SEMANTIC CONTENT ANALYSIS OF BROADCASTED SPORTS VIDEOS WITH INTERMODAL COLLABORATION

（インターモーダル協調による放送型スポーツ映像の意味内容解析）

Naoko Nitta

Department of Informatics and Mathematical Science

Graduate School of Engineering Science

Osaka University

Japan

January 2003

Preface

This thesis investigates the problem of efficiently describing sports videos for effective multimedia applications, and is summarized as follows:

Chapter 1 states the motivation and background of this research and its related works.

Chapter 2 discusses the problem of *what to describe*, addressing both syntactic and semantic structures of the sports videos. Considering the sports videos as a sequence of recurrent semantic *story units*, we propose a semantic description model for a sports video to represent itself based on its story.

The latter part of this thesis discusses *how to describe* the videos. The automatic generation of the descriptions requires 1)temporal segmentation of videos, and 2)semantic understanding of each segment. Since the several streams included in the video stream can be used collaboratively, this thesis proposes a method of using both the *closed-caption text*, which is the transcription of the broadcasted video speech, and the image stream, by compensating for limitations in both.

Chapter 3 proposes a method for automatically making a description of a play and its related players for each story unit as elements of the proposed semantic description model. The proposed method tries to acquire information for the descriptions from the closed-caption text by extraction of key phrases. Finding the corresponding segments of the video by means of template matching for the image stream attaches these textual descriptions

to the appropriate portion of the video. We verify the effectiveness of the method with experiments using American football and baseball videos.

Chapter 4 proposes a method for attaching the closed-caption segments, which correspond to the story units as more detailed descriptions. The proposed method restricts the use of much domain-dependent information and can be used to acquire more information, thus extending our method described in the preceding chapter. 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. Finally, associating the segmented closed-caption story units with the corresponding video story units attaches the appropriate closed-caption segment, which includes the detailed information about semantic content, to each video story unit. We also discuss experimental results using American football and baseball videos, and the potentiality for utilizing them for a video retrieval system.

Finally, Chapter 5 concludes the main contribution of this work and discusses possible subjects for future research.

Acknowledgements

This work was accomplished under the supervision of Professor Emeritus Tadahiro Kitahashi of Osaka University (now Professor of the Department of Informatics, Kwansei Gakuin University) and Associate Professor Noboru Babaguchi of the Institute of Scientific and Industrial Research, Osaka University (now Professor of the Graduate School of Engineering, Osaka University). I, the author, would like to thank them for helping and guiding me during my Ph.D. study.

I gratefully acknowledge the work of the members of my thesis committee – Professor Hideo Miyahara, who is the chairman of this committee, and Professor Kenichi Hagihara of the Graduate School of Information Science and Technology, Osaka University, and Professor Haruo Takemura of the Cybermedia Center, Osaka University, for their insightful comments on my research and for accommodating a tight defense schedule.

I would especially like to thank my advisor, Professor Noboru Babaguchi, who has been a source of great support and excellent guidance throughout my thesis. Without his continuous encouragement and advice, I could never have completed this thesis.

I also take pleasure in thanking Professors Kenichi Taniguchi, Katsuro Inoue, Toshimitsu Masuzawa, Makoto Imase, Toru Fujiwara, Toshinobu Kashiwabara, Tohru Kikuno, Teruo Higashino, Masaharu Imai, and Professor Hideo Matsuda of the Graduate School

of Information Science and Technology, Prof. Masayuki Murata of the Cybermedia Center, Prof. Shinichi Tamura of the Medical School, Osaka University, Professor Emeritus Nobuki Tokura, and Professor Emeritus Akihiro Hashimoto of Osaka University for providing the resourceful education throughout my Ph.D.

I also have had the benefit of interacting with a number of people during my doctoral program at Osaka University. I especially thank Assistant Professor Kouzou Ohara and Assistant Professor Fumihisa Shibata of Osaka University who provided me with helpful comments and invaluable discussions. I would also like to acknowledge the help and support of everyone in the Intelligent Media Laboratory at the Institute of Scientific and Industrial Research, Osaka University, especially my colleagues of the media group, Mr. Shigekazu Sasamori, Mr. Yukinobu Yasugi, Mr. Hironobu Yamasaki, Mr. Yoshihiko Kawai, Ms. Fumi Nishiue, Mr. Shingo Miyauchi, Mr. Akira Hirano, Mr. Hirotsugu Okamoto, Mr. Takehiro Ogura, Mr. Yoshihiko Fujimoto, and Mr. Tomoyuki Yao, for making my life more pleasant and my work more fun during my doctorate.

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

Last but not least, there is no way I can acknowledge enough the support from my family. I express my sincere thanks to my parents for their understanding and support.

Contents

Preface **i**

Acknowledgements **iii**

1 Introduction **1**

2 A Semantic Description Model for Sports Videos **9**

2.1 Introduction 9

2.2 Structurization of the Sports Video 10

2.3 Semantic Description Model with MPEG-7 15

2.4 Conclusion 17

3 Generating Descriptions for Live Scenes **19**

3.1 Introduction 19

3.2 Attaching Descriptions of Plays and Players 23

3.2.1 Text Stream Analysis 24

3.2.1.1 Extraction of Live Segments 25

3.2.1.2 Generation of Descriptions 27

3.2.2 Image Stream Analysis 29

3.2.3 Text-Image Streams Association 34

3.3 Experimental Results 35
3.3.1 Results of Text Stream Analysis 36
3.3.2 Results of Image Stream Analysis 38
3.3.3 Results of Text-Image Streams Association 39
3.3.4 Experiments with Baseball Videos 42
3.4 Discussion 45
3.5 Conclusion 47

4 Story Segmentation **49**
4.1 Introduction 49
4.2 Video Story Segmentation for Semantic Content Acquisition 50
4.2.1 Segmentation of Text Stream 51
4.2.2 Association of Text and Video Streams 56
4.3 Experimental Results 59
4.3.1 Experiments with American Football Videos 60
4.3.2 Experiments with Baseball Videos 66
4.4 Discussion 70
4.5 Conclusion 73

