osaka University, Osaka 565-0871, Japan (e-mail: babaguchi@comm.eng.osaka-u.ac.jp).

Abstract

This paper investigates the problem of efficiently describing broadcasted sports videos for effective multimedia applications. Considering the sports videos as a sequence of recurrent semantic *story units*, we propose a method for segmenting the sports videos into the story units and attaching the closed-caption segments, which correspond to the story units, as the detailed descriptions. This proposed method restricts the use of much domain-dependent information and can be used to acquire the semantic content. We first try to segment the closed-caption text into *scene units*, a set of which comprises 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 closed-caption text. Template matching in the image stream also segments the video stream into video story units. Finally, temporal association attaches the appropriate closed-caption story unit, which includes the detailed information about semantic content, to each segmented video story unit. We conduct experiments using American football and baseball videos, obtaining successful story segmentation results, a recall rate of 92.5% and a precision rate of 91.5%, and we also discuss the potentiality for utilizing them for a video retrieval system.

Keywords

video content analysis, broadcasted sports video, story segmentation, closed-caption, intermodal collaboration

I. Introduction

A continuous increase in the amount of *unstructured* multimedia data strongly requires a framework of simple but meaningful representation that enables efficient multimedia retrieval as well as a filtering system, in which multimedia databases can be searched with queries on the basis of semantic content. A primary concern of any video retrieval system is that a query be natural and easy to formulate for users. A text-based search is the first step in the searching stage; it serves as a straightforward and fast search filter. It is also important that the text descriptions of the attributes accurately reflect the characteristics of non-text multimedia data types to a certain extent and with the best capabilities.

As a scheme to realize the text descriptions, the MPEG-7 [1], formally known as Multimedia Content Description Interface, became an international standard for describing multimedia data. The MPEG-7 allows descriptions of audio-visual content at different perceptual and semantic levels. However, it does not specify what kinds of descriptions are

needed for a specific task or how the descriptors/users obtain these descriptions. Therefore, aiming for effective retrieval or filtering systems, we need to determine 1)how to represent multimedia data in a clear and concise way, and 2)how to automatically or semi-automatically acquire the needed descriptions of the semantic content. As shown in Fig. 1, the aim of this paper is to develop a system to assist the manual description of the multimedia data with semi-automatic semantic content analysis.

Broadcasted videos (hereafter simply called videos) are divided into genres according to their semantic content, such as news, dramas, documentaries, movies, and sports, and each has some typical semantic structures dependent upon its genre. For instance, a news video can be seen as a sequence of scenes beginning with an image frame presenting the anchor person followed by the variety of the news [2], [3], [4], [5]; a drama video or movie as an assembly of the semantically interrelated scenes [6], [7], [8], [9], [10]; and a sports video as a repetition of the play and break scenes [11], [12], [13]. As indicated, the temporal video segments, each of which corresponds to a single scene, are semantically and temporally related to one another and their assembly constructs every video giving it the semantic meaning. Therefore, structurizing videos according to their genres must be done as a step to understand their meaning. In this paper, focusing on the sports video, we first propose a semantic description model based on the typical semantic structures common to diverse sports.

We next discuss how the semi-automatic semantic content analysis for the sports videos can be accomplished in order to generate the proposed descriptions when given raw unstructured videos. The problem of semantic content analysis can be divided into two steps: 1)temporal video segmentation and 2)semantic content acquisition. While videos have several sources of information such as image, text, and audio streams, the image stream is mainly used in temporal video segmentation [2], [3], [7], [8], [9], [11], [12], [13], since its low level features, such as the color difference between adjacent frames or shots, help us find the content boundaries of the video. The image stream is also used in semantic content acquisition [14], [15], [16], although the audio or text stream can be a more effective source for acquiring detailed semantic content [17], [18], [19], [20]. Obviously, each of the streams gives us only a limited amount of information. Hence, semantic content analysis

should be accomplished by combining each result acquired from multimodal information streams. We call this strategy *intermodal collaboration* [19], [20]. Most of the research on sports videos has played its focus on image analysis and analyzed other streams only for a complementary use. In this paper, we attempt to integrate the image and the text stream called *closed-caption text*, which is the speech transcript of the announcers, taking advantage of each stream and making up for the weak point mutually.

As a way to achieve the semantic content analysis, we propose a method of segmenting videos according to the semantic structure of sports videos and attaching the semantically corresponding closed-caption segment to each video segment as a document that includes detailed information. Since the image stream and the closed-caption text are not necessarily synchronized either physically or semantically, they should be segmented separately considering the semantic structures of each stream. Association of the two streams after the segmentation of each stream should synchronize them more semantically. Our method exploits the superficial features thereby avoiding the use of many keywords or key phrases which will complicate the versatility of the method and segments the closed-caption text into the semantic units of the sports program called *scene units*, parts of which have information necessary to grasp the story, on the basis of the probabilistic Bayesian Network framework. Since a set of these scene units constructs the unit of the sports game called *story unit*, finding the boundary of the sets leads us to segment the closed-caption text into *story units*; that is, the semantic units of the sports game. Next, focusing on some image cues which are acquired with the knowledge of the sports videos, we also try to extract the story units with template matching in the image stream. Finally, after temporal synchronization of the results acquired from both streams, extracting only the significant scene units from each story unit of closed-caption text attaches the detailed text description to each video story unit. Note that, hereafter, we call it *story segmentation* to temporally segment a video into story units.

The paper is organized as follows. Section II introduces recent related work and its relationship to this research. Section III discusses the structure of broadcasted sports videos and proposes a semantic description model suitable for sports videos. Section IV proposes a mthod to generate the descriptions proposed in Section II, integrating the text and the

image streams. In Section V, experiments are conducted to demonstrate the effectiveness of the proposed method. Section VI discusses the effectiveness and potentiality of the method. Section VII concludes this paper and gives the future work.

II. Related Work

Let us first introduce some related work. First of all, there are several researchers working on temporal video segmentation. Most of the research with news videos tried to semantically segment such videos with the detection of anchor shots, since the configuration of the anchor-person frame obeys a certain spatial structure [2], [3]. For movies or drama videos, Hanjalic et al. [7], Kwon et al. [8], and Javed et al. [9] attempted to segment the videos into semantically related scenes based on the visual similarity of each shot. For sports videos, Zhong et al. [11] and Li et al. [12] tried to extract patterned event boundaries from the image stream, and Xu et al. [13] also tried to segment a soccer video into play/break scenes with frame-based image analysis. Note that these systems used the image stream to realize the temporal video segmentation and the attainable semantic content is limited to the kind of scenes such as "anchor scene" and "play scene".

More detailed semantic content acquisition has been accomplished by detecting object, motion, and event with visual features. For example, for sports videos, Zhou et al. [14], Gong et al. [15], and Sudhir et al. [16] respectively proposed a method of classifying the shots into 9 classes such as Left/Right Offense and Left/Right Scores for basketball; into 15 classes such as Left/Right Penalty Area and Midfield for soccer; and into 4 classes, Baseline-rallies, Passing-shot, Serve-and-Volley, and Net-game for tennis with line detection, player/ball motion detection, and court/field color detection from the image stream.