5 Conclusion **75**

Bibliography **81**

List of Tables

3.1 Examples of key phrases 26
3.2 Procedure to determine players 28
3.3 Relationship between plays and beginning images 32
3.4 How to determine plays integrating text and image streams 35
3.5 Results of a live scene extraction from CC text 36
3.6 Details of generated descriptions 37
3.7 Results of extracting live scenes from the image stream 38
3.8 Results of acquisition of plays from the image stream 39
3.9 Results of generating descriptions 40
3.10 Results of attachment of descriptions to videos 42

4.1 Characteristics of each scene in CC text 52
4.2 American football videos 61
4.3 Results of CC scene categorization (American football) 62
4.4 Results of CC story unit generation (American football) 63
4.5 Results of video story segmentation (American football) 65
4.6 Results of CC and video integration (American football) 66
4.7 Results of CC scene categorization (baseball using learned data from baseball videos) 68

4.8 Results of CC scene categorization (baseball using learned data from American football videos) 68
4.9 Results of CC story unit generation (baseball) 69
4.10 Results of video segmentation (baseball) 69
4.11 Results of CC and video integration (baseball) 70
4.12 Examples of semantic content acquisition 72

List of Figures

1.1 Environment of our system 2

2.1 Structure of the video stream 11
2.2 Structure of a sports TV program 12
2.3 Characteristics of the image stream 13
2.4 Structure of a sports game 14
2.5 Overall structure of the sports video 15
2.6 Description model with MPEG-7 18

3.1 Outline of our system 22
3.2 Example of CC text 23
3.3 Outline of the proposed method 24
3.4 Characteristics of the image stream 30
3.5 Examples of beginning images 31
3.6 Block matching 33
3.7 Text-image streams association 34
3.8 Example of final description results 41
3.9 Examples of beginning images for baseball 43

4.1 Outline of the proposed method 51

4.2 Bayesian Network 54
4.3 Bayesian Network for categorizing CC segments 55
4.4 Association of CC and video: the CC text in the lower part shows the actual boundaries of the corresponding CC text. Finding the boundaries at the position with the same time lag doesn't necessarily provide the correct boundaries (the CC text in the upper part). 57
4.5 Association of CC and video 59
4.6 Examples of beginning images (American football) 64
4.7 Examples of beginning images (baseball) 67
4.8 Example of the final description result 71

Chapter 1

Introduction

A continuous increase in the amount of *unstructured* multimedia data has strongly required a novel framework of simple but meaningful representation that enables efficient multimedia retrieval or a filtering system [SFChang99, Dimitrova99], in which multimedia databases can be searched with queries on the basis of some sort of organization of the time-oriented structure of the data, and more interestingly, on the basis of the semantic content. The prime 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 and is an extremely important search tool that should by no means be neglected in the video retrieval problem. It is also important that the text descriptions of the attributes accurately reflect the characteristics of nontext multimedia data types to a certain extent, and with the best capabilities. These text descriptions should also describe the content sufficiently to help the users to locate segments of interest or, at least, not exclude potential segments from the candidate list.

As a scheme to realize the text descriptions, the MPEG-7 [MDS01, Benitez01, SFChang01], formally known as Multimedia Content Description Interface, has become an international

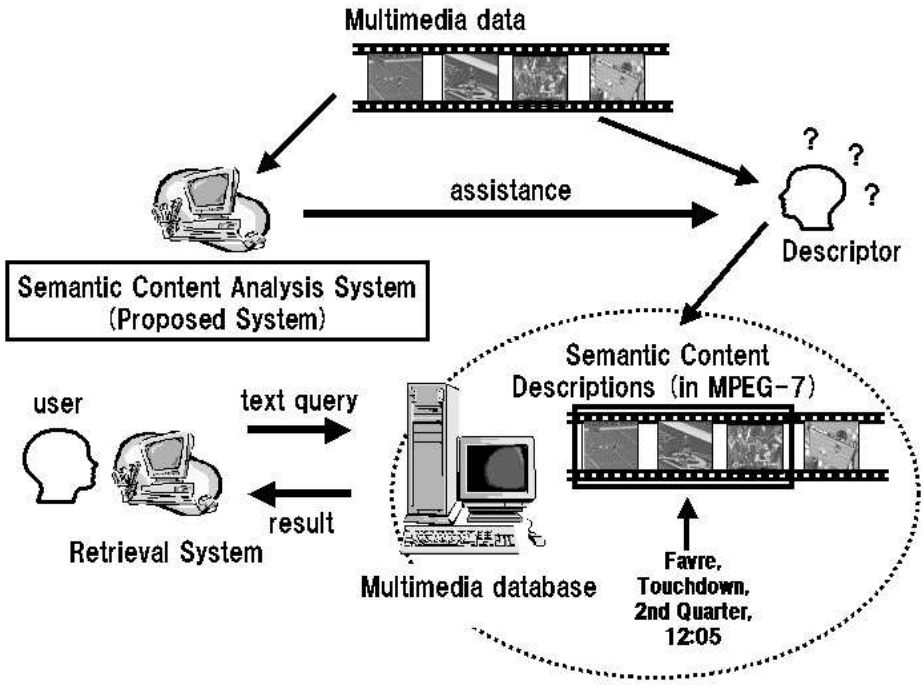


Figure 1.1: Environment of our system

standard for describing multimedia data in October 2001. 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 and how the descriptors/users obtain these descriptions. Therefore, aiming at effective retrieval or filtering systems, we need to determine how to represent multimedia data in a clear and concise way, and how to automatically or semi-automatically acquire the needed descriptions of the semantic content. As shown in Figure1.1, the aim of this thesis is to develop a system to assist the manual description of the multimedia data with (semi-)automatic semantic content analysis.

Note that the *multimedia data* can be defined as the data that is comprised of several information streams: image, audio, and text streams. There are many categories among the multimedia data such as surveillance, satellite, medical, industrial inspection, video-conference, home-made, broadcasted videos, etc [Roach02, Alatan01]. Among these video

categories, broadcasted videos, more specifically TV programs, have been one of the most popular and challenging targets of content analysis due to their unlimited and unpredictable variability in their content. This is also the area with which this thesis is concerned, and we use the term *video* to refer to broadcasted video hereafter.

We now proceed to discuss the characteristics of videos. Videos have genres according to their semantic content, such as news, dramas, documentaries, movies, and sports. Regardless of its genre, every video is a sequence of *shots*, which are contiguously recorded sequences with a single camera, and each shot is a sequence of *frames*, each of which represents an image. However, it is highly unlikely that users would think in terms of single frames or single shots when interacting with a video-retrieval system. Instead, they would view a video as a sequence of *scenes*, each of which is any series of shots unified by their semantic content. Obviously, a video needs to be segmented based on these scenes, however, no concrete definition of semantic content can be available.

Now, let us consider how these scenes construct the videos. Videos have some typical semantic structures which depend on their genres. Due to the inability of a universal semantic content definition to account for all genres, many applications have been developed specifically to each genre according to its structure. For instance, a news video can be considered as a sequence of scenes which start with an image frame presenting the anchor person followed by the variety of the news [Gao00, Blumin, Shearer00, Eickeker99], a drama video or a movie as an assembly of the semantically interrelated scenes [Alatan01, Hanjalic99, Kwon00, Javed01, YLi01], and a sports video as a repetition of the play and break scenes [Zhong01, BLi01, Xu01]. Each element of the sequence is called *LSU(Logical Story Unit)* [Hanjalic99] or simply *story unit* and their assembly constructs every video and gives it the semantic meaning or the story. Therefore, video structurization based on these story units and understanding their semantic content must be done as a step toward semantic

representation. In this thesis, we focus on the sports video, which is one of the most challenging targets of semantic representation because of its wide variety in domains such as football, baseball, soccer, etc. [Kittler01], and propose a semantic description model which can be useful to understand the story of a video based on the typical semantic structures common to diverse sports.

We next discuss how to acquire the semantic content of the videos and generate the proposed descriptions when given the raw unstructured videos. The descriptions have usually been acquired solely by hand, if they have been described at all, with meticulous previewing of the video. Ideally, the descriptions should be automatically acquired as a result of machine interpretation of the semantic content of the video; however, given the state of the art in video processing, such sophisticated data analysis may not be feasible in practice. Rather, the use of automatic video processing, despite its limited semantic interpretation, may offer intelligent assistance in the manual description of videos.

Let us first discuss the recent works that attempt semantic content acquisition. As defined above, videos consist of several multimodal information streams closely related to one another: image, audio, and text streams, and the content of videos is usually analyzed processing these streams. As a study for news and drama videos, specific scenes such as the closed-up scenes of a person and interview scenes have been acquired with image stream analysis and attached their semantic indexes such as the person's name, the location, the subject of the talk, and the scene category acquired from the text/audio stream [Satoh99, Shearer99, Nakamura97, Jasinschi01, Mani97, Huang99, Lienhart97, Smith97]. The researchers working with sports videos tried to extract the play scenes and the movement of the player, the ball, the camera, etc. from the image stream, the cheering, the sound of the hitting, etc. from the audio stream, and the kind of events from the text stream [Lazarescu99, YChang96, Babaguchi01, Miyauchi02, Rui00].

Considering the differences between the news, or drama and sports videos, we believe that while the scene category and the main person of each scene is usually considered to be enough information for indexing of the content for the news and drama videos, more detailed information about the content of each scene such as the player's actions and the events is necessary for sports videos. However, what these methods were able to attain was limited to special events like the score events, and other events, players, etc. have been neglected. Moreover, since the image stream of news and drama videos has some distinctive features for each scene, it is relatively easy to achieve image analysis. On the contrary, the image stream of sports videos is constructed of similar image frames, and moreover, the players are not usually perceived as closed-up in the play scenes since their movements are the center of the viewers' interests in those scenes. Therefore, it is much more difficult to acquire semantic information from the image stream for sports videos.

Here, we proceed to examine how the content analysis for the sports videos can be accomplished. The problem of semantic content analysis can be divided into two steps: 1)temporal video segmentation, and 2)semantic content acquisition. In most of the work discussed above, the image stream is mainly used in temporal video segmentation, 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, although the audio or the text stream can be a more effective source for acquiring detailed semantic content [Toklu00]. Obviously, each of the streams gives us only a limited amount of information. Therefore, semantic content analysis should be accomplished by combining each result acquired from multimodal information streams. We call this strategy intermodal collaboration [Babaguchi01, Miyauchi02, Nitta01]. Nevertheless, most of the researches for sports videos have been putting their focus on the image analysis and have analyzed other streams only for a complementary use. In this thesis, we try to

acquire the descriptions needed for our proposed description model by putting more focus on the *closed-caption text*, which is the speech transcript of the announcers, as an important information source of semantic content.

We first propose a method for segmenting a sports video into play/break scenes, and attach the semantic descriptions to all the play scenes, using the closed-caption text for semantic content acquisition and the image stream for video segmentation [Nitta01, Nitta00a, Nitta00b, Nitta00c, Nitta99]. The proposed method first segments the closed-caption text to extract the parts, each of which corresponds to the play scene unit by utilizing key phrases which are determined beforehand, and then acquires the main elements of the story of each unit such as the plays and the players as the text descriptions. Note that the use of the key phrases restricts the location of where to look for the semantic information, filtering out other insignificant parts from the whole closed-caption text, in order to acquire only the most significant information. Next, focusing on some image cues which are acquired with the knowledge of the sports videos, we also try to extract the corresponding segments from the image stream. Finally, by temporally integrating the results acquired from both streams, we attach the text descriptions to the proper segments of the video.