Although the research discussed above is based on only the image stream, the visual features cannot always be easily mapped into semantic concepts. Therefore, the text stream and the audio stream, which can be important sources of semantic information and are computationally much cheaper to analyze, have been the next major targets for semantic content analysis. For sports videos, Lazarescu et al. [17] tried to make annotations about the movement of the player by searching keywords from the text stream and analyzing the image stream. Chang et al. [18] tried to detect important events by integrating the audio and the image streams. Babaguchi et al. [19] proposed event scene detection by

integrating text, audio, and image streams. As you can see, more detailed semantic content can be acquired with intermodal collaboration. However, what these methods were able to describe was limited to the scenes where special events such as the score events occur, and other scenes or other information such as players have been neglected. Nitta et al. [20] attempted to annotate about plays and players for every play scenes; however, their method useed too many domain-dependent key phrases to be applied to several kinds of sports.

Placing more focus on text analysis for acquiring semantic video content, Shahraray et al. [21] proposed an automatic authoring method of hypermedia documents for news videos by segmenting the CC text correspondingly to the video units. Takao et al. [22] tried to find the topic boundary of the news speech and to summarize the speech using TF-IDF with the speech transcript. Greiff et al. [23] used the Hidden Markov Model associated with the parameters which reflected the occurrence of words for segmentation of news videos. They all used the characteristics of word occurrence for each topic or topic boundary. Due to the relative uniformity of the topics for the sports videos, however, few researchers have succeeded in semantic topic segmentation of the CC text of sports videos. Considering the discussion above, we try to acquire more semantic content by achieving semantic topic segmentation of the CC text of sports videos without much domain-dependent information.

III. A Semantic Description Model for Sports Videos

The sports video has two kinds of structures according to different points of view: sports TV program and sports game. In this section, we discuss both structures and present a semantic description model for sports videos summarizing these different structures.

(i)Structure of a Sports Program

A TV program of a sports game can be regarded as a sequence of specific scenes. We define a "Live" scene as the time interval beginning with the players' first move and ending with an event such as a goal being scored or the players or a ball leaving the field. The "Replay" scenes can be defined as the scenes where detailed explanation of "Live" scenes is given. Other scenes such as "Report," "Spectators," and "CM (Commercial Message)" are considered semantically unrelated to the game. We call

these logical elements which construct a sports program *scene units*, and also regard the time interval between one "Live" scene unit and the next "Live" scene unit as a *story unit*, which is the logical element of the whole story. Note that the "Live" scenes usually start with some characteristic images and end with other kinds of images.

(ii)Structure of a Sports Game

A sports game can be mostly expressed as a tree. When considered in a form of a tree, a sports game can be considered as a sequence of fundamental elements of the tree. These fundamental elements correspond to the *story units*, which were discussed in the "Structure of a Sports Program," and can be viewed as a basic logical unit describing a sports game. Therefore, the sub-story in each unit constitutes the whole story of the game. The information needed to explain the sub-story is the upper node of the tree (which is composed of several story units; e.g., 1st–4th Quarter, offense team name, and 1st–4th Down for American football; 1st-9th inning and top/bottom for baseball) and about the unit itself (the attributes such as play, player, time-in-game, score). In general, the scene descriptions basically should indicate 5Ws1H, which are the WHEN, WHERE, WHY, WHAT, WHO, and HOW to satisfy this requirements. The information discussed above will satisfy the parts of the 5Ws1H requirement.

Considering these two structures, a sports video is regarded as a sequence of the story units, each of which is constructed with several scene units, starting with the "Live" scene. Indexing each unit according to its semantic content will help us understand the whole story of a video. Fig. 2 shows the overall structure of a sports video we try to construct.

IV. Story Segmentation

We have proposed a semantic description model for sports videos in Section 2. Here, we face another problem: How do we obtain the information necessary for the descriptions? Obviously, it is extremely time-consuming to manually generate entire descriptions for each video. Therefore, some systematic way to acquire the needed information should be developed to reduce the labor involved in manual generation of the descriptions.

Many researchers have tackled this problem, *Automatic Indexing/Annotation*, which contains two main sub-problems: *Temporal Video Segmentation* and *Semantic Content*

Acquisition; that is, the acquisition of semantic information such as objects, motions, and events from each video segment. Although videos can be analyzed with several information streams, each of these streams has its own advantage and disadvantage. While the image stream is the most reliable information source for temporal video segmentation, semantic content acquisition requires more high-level content analysis and more researchers are paying attention to the text/audio stream (the superimposed text, speech, closed-caption text, etc.) because such streams include more semantic information and are much less costly to process. Moreover, since the announcers talk about the situation of the game in a sports game program, we have a good chance of acquiring information about the game from the commentary. Therefore, we propose a method of segmenting the commentary of sports videos into semantic segments and associating the segmented commentary to the corresponding video segments as an annotation to videos. Here, we use the text stream called Closed-Caption (CC) text, which is a transcript of commentary and sound.

The developed system takes a raw video as the input and outputs the semantic MPEG-7 descriptions. Fig. 3 shows the outline of our proposed method. Given a video stream which has CC text and image stream, the system first analyzes each stream separately. For the CC text, in order to make applying this method to several kinds of sports easier, we restrict the use of the domain-dependent key phrases and, with mostly the features common among many kinds of sports videos, probabilistically segment the CC text into both the scene units and the story units with a Bayesian Network [24]. The video stream is segmented into the story units with template matching in the image stream. Then the segmented video story units are associated with the segmented story units in the CC text. The details of each step are discussed below.

A. Segmentation of Text Streams

Semantically synchronizing the CC text and video stream is not such a simple task, since the CC text usually lags approximately 0 to 20 seconds behind the actually spoken words in sports videos. Moreover, the announcers do not necessarily talk about the present scene. Therefore, we try to segment the CC text separately from the video stream so that we can realize the semantic synchronization between the CC text and video stream with segment-to-segment association, rather than with word-to-word association.

Here, let us consider the semantic element in the CC text. The CC text is just a sequence of words and does not have prominent indicators of scene changes. However, the change of the speaker, which is marked with "NAME:" (NAME indicates the name of the speaker) in the CC text, can be a boundary of the topic. Though if the same speaker talks throughout several kinds of scenes, there is no pronounced marking. By calculating the time interval between the sentences, therefore, we also consider the blank portion of the speech as the boundary of the CC text. Here, we define the segment between the boundaries as a *CC segment*.

Each CC segment is supposed to belong to one of the scene categories. Based on the structure of the sports TV program, the four scene categories we attempt to categorize are "Live," "Replay," "CM," and "Others" ("Report," "Spectators," etc.). Since the content of the talk in each scene category depends on the speaker, the production company, etc., more general characteristics for each scene category are considered in TABLE I. The "Speakers" column shows the speakers who usually talk in each scene. The "Length of Sentences" and the "♯ of Sentences" respectively, show the general length and the number of the sentences in each scene. For example, in live scenes, since the announcers usually make simple comments about the on-going play, the number of sentences tends to be few, and the length of the sentences tends to be short. "Players' Names," "Situational Phrases," and "Numbers" respectively, show the likeliness of the appearance of the players' names, *situational phrases*, and numbers (which possibly represent the score, yards, etc.). Here, the situational phrases can be defined 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.