While the proposed method aims at generating the descriptions in a form such as <Play, Player>, another form of the descriptions can be the sentences including the semantic content such as “First down and 10. Player A runs to 20-yard line and is taken down by Player B”, which can be used for the video retrieval system in the same way as the text retrieval system [Oard97]. For the news videos, many researchers have been trying to attach the semantic documents for new videos generating the documents with the topic segmentation of the automatically transcribed or the closed-caption text [Shahraray95, Takao01, Zhu01, Greiff01, Mulbregt99, Ponte97]. These authors 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 thesis, we also present a method for segmenting the closed-caption text according to the semantic structure of sports videos [Nitta02a, Nitta02b, Nitta02c, Nitta02d]. 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 avoiding the use of many keywords or key phrases which will complicate the versatility of the method, and segment 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 construct 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. The synchronization with the video stream is accomplished by association with the video story units obtained in the same way as proposed in Chapter 3. After the synchronization, extracting only the significant scene units from each story unit of closed-caption text attaches the detailed text description to each video story unit.

The thesis is organized as follows. Chapter 2 discusses the structure of broadcasted sports videos and proposes a semantic description model suitable for sports videos. Chapter 3 proposes a method for attaching the description about plays and players to each story unit. Chapter 4 proposes a method for segmenting the closed-caption text into the story units to attach more detailed description to each story unit. Chapter 5 concludes this thesis and gives the future work.

Chapter 2

A Semantic Description Model for Sports Videos

2.1 Introduction

Many TV viewers often face difficulties when they are too busy to watch a whole program, or when they want to watch only a part of the program, but do not know its location within the program. What viewers usually do is tape the video, and afterward, watch it with fast-forwarding or fast-rewinding and occasionally stopping the video at the location of their interest. In order to save the TV viewers' time, indexing a video according to its content is one of the desired solutions, so that the viewers can easily search the location of what they are looking for.

Video content is usually characterized based on visual content, and can be grouped into two types: low-level visual content and semantic content. Low-level visual content is characterized by primitive visual features such as colors, shapes, textures, etc., and the semantic content contains higher-level concepts such as objects and events. Moreover,

the visual content of videos differs from that of the static images in containing dynamic temporal content in a sequence of images as well as static content in a single image. For the diversity in the video content, the text descriptions, which can be used for the textual query, accurately reflect the characteristics of nontext multimedia data types to a certain extent providing the users simple and understandable information, while the visual descriptions, which can be used for the query by example and visual sketches, allows more broad, rather redundant information. Note that the standardization of the MPEG-7 [MDS01, Benitez01, SFChang01], formally known as Multimedia Content Description Interface, has realized these text descriptions of multimedia data.

Next, we have to consider what kind of semantic information would be enough to satisfy the needs of users and how concisely we need to index the video. In this chapter, we discuss both syntactic and semantic structures of videos, and based on these structures, present the semantic description model especially dealing with the broadcasted sports video, which is not only one of the popular genres among TV viewers but also the genre for which this kind of technology is highly desired [Li00].

2.2 Structurization of the Sports Video

In order to represent videos, the primal concern will be the granularity of video segments to index. Videos have some common structures regardless of their genres. Note that we now focus on the visual features since they are the most important and instinctively understandable factors for the semantic structures of videos.

As you can see in Figure2.1, at the most basic levels, a video is a sequence of *frames*, each of which is the static image. At the next level, a sequence of frames construct a segment called a *shot*. A shot is a sequence that is contiguously recorded with a single

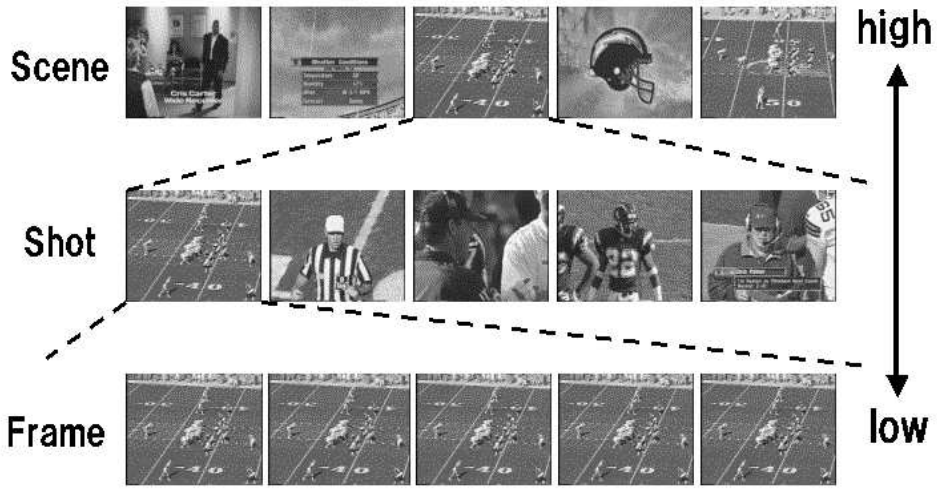


Figure 2.1: Structure of the video stream

camera, and represents a continuous action in time or space which is often considered as the finest level for semantic descriptions. At the higher level, a sequence of shots is called a *scene*, which has more semantic meaning for the story of the video compared to the other levels. Note that the semantic scene level is recursive. That is, the sequence of scenes at level *L* can be a scene at level *L+1*. For example, the scene of flames breaking out and the scene of extinguishing fire make a news scene of fire at a one step higher level. While the frame and the shot levels can be constructed in the same way for every video, the scene level cannot be, since it is related to the story of the video rather than just the physical features.

Some typical structures depend on their genres. For instance, a news video can be considered as a sequence of units which start with an image frame presenting the anchor person followed by a variety of news [Gao00, Shearer00]. Hauptmann et al. [Hauptmann98]

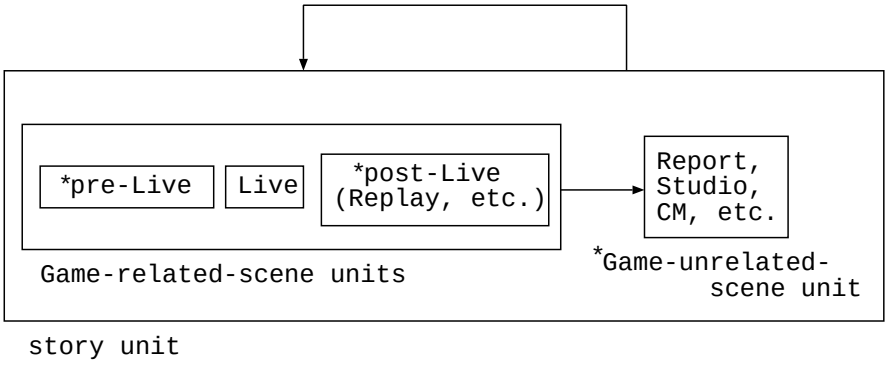


Figure 2.2: Structure of a sports TV program

viewed news videos as a sequence of the news scenes and the commercials, and Eickeker et al. [Eickeker99] and Blumin et al. [Blumin] respectively viewed them as a sequence of scenes of six classes: Anchor, Begin, End, Interview, Weather Forecast, and Report, and four classes: Flying logo, Anchorperson, Newsreel, and Graphic. For other genres, Hanjalic et al. [Hanjalic99], Kwon et al. [Kwon00], Li et al. [YLi01], and Alatan [Alatan01] regarded movies or dramas as an ensemble of semantically related scenes with a similar appearance (scenes at the same location, with the same people, etc.), and Javed et al. [Javed01] considered talk and game shows as a repetition of commercials, hosts, and guests scenes. Such flows of the scenes construct every video and give it the semantic meaning or the story. Therefore, in order to represent video effectively, the structure of the video should be the key to grasping the story of the video. Since each scene of the video is an element of the structure, segmenting the video into these scenes and understanding their content must be done as a step toward semantic representation.

The sports video has two kinds of structures seen from different point of views: sports TV program, and sports game. Here, we discuss both structures and present a semantic description model for sports videos summarizing these different structures.

(i)Structure of a Sports Program

Figure 2.3: Characteristics of the image stream

A TV program of a sports game can be regarded as a sequence of the units shown in Figure2.2. We define a “Live” scene as the time interval which begins with the players starting to move and ends with an event such as getting the score, getting out of the field. The “Pre-Live” and the “Post-Live” scenes can be defined as the time before and after the “Live” scene which have a semantic relationship to the in-between “Live” scene. Other scenes such as “Report” and “CM(Commercial Message)” are considered semantically unrelated to game[1]. We call each of these scenes a *scene unit*, and also regard the time interval between a “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

[1]The “*” attached to each segment in Figure2.2 indicates the possibility of 0 occurrence of the corresponding segment.

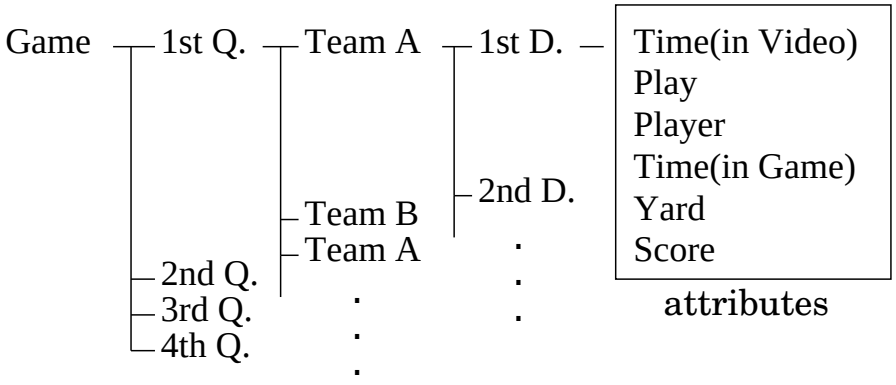


Figure 2.4: Structure of a sports game

kinds of images. Figure2.3 shows an example of these images for American football, baseball, and basketball. Each image shown here is the first frame of each shot, and those framed by the rectangles represent the "Live" scenes.

(ii)Structure of a Sports Game

A sports game in the case of American football can be expressed as a tree, as shown in Figure2.4. A sports game is actually going on by repeating the fundamental element of the tree such as "1st D.(Down)" and "2nd D." in Figure2.4. 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 about the upper node(1st–4th Quarter, offense team name, 1st–4th Down, etc. for American football, 1st-9th inning, top/bottom, etc. 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 requirement. The information discussed above will satisfy the parts of the 5Ws1H requirement. Although Figure2.4 shows the structure of American football

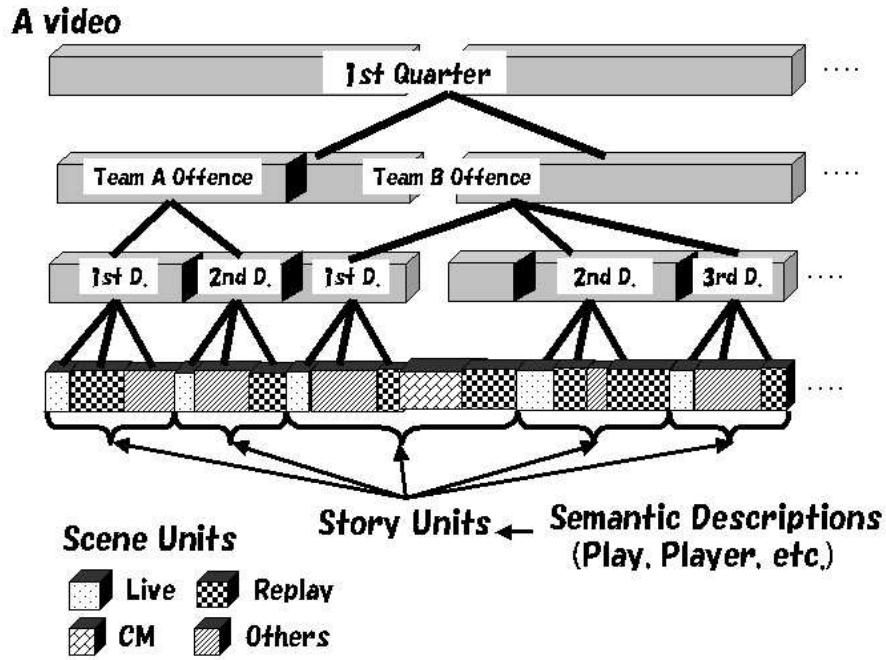


Figure 2.5: Overall structure of the sports video

game, such a tree-like structure is general to many kinds of sports.