Since the characteristics discussed above are highly ambiguous in categorizing CC segments with their corresponding scenes, more precise patterns in the CC text for each scene should be learned. Here, based on the characteristics above, we extract 6 features for each CC segment: **the name of the speaker** (announcer, commentator, referee, and others), **the number of sentences**, **the length of sentences**, **the appearance of the players' names**, **the appearance of the situational phrases**, and **the appearance of the numbers** in order to categorize the CC segments into four kinds of scenes. Note that

the names of the announcer, commentator, players, and referee are given beforehand.

Moreover, the structure of the sports TV program shows that the scenes have some rules in how they line up, that is, the scene category of a CC segment depends on the scene category of the previous CC segment as well as on its own features. In our method, to tackle the uncertainty of the information, we use a probabilistic framework which can handle such information, namely, the Bayesian Network (BN), in order to categorize each CC segment.

The BN in Fig. 4 shows the relationship between the scene category of the present CC segment and other factors. Node C_B represents the scene category of the previous CC segment and is the parent of node C_X, which represents the scene category of the present CC segment. Nodes F_j, which are the jth children nodes of C_X, represent the features of the present CC segment. $P(c_x|c_b)$ and $P(f_j|c_x)$ represent the probability of the transitions, where c_x represents the values of variables on nodes C_X, and c_b and f_j 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 for the present scene categories as

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

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

$$\begin{aligned} P(live|e) &= [P(live|live)P(live) + P(live|replay)P(replay) + \cdots] \\ &\quad \times P(announcer|live) \times \cdots. \end{aligned} \quad (2)$$

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

[Procedure to categorize CC segment]

1) Learn the conditional probability distributions for every arc of the BN from the sample CC text data.

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 present CC segment and determine the scene as the one which has the maximum value.
3) Update the $P(b)$ with the calculated values in step 2), and then repeat 2) and 3) for the next CC segment.

After categorizing all CC segments, the *CC 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 consecutive CC segments have been determined as live segments, we consider those with an interval more than a threshold between them as consecutive separate live scenes and those with less interval as the ones in the same live scene.

B. Image Stream Analysis

In order to synchronize the CC text and the video stream, the video stream should be segmented into the corresponding story units. Generally speaking, since players usually take their stances at the beginning of each live scene of the game, we can often see the stationary images which are captured by cameras positioned at the fixed locations at that time. For example, American football has 1)an image of players of each team lining up face-to-face at a standstill for a while, which is taken horizontally to the lines; 2)the same scene which is taken vertically to the lines; 3)an image of players lining up before the goal line, which is taken from the end of the field; and 4)an image of players lining up at the end of the field taken as to show the whole field, as the beginning images of live scenes. Most sports have these kinds of characteristic beginning images, and the images are common throughout videos for the same kind of sports. Therefore, finding these images from the image stream of each video enables us to find the beginnings of live scenes, which are also the beginnings of the story units, and, consequently, to segment the video stream into *video story units*. The method is realized with the template-matching of color histograms of initial frames of each shot with those of example images which are given beforehand. Refer to [20] for the details of this method.

C. Association of Text and Video Streams

Finally, we temporally associate the segmented CC story units with the video story units. The CC text and video stream have some time lag between them. However, since the CC text has already been segmented according to the semantic content, the association can be achieved just by finding the CC story unit temporally closest to each video story unit. Note that the approximate beginning time of each CC story unit can be calculated from the frame number of the image in which the initial character is embedded. Moreover, it should be noted that the "Live" scene usually occurs during a certain interval. Therefore, consecutively extracted beginnings of the "Live" scene from the video stream should not exist too closely to each other.

Considering the characteristics mentioned above, we associate the CC story units with the video story units as follows:

[Procedure to associate CC and Video Stream]

1) Add Th_1 seconds to the beginning of each video story unit to fill the time lag between the CC text and the video stream.
2) For every video story unit,
   2-a) Search the CC story unit whose beginning is closest to the beginning of the current video story unit.
   2-b) If there is no CC story unit to associate, search a CC segment whose beginning is closest to the beginning of the current video story unit. Then, after dividing the CC story unit including the found CC segment at the beginning of the found CC segment, associate the generated CC story unit to the current video story unit.
   2-c) If there are consecutive video story units which have less than Th_2 seconds interval between them, leave only the one closer to corresponding CC story unit, and discard the other one.
3) After associating all the video story units to CC story units, merge the remaining CC story units with their previous CC story units.

Fig. 5 illustrates the behavior of this procedure.

V. Experimental Results

We have experimented with 10 broadcasted American football videos (Video1-Video10) and 5 baseball videos (VideoI-VideoV) by extracting 20 minutes from each video stream to see the difference in effectiveness between different kinds of sports. We also selected videos which have different announcers and were produced by different companies in different years for each kind of sport to see if the method works the same for several kinds of videos for the same sport.

Each of the results of the CC scene categorization, the CC story unit generation, the video story segmentation, and the association of the two streams are shown below. Comparing the results after the association and those of the analysis of each stream shows the effectiveness of the integration of different streams.

Here, the results of the CC scene categorization are evaluated in terms of the accuracy as defined below.

$$Accuracy = \frac{\sharp\ of\ correctly\ categorized\ CC\ segments}{total\ \sharp\ of\ CC\ segments\ in\ test\ data}$$

Since the extraction of the "Live" scene units is most important for the story segmentation afterwards, we also evaluate the results of the "Live" scene categorization in terms of the CC Live scene recall rate and the CC Live scene precision rate as defined below.

$$CC\ Live\ Scene\ Recall = \frac{\sharp\ of\ correctly\ categorized\ ``Live''\ CC\ segments}{\sharp\ of\ the\ actual\ ``Live''\ CC\ segments}$$

$$CC\ Live\ Scene\ Precision = \frac{\sharp\ of\ correctly\ categorized\ ``Live''\ CC\ segments}{\sharp\ of\ the\ CC\ segments\ categorized\ as\ a ``Live''\ scene}$$

We also evaluate the results of detecting the CC/Video story units with the CC/Video story unit precision and the CC/Video story unit recall rate which are calculated as:

$$CC/Video\ Story\ Unit\ Recall = \frac{\sharp\ of\ correctly\ segmented\ CC/Video\ story\ units}{\sharp\ of\ the\ actual\ CC/Video\ story\ units}$$

$$CC/Video\ Story\ Unit\ Precision = \frac{\sharp\ of\ correctly\ segmented\ CC/Video\ story\ units}{\sharp\ of\ all\ the\ segmented\ CC/Video\ story\ units}$$

A. Experiments with American Football Videos

TABLE II shows the production company, the production year, and the names of the speakers (the announcer and the commentators) for each of 10 video streams.

The CC text was segmented with a speaker change and at the beginning of every blank portion that lasts more than 3 seconds. The situational phrase for American football is "num_1 down (and num_2)," where num_1 represents either of "1st," "2nd," "3rd," or "4th," and num_2 represents any number.

Here, after learning the patterns of each scene from the CC texts of 9 videos, we used the learned data to categorize CC segments of the remaining 1 video (cross validation). Although the learned probability for Bayesian Networks varied depending on the sample video stream, the overall tendency is shown in TABLE III. Here, "Transition" compares the probability of transition from one scene to another scene; "Speaker" shows who is likely to speak in each scene; and "Players' names," "Situational phrases," and "Numbers" respectively compare the probability of the appearance of players' names, situational phrases, and numbers.

On the other hand, the apparent patterns for sentences are:

- CMs tend to have many sentences, most of which are short. Live scenes also have many short sentences, while in replay scenes, long and short sentences are somewhat evenly distributed. Others tend to have only a single sentence, which is often relatively longer then those in other scenes.
- Since the CC text was segmented with blank-portion among the speech as well as the speaker change, a single scene was divided into several CC segments, which had at most only a couple of sentences for all kinds of scenes. Therefore, in order to acquire the patterns in duration of each scene, the segmentation should be done with longer blank portion. However, with the longer blank, several scenes can be included in a single CC segment, which makes categorization difficult. Therefore, we have to set the length of blank portion carefully to succeed in categorization.

Overall, these results showed us the expected tendency for each scene. According to the results, however, the main speakers in CMs depend on the video stream and were often erroneously learned as the announcer. This discrepancy is due to the errors in the CC text, the omission of the speaker change in the transition from programs to CMs. Moreover, since the features used in this experiment are not necessarily the best selected ones, more experiments with other features should be conducted in the future.

We show the results of the CC scene categorization for each video in TABLE IV. These results indicate that:

- The accuracy was 59% on average ranging from 50% to 66%. For example, Video6, Video7 and Video8 were produced by the same production company, in the same year, with the same announcer/commentators but differed in the results of CC scene categorization. That is, the differences in the accuracy among the videos can be inferred to be caused by not the differences in the way the videos were produced, but by the errors in the CC text. The most common error found in the CC text was the omission of the speaker changes at the time when the program scene changes to the CM scene. Consequently, "CMs" were indistinguishable from the scenes in the program and were often confused with the "Others."
- The "Live" CC scene units usually have few errors, since the information in the scenes is important for the viewer, and additionally, the announcers tend to make simple short comments. Therefore, among the scene categories, the "Live" was most successfully categorized with a CC live scene recall rate of 76% and a CC live scene precision rate of 69%.

TABLE V shows the results of CC story unit generation. The segmentation sometimes shifts slightly from the actual story boundaries. However, we have segmented the CC text to acquire semantic information, and from this point of view, we do not have to be so strict about the location of the boundaries. Therefore, we evaluated the results allowing for shifts up to 1 segment.

A comparison of TABLE IV and TABLE V shows that the results of the CC story unit generation were obviously better than those of the CC scene categorization. When a "Live" scene consists of several CC segments, features such as the situational phrases and the players' names often appear only in some of the corresponding CC segments. Consequently, only a part of the CC segments in a "Live" scene could be categorized as "Live," and the others could be categorized as "Other" scenes. Although the erroneously categorized CC segments deteriorated the results of the CC scene categorization, with the CC segments correctly categorized as "Live," the CC story units were able to be correctly generated.

We next experimented with the video segmentation method. The sampling rate for the video was six frames per second, and we provided four kinds of images shown in Fig. 6 as the beginning images for each video stream. We selected the example images considered to be taken from up front of the field as the general camera direction. TABLE VI shows the results of the segmentation. We compared the initial 6 frames of each shot with the example images and determined the shots with more than 5 similar frames as the beginning shots of video story units.

TABLE VII shows the results of the association of the CC text and the video stream with the parameters $Th_1 = 10$ and $Th_2 = 9$. The "Shifted VSU (Video Story Units)" column shows the number of the extracted video story units which are temporally shifted from the actual video story units, and the "Shifted CSU (CC Story Units)" shows the number of the extracted CC story units whose beginnings are shifted from the actual CC story units. The "Discarded VSU" column represents (♯ of discarded video story units / ♯ of excessive video story units extracted with the video segmentation), the "Discarded CSU" column represents (♯ of discarded CC story units / ♯ of excessive CC story units generated with CC story unit generation), and the "Generated CSU" represents (♯ of added CC story units / ♯ of the insufficient extractions from CC story unit generation).

These results imply the following information:

- The live scenes sometimes include other kinds of shot in the midst of the beginning images. In this case, we extracted both shots on either side of the insignificant shot and considered the first one a false detection. As a consequence of association with the CC text, however, we were able to discard the unnecessary extractions (See "Discarded VSU" in TABLE VII). Moreover, association with the most appropriate CC story units prevented the false deletion of the correct video story units and, as a result, the precision rate improved without degrading the recall rate.
- As the "Discarded CSU" and "Generated CSU" columns indicate, the shifted CC story units generated with the CC story unit generation can be discarded or changed to the correct ones as a result of the association which considers the beginning time of the video story units.
- Since we have achieved the association based on the results of the video segmentation,

the video story units which we failed to extract in the video segmentation cannot be recovered with the association. Therefore, the recall rate in the video segmentation should be emphasized more than the precision rate.

B. Experiments with Baseball Videos

We have also experimented with 5 baseball videos using our method. All of these videos were broadcasted in 2001 by FOX. Two of them (Video I and Video II) have Thom Brennaman as the announcer and Steve Lyons as the commentator, and the other three (Video III, Video IV, and Video V) have Joe Buck as the announcer and Tim McCarver as the commentator. The situational phrases for baseball were changed to "Num_1 ball(s) and Num_2 strikes," "full count," and "Num_3 away (out)," where Num_1 represents the integer from 0 to 4, Num_2 the integer from 0 to 3, Num_3 the integer from 1 to 3. Baseball videos have two kinds of images as the characteristic beginning images of a "Live" scene, in which the pitcher posing before throwing the ball was taken from (1)the back of the pitcher and (2)the front of the pitcher so that the image shows both the pitcher and the player at the base (See Fig. 7).

TABLE VIII shows the results of CC scene categorization for each video using learned data from the other 4 video streams. As we have discussed earlier, applying the method to other kinds of sports requires changing the situational phrases, the kind of characteristic beginning images of the "Live" scene, and the sample CC text data for the CC scene categorization. Of these three, changing the sample CC text data requires much more work than the other two. However, since the sports videos generally consist of four kinds of scenes, – "Live," "Replay," "Others," and "CM" – and the characteristics of these four scenes discussed in Section 3.1 are common among many kinds of sports, the characteristics in the CC text are assumed to be similar between American football and baseball videos. Based on this assumption, we have also tested the CC scene categorization for baseball videos using data learned from American football videos. The results are shown in TABLE IX.

Comparing TABLE VIII and TABLE IX shows us that there is no significant difference between these two experiments. We can infer from the comparison that since the method uses few domain-dependent features, the sample data for one sport can be applied to other

kinds of sports without creating the sample data for each kind of sports. This fact indicates the generality of our method. Note that we used the results with the data learned from the baseball videos for the following experiments.