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. Figure2.5 shows the overall structure of a sports video we try to construct.

2.3 Semantic Description Model with MPEG-7

In order to describe the content of the video, we can use the *content description tools* which are part of the MPEG-7 MDS tools [Benitez01, SFChang01, MDS01]. The *content description tools* are divided into two parts: the *structure description tools* and the *semantic description tools* which describe the structure and the semantics of multimedia data,

respectively. The *structure description tools* represent the structure of multimedia data in space, time, and the media source by describing general and application-specific segments of multimedia data together with their attributes, hierarchical decompositions, and relations. The *semantic description tools* represent *narrative worlds* depicted in, or related to, multimedia content by describing *semantic entities* such as *objects*, *events*, and so on. A *narrative world* refers to the reality in which the description makes sense, which may be the world depicted in the multimedia data. *Objects* and *events* are perceivable entities that exist or take place in time and space in the *narrative world*, respectively. *Agent Objects* are *objects* that are persons, group of persons, or organizations.

The *semantic relation* description tools describe general relations among semantic entities and other entities. For example, as the relations among semantic entities, we have: "hasAgentOf" – *object* that initiates an *action* or an *event*, "hasDestinationOf"– *object* that is the finishing point for the transfer or the motion of an *event*, "hasPatientOf"– *object* whose state is affected by an *event*, and so on.

Figure2.6 shows an example of the MPEG-7 description representing the tree structure of an American football video. "subseg" segments the whole video into first half and second half, "offenseteam" divides each half for every offense team, "DOWN" divides each team segments for every "down", and "scene" divides each down into the scene units. Each "down" which corresponds to a story unit has the annotations or the semantic descriptions. Here, we consider the plays as the *events* and the players as the *agent objects*. For example, in a scene such as "X passes the ball to Y.", "X" and "Y" are the *agent objects*, "Pass" is the *event*, and as the relation between them, "X" initiates the "PASS", and "Y" is the finishing point of the "PASS". Figure2.6 shows a MPEG-7 description representing a live scene from 0:20:42 to 0:20:47 in the video time. We describe the player as the *agent object* separately from the play - the *event* - and locate the actual scene as the *event* scene with a relationship

between the play and the player.

2.4 Conclusion

This chapter proposed a semantic description model for a sports video to effectively represent its story based on the structures common to many kinds of sports. The proposed model defines the semantic units for sports videos, and gives the semantic content which is usable for semantic video retrieval system. We also showed that the proposed model is easily applicable to the MPEG-7, which has been standardized to describe the content of multimedia data in a textual form. This chapter answers the question: *What should be described?*, and now we proceed to the next question: *How do we obtain the descriptions?*

```
<AudioVisualSegment id=''whole''>
  <TemporalDecomposition id=''subseg''>
  <AudioVisualSegment id=''1sthalf''>
    <TemporalDecomposition id=''offenseteam''>
    <AudioVisualSegment id=''team1-1''>
      <TemporalDecomposition id=''DOWN''>
      <AudioVisualSegment id=''1std-1>
        <TextAnnotation>
          <FreeTextAnnotation>
               1st Half, Offense of Team A,
               1st Down, 15:00 left,
               Kickoff by X, 0-0 ...
          </FreeTextAnnotation>
        </TextAnnotation>
        <!-- Events -->
        <Semantics>
          <SemanticsBase xsi:type=EventType''
                                id=''kickoff1_ins-eve''>
            <Relation xsi:type=''ObjectEventRelationType''
              name=''hasAgentOf'' target=''#x_ins-obj''/>
          </SemanticsBase>
        </Semantics>
        <TemporalDecomposition id=''scene''>
        <AudioVisualSegment id=''live-1''>
          <MediaTime>
            <MediaRelTimePoint>
                 T0:20:42
            </MediaRelTimePoint>
            <MediaDuration>
                 PT5S
            </MediaDuration>
          </MediaTime>
        </AudioVidualSegment>
        <AudioVisualSegment id=''replay-1''>
                .
                .
        </TemporalDecomposition>
      <AudioVisualSegment id=''2ndd-1''>
              .
              .
      </TemporalDecomposition>
    <AudioVisualSegment id=''team2-1''>
            .
            .
    </TemporalDecomposition>
  <AudioVisualSegment id=''2ndhalf''>
          .
          .
  </TemporalDecomposition>
</AudioVisualSegment>

<!--  Objects -->
<Semantics>
<SemanticsBase xsi:type=''AgentObjectType''
                                id=''x_ins-obj ''>
  <Agent xsi:type=''PersonType'' id=''x-per''>
    <Name>
      X
    </Name>
  </Agent>
</SemanticsBase>
</Semantics>
```

Figure 2.6: Description model with MPEG-7

Chapter 3

Generating Descriptions for Live Scenes

3.1 Introduction

We have proposed a semantic description model for sports videos in Chapter 2. Here, we face another problem: How do we obtain the information necessary for the descriptions? Obviously, it is extremely time-consuming to generate entire descriptions for each video manually, because such descriptions usually take as many as a few days for a single program. 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 is composed of two main problems: *Temporal Video Segmentation* and *Semantic Content Acquisition*, that is, acquisition of the semantic information such as objects, motions, events from each video segment.

While videos have several sources of information such as image, text, and audio streams, most of these research papers have focused on the image stream paying attention to low

level visual features such as colors, textures, and shapes, since the methods for *image analysis and retrieval* are easily applicable to the selected representative frames extracted from video segments.