TABLE X, TABLE XI, and TABLE XII each shows the results of the CC story unit generation, the video segmentation, and the association of the CC text and video stream. Although these results held little difference with the results using American football videos, the differences between the two sports were the following:

- While the announcers almost always make the play-by-play commentary in every "Live" scene of American football videos, it is not necessarily the case with baseball. In baseball videos, the announcers sometimes skip the explanation of the plays which are insignificant to the story such as simple strikes and balls and talk about other unrelated subjects. In that case, there appears no "Live" scene in the CC text, and as a result, the recall rate of the CC story unit generation degraded.
- A story unit for baseball is usually shorter than that of American football. Consequently, for baseball, these rather short story units were more erroneously discarded than American football in the step of the CC-video association, confused with the falsely generated story units in the CC story unit generation, and the recall rate after the association also degraded.

However, these were minor differences when viewing the overall results, and these results also indicated the generality of our method.

VI. Discussion

Here, we discuss the effectiveness of our method compared to the prior work: annotations of plays and players with intermodal collaboration [20]. Both methods predefined information which depends on 1)the kind of sports and 2)each video stream. The method proposed in [20] requires about 100 "key phrases explaining plays" for American football and 60 of them for baseball as the information dependent on the kind of sports. Moreover, since these key phrases can often be slightly changed according to the speakers, they have to be defined by studying many sample video streams.

On the other hand, our method requires 1 "situational phrase" for American football and 3 of them for baseball as the information dependent on the kind of sports. Moreover, they

can be easily defined since these phrases are the same throughout all the video streams for the same kind of sport. These facts indicate more applicability of our method to a variety of sports.

Moreover, both methods first analyze the CC text and the image stream separately and integrate them afterwards. However, the effectiveness of the method in [20] highly depends on the result of analyzing each stream and ends up in degrading both the precision and the recall rate after the integration, since it deletes the live segments which have been correctly obtained from one stream if the corresponding segments have not been obtained from the other stream. On the other hand, our method succeeded in improving the precision rate without degrading the recall rate much, by trying not to delete the obtained segments from either stream. This fact also indicates the effectiveness of our method.

We next discuss the potentiality of our method for semantic content acquisition. Fig. 8 illustrates an example of the generated MPEG-7 descriptions. The text descriptions are the CC segments which correspond to the "Live" and "Replay" CC scene units included in the attached CC story unit. TABLE XIII shows the usability of the CC segments attached to each story unit. In TABLE XIII, "actual time" is the actual time within the video; "extracted time," the video time of the segmented video story units; "included words," the words which can be used as the semantic descriptions of the units included in the "Live" and "Replay" scene units within the associated CC story units; and "actual content," the actual semantic content of the units. In the "included words," the "player" and "situation" indicate respectively, the words corresponding to the players' names and the situational phrases used in the CC scene categorization. Moreover, we added the information about "play" by extracting the predefined general key words related to the plays for the sports ("Touchdown," "Punt," "Extrapoint," "Flag," etc. for American football). Underlined in the "included words" and "actual content" are the common key words between them. As shown in TABLE XIII, segmentation of the CC text and its association with the video stream allows us to attach semantic content descriptions in a text form.

Note that the attached text descriptions are much more redundant compared to the ones attainable with the method in [20] (at most one play and two players for each live scene), since our method obtained them from the commentary of announcers without

any modification. However, as shown in "included words" in TABLE XIII, our method enabled us to obtain more information than [20]. In the actual retrieval system, these "included words" should work as effective queries, and searching the given words through the attached descriptions will find potential video segments. Moreover, analyzing the attached descriptions with the knowledge about the flow of a game will help us to obtain more simple descriptions and this remains as our future work.

VII. Conclusion

This paper proposed a method for segmenting the closed-caption text into scene and story units and for attaching each segment to the corresponding video segment as text semantic descriptions. As a result of the experiments with 10 American football videos, we accomplished correct video story segmentation and attached the segmented video story units to the semantically corresponding closed-caption segments with a recall rate of 92% and a precision rate of 89%, with two different information streams helping each other obtain better results than would be achieved with their individual results. Moreover, the experiments with baseball videos indicated the possibility of the applicability of our method to several kinds of sports with little additional work. We also discussed the applicability of the attached closed-caption segments to the video retrieval system; however, for more concise descriptions, a method of acquiring only the significant information from the attached closed-caption segments should be examined in the future.

References

[1] MPEG MDS Group:"Text of 15938-5 FCD Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes," ISO/IEC JTC1/SC29/WG11 MPEG01/M7009, Singapore, March 2001.

[2] X.Gao and X.Tang, "Automatic Parsing of News Video Based on Cluster Analysis," *Proc. 2000 Asia Pacific Conference on Multimedia Technology and Applications (APCMTA'00)*, 2000.

[3] J.Blumin, L.Cserey, D.Holcomb, P.Kelly, D.Nadeau, T.Nguyen, and D.Swanberg, "Towards A Personalized News Service," *Report of University of California, San Diego.*

[4] K.Shearer, C.Corai, and S.Venkatesh, "Incorporating Domain Knowledge with Video and Voice Data Analysis in News Broadcasts," *Proc. ACM International Conference on Knowledge Discovery and Data Mining (KDD'00)*, pp.46–53, 2000.

[5] S.Eickeker, and S.Muller, "Content-Based Video Indexing of TV Broadcast News Using Hidden Markov Models," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99)*, pp.2997-3000, 1999.

[6] A.A.Alatan, A.N.Akansu, and W.Wolf, "Multi-Modal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing," *Multimedia Tools and Applications*, vol.14, no.2, pp.137–151, 2001.

[7] A.Hanjalic, R.L.Lagendijk, and J.Biemond, "Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.9, No.4, pp.580–588, June, 1999.

[8] Y-M.Kwon, C-J.Song, and I-J.Kim, "A new approach for high level video structuring," *Proc. IEEE Conference Multimedia and Expo (ICME'00)*, pp.773–776, 2000.

[9] O.Javed, Z.Rasheed, and M.Shah, "A Framework for Segmentation of Talk & Game Shows," *Proc. IEEE International Conference on Computer Vision (ICCV'01)*, pp.532–537, 2001.

[10] Y.Li, W.Ming, and C-C.J.Kuo, "Semantic Video Content Abstraction Based on Multiple Cues," *Proc. IEEE International Conference on Multimedia and Expo (ICME'01)*, 2001.

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

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

[13] 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," *IEEE ICME'01*, pp.928-931, Aug. 2001.

[14] W.Zhou, A.Vellaikal, and C-C-J.Kuo, "Rule-Based Video Classification System for Basketball Video Indexing," *Proc. ACM Multimedia 2000 Workshops*, pp.213–216, 2000.

[15] Y.Gong, L.T.Sin, C.H.Chuan, H.Zhang, and M.Sakauchi, "Automatic Parsing of TV Soccer Programs," *Proc. IEEE International Conference on Multimedia Computing and Systems (ICMCS'95)*, pp.167–174, 1995.

[16] G.Sudhir, J.C.M.Lee, and A.K.Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval," *Proc. IEEE International Workshop on Content-Based Access of Image and Video Databases (CAIVD'98)*, pp.81–90, 1998.

[17] 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, June 1999.

[18] 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, June 1996.

[19] 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 2002.

[20] N.Nitta, N.Babaguchi, and T.Kitahashi, "Generating Semantic Descriptions of Broadcasted Sports Video Based on Structure of Sports Game," Multimedia Tools and Applications, Kluwer (to be published as a full paper).

[21] B.Shahraray and D.C.Gibbon, "Automated Authoring of Hypermedia Documents of Video Programs," *Proc. 3rd ACM international Conference on Multimedia (MM'95)*, pp.401–409, 1995.

[22] S.Takao, T.Haru, and Y.Ariki, "Summarization of News Speech with Unknown Topic Boundary," *Proc. IEEE Conference Multimedia and Expo (ICME'01)*, Aug. 2001.

[23] 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.

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

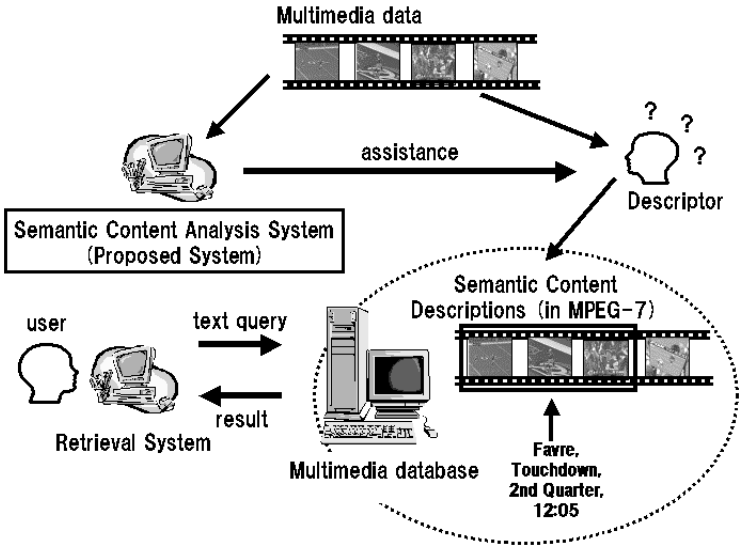


Fig. 1. Environment of our system

TABLE I
CHARACTERISTICS OF EACH SCENE IN CC TEXT

	Live	Replay	Others	CM
Speakers	Announcer	Announcer, Commentator Referee	Announcer, Commentator, Others	Others
Length of Sentences	Short	Long	cannot be determined	cannot be determined
♯ of Sentences	A few	Many	cannot be determined	cannot be determined
Players' names	likely	likely	cannot be determined	rarely
Situational Phrases	highly likely	less likely	probably	rarely
Numbers	likely	likely	cannot be determined	cannot be determined

TABLE II
AMERICAN FOOTBALL VIDEOS

Video	company	year	main speakers
Video1	abc	1997	Al Michaels, Frank Gifford, Dan Dierdorf
Video2	abc	1999	Al Michaels, Boomer Esiason
Video3	CBS	1999	Don Criqui, Brent Jones
Video4	CBS	1999	Gus Johnson, Brent Jones
Video5	CBS	1999	Greg Gumbel, Phil Simms
Video6	FOX	1997	Pat Summerall, John Madden
Video7	FOX	1998	Pat Summerall, John Madden
Video8	FOX	2000	Pat Summerall, John Madden
Video9	FOX	2000	Sam Rosen, Bill Maas
Video10	FOX	2000	Kenny Albert, Tim Green