Two challenges for temporal video segmentation exist: *Shot Detection* and *Video Structure Parsing* [Dimitrova99]. Shot detection is literally to detect shot boundaries and such detection has been done with a simple comparison of pixel differences between frames, histograms, edge content, or DCT coefficients. Video structure parsing is to detect the more semantic boundaries based on temporal structures and the relationship of several shots. Therefore, while shot detection involves only detection of temporal boundaries, video structure parsing additionally involves identification of the meaningful composition of temporal elements in the video. For instance, 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 [Gao00, Blumin]. For movies or drama videos, Hanjalic et al. [Hanjalic99], Kwon et al. [Kwon00], and Javed et al. [Javed01] attempted to segment the videos into semantically related scenes based on the visual similarity of each shot. For sports videos, Zhong et al. [Zhong01] and Li et al. [BLi01] tried to extract patterned event boundaries from the image stream, and Xu et al. [Xu01] also tried to segment a soccer video into play/break scenes with frame-based image analysis.

Semantic content acquisition has also been accomplished by detecting *object*, *motion*, and *event* with the visual features. For example, Zhou et al. [Zhou00], Gong et al. [Gong95], and Sudhir et al. [Sudhir98] 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 researches discussed above are 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 news or drama videos, Nakamura et al. [Nakamura97], Shearer et al. [Shearer00], Eickeker et al. [Eickeker99], and Janinschi [Jasinschi01] proposed methods for segmenting news videos and semantically classifying each segment by combining the features of the text/audio and the image streams. Satoh et al. [Satoh99] have developed a system that identifies faces, by associating the faces extracted from the image stream and the names from the text stream. Shearer et al. [Shearer99] detected the interview scenes based on the structure of the camera shots, and acquired the semantic information analyzing the words' occurrence in the narration. Li et al. [YLi01] proposed a method of detecting semantically related scenes based on the similarity in the image and the audio stream, and acquired semantic content from the text stream. Mani et al. [Mani97], Huang et al. [Huang99], and Lienhart et al. [Lienhart97] each tried to make video summary using image, audio, and text streams, and Smith et al. [Smith97] also proposed video skimming by selecting video segments based on TF-IDF, detecting camera motions, faces, etc. For sports videos, Lazarescu et al. [Lazarescu99] 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. [YChang96] tried to detect important events by integrating the audio and the image streams. Babaguchi et al. [Babaguchi01] and Miyauchi et al. [Miyauchi02] proposed event scene detection by integrating text, audio, and image streams.

As just described, although videos can be analyzed with several information streams,

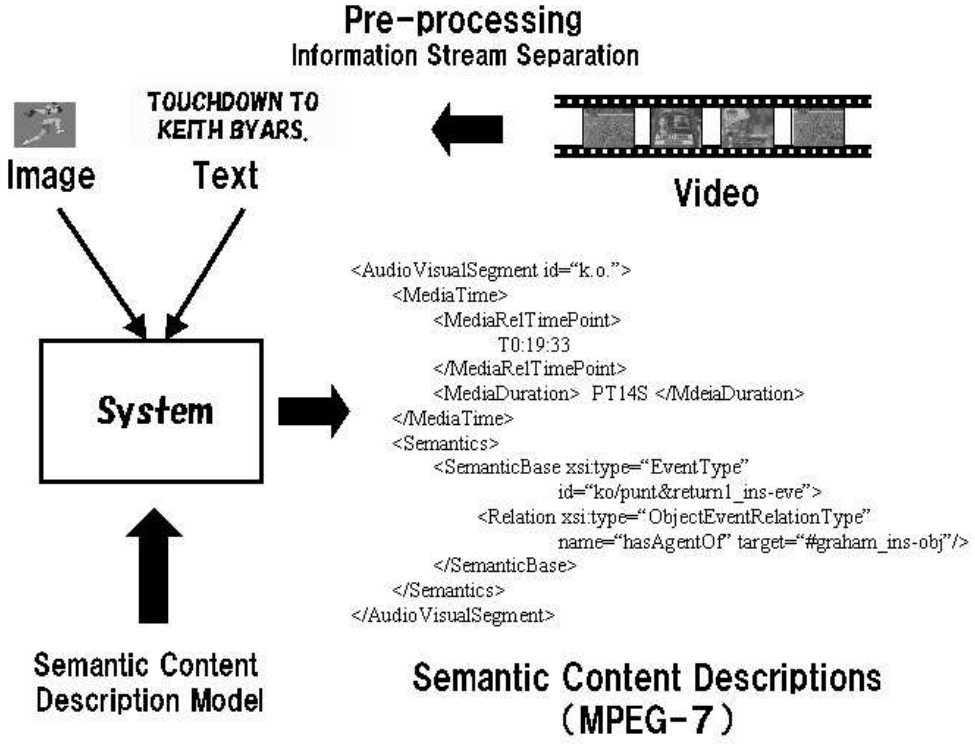


Figure 3.1: Outline of our system

each of these streams has its own advantage and disadvantage. While the image stream is the most reliable information source for video shot segmentation, video structure parsing and semantic content acquisition require 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 be processed [Toklu00]. Taking advantage of each stream and making up for the weak point mutually, we try to integrate these multiple simultaneous streams. In this chapter, we propose a method for acquiring the information about plays and players as semantic content of the story units and attach them to each corresponding live scene by integrating the text and the image streams. This developed system takes a raw video as the

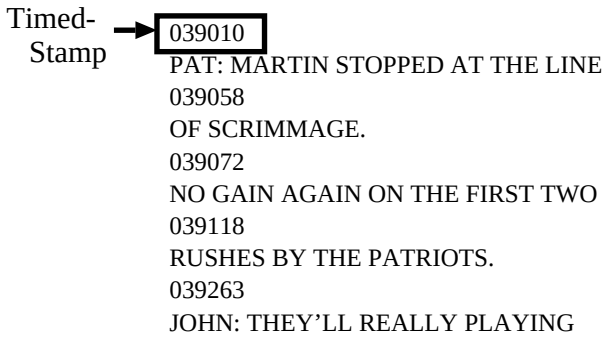


Figure 3.2: Example of CC text

input and outputs the semantic MPEG-7 descriptions as shown in Figure3.1.