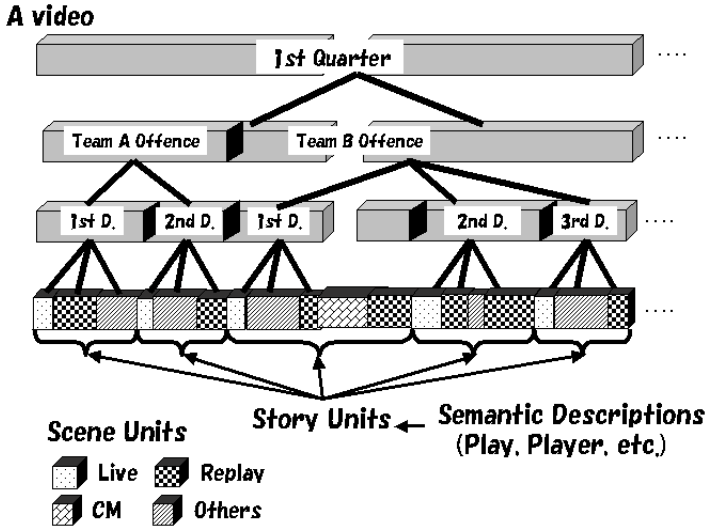


Fig. 2. Overall structure of the sports video

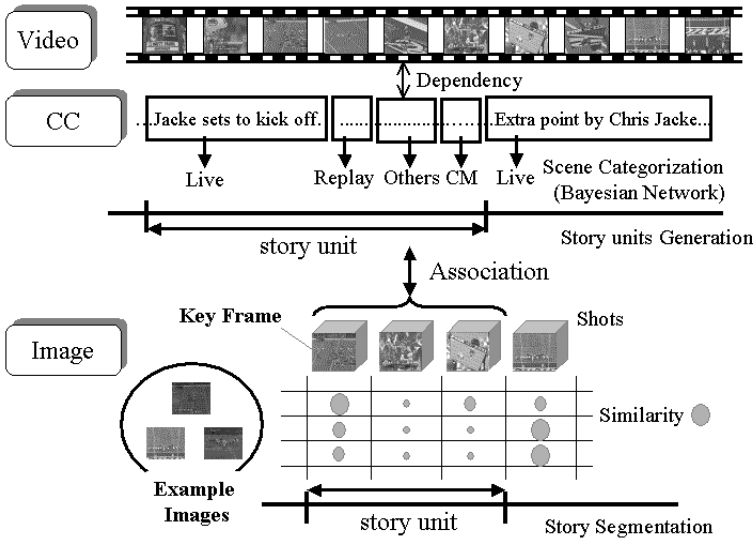


Fig. 3. Outline of the proposed method

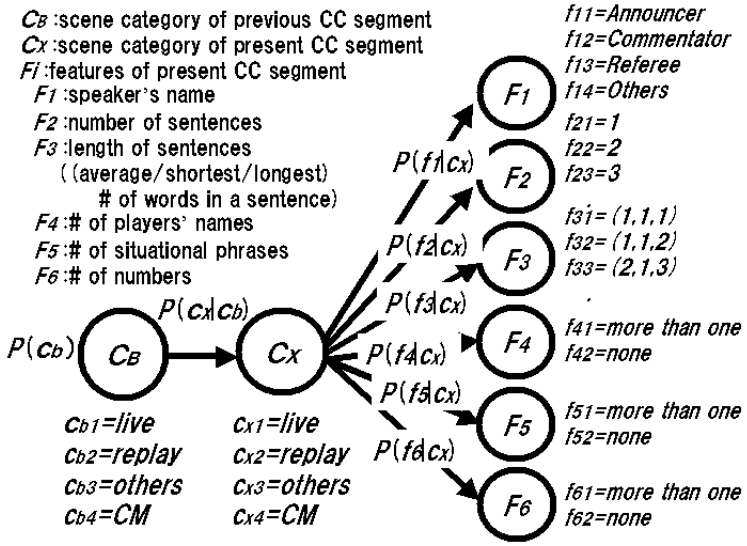


Fig. 4. Bayesian Network for categorizing CC segments

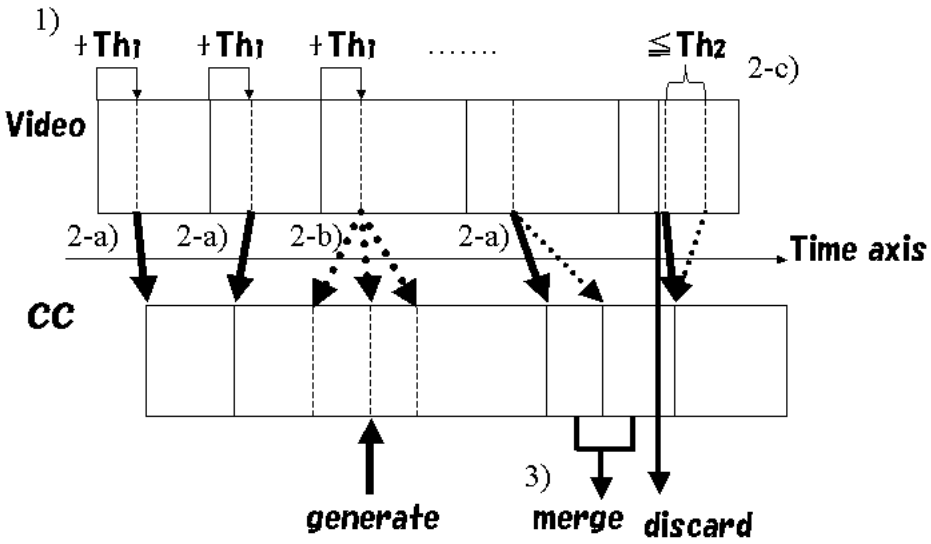
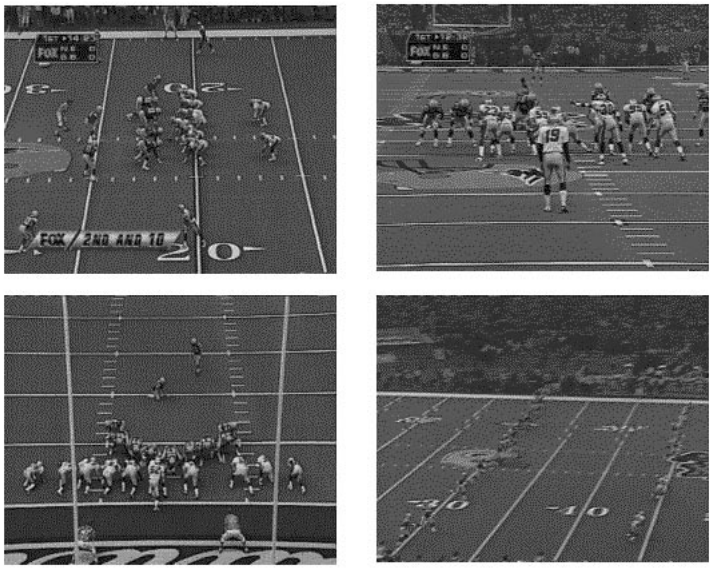
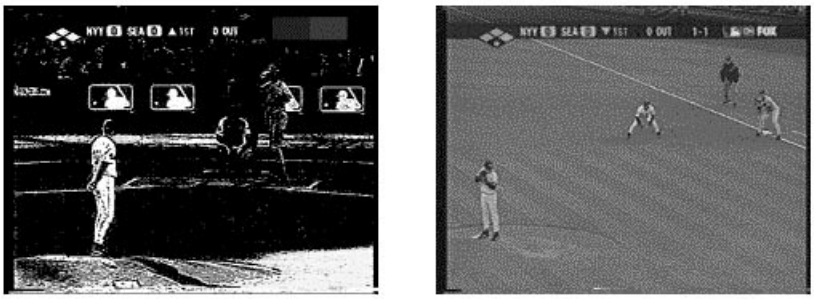


Fig. 5. Association of CC and video





(1) (2)

Fig. 7. Examples of beginning images (baseball)

TABLE III
LEARNED CHARACTERISTICS OF EACH SCENE IN CC TEXT

Transition	Live	Live $\simeq$ Replay $\simeq$ Others $>$ CM $\simeq$ 0%
	Replay	Live $\simeq$ Replay $\simeq$ Others $>$ CM $\simeq$ 0%
	Others	Live $\simeq$ Others $>$ Replay $>$ CM $\simeq$ 0%
	CM	CM $\gg$ Others $>$ Live $\simeq$ Replay $\simeq$ 0%
Speaker	Live	Announcer($\simeq$100%) $\gg$ Commentator $\simeq$ Referee $\simeq$ Others $\simeq$ 0%
	Replay	Commentator($\simeq$65%) $\gg$ Announcer($\simeq$30%) $\gg$ Referee($\simeq$5%) $>$ Others $\simeq$ 0%
	Others	Announcer($\simeq$50%) $\simeq$ Commentator($\simeq$50%) $\gg$ Referee $\simeq$ Others $\simeq$ 0%
	CM	Announcer $\gg$ Commentator $\simeq$ Referee $\simeq$ Others
Players' names		Live $\simeq$ Replay $>$ Others $\gg$ CM $\simeq$ 0%
Situational phrases		Live $>$ Replay $>$ Others $\simeq$ CM $\simeq$ 0%
Numbers		Live $\gg$ Replay $\simeq$ Others $>$ CM $\simeq$ 0%

TABLE IV
RESULTS OF CC SCENE CATEGORIZATION (AMERICAN FOOTBALL)

	accuracy	Live Recall	Live Precision
Video1	62%	79%(48/61)	79%(48/61)
Video2	57%	93%(41/44)	64%(41/64)
Video3	50%	68%(38/56)	54%(38/71)
Video4	60%	77%(51/66)	64%(51/80)
Video5	62%	78%(40/51)	67%(40/60)
Video6	66%	72%(42/58)	79%(42/53)
Video7	56%	59%(48/81)	79%(48/61)
Video8	55%	70%(28/40)	60%(28/47)
Video9	59%	98%(48/49)	66%(48/73)
Video10	62%	81%(55/68)	79%(55/70)
Total	59%	76%(439/574)	69%(439/640)

TABLE V
RESULTS OF CC STORY UNIT GENERATION (AMERICAN FOOTBALL)

	CC Story Unit Recall	CC Story Unit Precision
Video1	96%(23/25)	82%(23/28)
Video2	100%(23/23)	88%(23/26)
Video3	56%(9/16)	50%(9/18)
Video4	68%(13/19)	68%(13/19)
Video5	90%(18/20)	78%(18/23)
Video6	91% (20/22)	91%(20/22)
Video7	79%(19/24)	86%(19/22)
Video8	100%(19/19)	73%(19/26)
Video9	94%(17/18)	68%(17/25)
Video10	100%(21/21)	84%(21/25)
Total	88%(182/207)	78%(182/234)

TABLE VI

RESULTS OF VIDEO STORY SEGMENTATION (AMERICAN FOOTBALL)

	Video Story Unit Recall	Video Story Unit Precision
Video1	84%(21/25)	66%(21/32)
Video2	87%(20/23)	71%(20/28)
Video3	100%(16/16)	80%(16/20)
Video4	95%(18/19)	75%(18/24)
Video5	90%(18/20)	82%(18/22)
Video6	95%(21/22)	88%(21/24)
Video7	96%(23/24)	92%(23/25)
Video8	89%(17/19)	89%(17/19)
Video9	100%(18/18)	86%(18/21)
Video10	90%(19/21)	95%(19/20)
Total	92%(191/207)	80%(191/240)

TABLE VII

RESULTS OF CC AND VIDEO INTEGRATION (AMERICAN FOOTBALL)

	Video Story Unit Recall	Video Story Unit Precision	Shifted VSU	Shifted CSU	Discarded VSU	Discarded CSU	Generated CSU
Video1	84%(21/25)	88%(21/24)	5	0	5/11	2/5	1/2
Video2	87%(20/23)	87%(20/23)	0	0	5/8	3/3	0/0
Video3	100% (16/16)	84%(16/19)	0	3	1/4	5/9	4/7
Video4	95%(18/19)	78%(18/23)	0	2	1/6	4/6	4/6
Video5	90%(18/20)	90%(18/20)	1	1	2/4	4/5	0/2
Video6	95%(21/22)	95%(21/22)	1	0	2/3	2/2	2/2
Video7	96%(23/24)	96%(23/24)	0	6	1/2	1/3	4/5
Video8	89%(17/19)	89%(17/19)	0	1	0/2	6/7	0/0
Video9	100%(18/18)	90%(18/20)	0	0	1/3	7/8	1/1
Video10	90%(19/21)	95%(19/20)	0	1	0/1	2/4	0/0
Total	92%(191/207)	89%(191/214)	7	14	18/44	36/52	16/25

TABLE VIII

RESULTS OF CC SCENE CATEGORIZATION (BASEBALL USING LEARNED DATA FROM BASEBALL VIDEOS)

	Accuracy	Live Scene Recall	Live Scene Precision
VideoI	66%	77%(23/30)	51%(23/45)
VideoII	62%	77%(48/62)	61%(48/79)
VideoIII	61%	93%(67/72)	52%(67/129)
VideoIV	65%	83%(55/66)	49%(55/112)
VideoV	59%	70%(44/63)	59%(44/74)
Total	63%	81%(237/293)	54%(237/439)

TABLE IX

RESULTS OF CC SCENE CATEGORIZATION (BASEBALL USING LEARNED DATA FROM AMERICAN FOOTBALL VIDEOS)

	Accuracy	Live Scene Recall	Live Scene Precision
VideoI	73%	83%(25/30)	61%(23/38)
VideoII	61%	77%(48/62)	63%(48/76)
VideoIII	62%	92%(66/72)	61%(67/110)
VideoIV	66%	72%(48/66)	59%(48/81)
VideoV	54%	67%(42/63)	61%(44/72)
Total	63.2%	78%(229/293)	61%(229/377)

TABLE X
RESULTS OF CC STORY UNIT GENERATION (BASEBALL)

	CC Story Unit Recall	CC Story Unit Precision
VideoI	83%(25/30)	93%(25/27)
VideoII	81%(21/27)	84%(21/25)
VideoIII	77%(27/35)	84%(27/32)
VideoIV	83%(29/35)	74%(29/39)
VideoV	63%(22/35)	79%(22/28)
Total	77%(124/162)	82%(124/151)

TABLE XI
RESULTS OF VIDEO SEGMENTATION (BASEBALL)

	Video Story Unit Recall	Video Story Unit Precision
VideoI	100%(30/30)	86%(30/35)
VideoII	100%(27/27)	87%(27/31)
VideoIII	91%(32/35)	82%(32/39)
VideoIV	100%(35/35)	88%(35/40)
VideoV	100%(35/35)	97%(35/36)
Total	98%(159/162)	88%(159/181)

TABLE XII
RESULTS OF CC AND VIDEO INTEGRATION (BASEBALL)

	Video Story Unit Recall	Video Story Unit Precision	Shifted VSU	Shifted CSU	Discarded VSU	Discarded CSU	Generated CSU
VideoI	97%(29/30)	91%(29/32)	0	1	2/5	2/2	4/5
VideoII	93%(25/27)	96%(25/26)	1	6	3/4	1/4	2/6
VideoIII	86%(30/35)	91%(30/33)	1	5	4/7	1/5	3/8
VideoIV	94%(33/35)	92%(33/36)	0	5	2/5	5/10	3/6
VideoV	94%(33/35)	100%(33/33)	0	5	1/1	5/6	8/13
Total	93%(150/162)	94%(150/160)	2	22	12/22	14/27	20/38

```
<AudioVisualSegment id=''1std''>
    <MediaTime>
        <MediaRelTimePoint>
              T0:23:29
        </MediaRelTimePoint>
        <MediaDuration> PT50S </MediaDuration>
    </MediaTime>
    <TextAnnotation>
        <FreeTextAnnotation>
              AIKMAN STARTS A MAN IN MOTION.
              HANDS TO EMMITT SMITH,
              AND THERE IS NOTHING THERE.
              SIRAGUSA WAS THE FIRST MAN TO
              MAKE CONTACT.
              AND LET'S LOOK AT THE DEFENSE.
        </FreeTextAnnotation>
    </TextAnnotation>
</AudioVisualSegment>
```

Fig. 8. Example of the final description result

TABLE XIII
EXAMPLES OF SEMANTIC CONTENT ACQUISITION

Time	Extracted Time	Included Words	Actual Content
0:23:29–	0:23:29–	player: AIKMAN, EMMITT SMITH, SIRAGUSA	Aikman hands the ball to Emmitt Smith who is taken down by Siragusa
0:24:20–	0:24:20–	player: AIKMAN, BARRY CANTRELL, JERMANE LEWIS situation: THIRD AND SIX play: PUNT	third and six, Aikman's pass incomplete to Chris Warren
0:25:5–	0:25:5–	player: LEWIS, IZELL REESE	punt by Barry Cantrell Lewis's return taken down by Izell Reese
0:27:21–	0:27:21–	player: QUDRY ISMAIL, DILFER, PRIEST HOLMES, JAMAL LEWIS, RYAN McNEIL play: FLAG, FIVE-YARD PENALTY	Dilfer's pass complete to Priest Homes
0:27:50–			Jamal Lewis runs, flag for Ryan McNeil five-yard penalty
0:28:57–	0:28:57–	player: OGDEN, BRANDON NOBLE	Lewis came from behind Ogden stopped by Brandon Noble
0:29:36–	0:29:36–	player: DILFER, LEWIS, EDWIN MULITALO situation: SECOND AND SEVEN	second and seven Lewis runs
0:30:18–	0:30:18–	player: IZELL REESE, GREGG MYERS, DARREN WOODSON	Lewis is taken down by Izell Reese and Gregg Myers

List of Figures

List of Tables