3.2 Attaching Descriptions of Plays and Players

The text stream we use here is called Closed-Caption(CC) text, which is a transcript of speech and sound. This stream also contains several markers, such as "$>>$" to indicate the change of the speaker, "NAME:" to indicate the speaker. It is broadcasted together with the video as character code and is easily reproduced in text form. We add the "Timed-Stamp" to the original CC text in advance as shown in Figure3.2. The "Timed-Stamp" increases per 1/30 seconds and represents the image frame in which the first character of each line appears.

Figure3.3 shows the outline of the proposed method. We first generate the semantic descriptions by acquiring the information about the plays and the players from the text stream, then segment the video stream to extract the live scenes using the image stream, and finally, attach the generated descriptions to the video stream by integrating these results. Each of these steps is explained below.

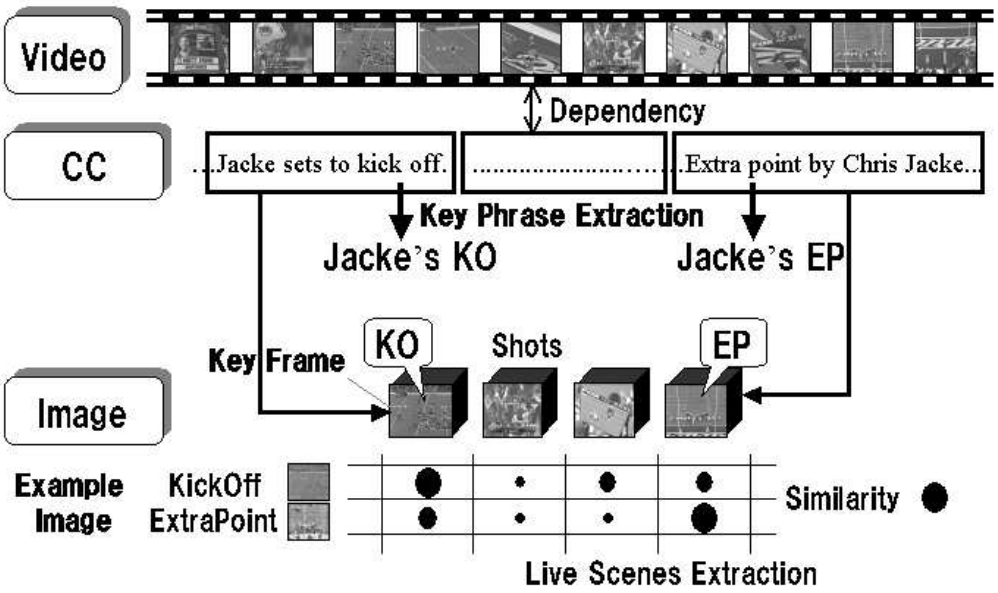


Figure 3.3: Outline of the proposed method

3.2.1 Text Stream Analysis

Since the announcers talk about the situation of the game in a sports game program, we have a good chance to acquire information about the game from the CC text. Specifically, the announcers usually explain the current situation simultaneously with the video stream in the live scenes. Therefore, we try to extract the segments corresponding to these live scenes from the whole CC text, and then acquire the information about plays and players from each extracted CC live segment.

Here, we define two kinds of plays: “action” and “event”.

action: The fundamental movement in the corresponding sports, i.e., Run, Throw, Kick, etc.

event: The result of the actions which has more meaning to the story of the game than the action, i.e., Touchdown, Home Run, etc.

At least one action has to occur and an event may or may not occur in a live scene. Note that although there can be no event in a live scene, if there is one, the event has more importance than the actions in the live scene.

3.2.1.1 Extraction of Live Segments

Since the announcers usually talk in colloquial language in sports videos, natural language processing such as syntactic parsing is hard to apply to their speech. However, in the live scenes, the announcers use some characteristic phrases to specify the situation of the game. These phrases can be determined by the kind of play. We define these phrases as the key phrases, and try to extract those segments which match the live scenes in the CC text by searching these key phrases. An example of these key phrases for American football are shown in Table3.1. We restrict the play of American football to twelve plays: Run, Pass, Kick, Taken Down as the actions, and TouchDown(TD), ExtraPoint(EP), FieldGoal(FG), Return, Punt, KickOff, Interception, Fumble as the events, and each key phrase corresponds to only one play. Here, we determine the key phrases like “runner’s name” and “punter’s name” with knowledge about the players’ roles. In addition to the key phrases indicating the play itself, we determine the key phrases which indicate the beginning or the end of the play, such as, “FIRST DOWN AND yards”. Table 3.1 also shows the role of players such as: the agent, the patient and the destination that should be involved in each play based on the MPEG-7 definition.

Moreover, since an announcer always does a play-by-play commentary of the game in sports videos, we can eliminate the segments in which every speaker except that announcer is speaking as the segments other than the live segments. Following these characteristics,

Table 3.1: Examples of key phrases

plays	key phrases	player role
Run	runner's name, GETS 5 YARDS, etc.	agent
Pass	MAKES CATCH, THROWS TO PLAYER, etc.	agent and destination
Taken Down	TAKEN DOWN, STOPPED BY, etc.	patient and agent
Kick	punter's name, etc.	agent
TouchDown	TOUCHDOWN	agent
ExtraPoint	EXTRA POINT, etc.	agent
FieldGoal	FIELD GOAL	agent
Return	RETURN	agent
Punt	PUNT	agent
KickOff	KICKOFF, etc.	agent
Intercept	INTERCEPTED, etc.	agent
Fumble	FUMBLE, etc.	agent

we extract the live segments as follows:

[Procedure to extract live segments from CC text]

1. Segment the CC text into its segments spoken by a single speaker, and identify the speaker of each segment.
2. Search the key phrases in each segment which starts with the "NAME:" corresponding to the one who makes the play-by-play commentary. If any key phrases are found, identify the segment as a live segment.

Here, the following knowledge should be given beforehand.

Knowledge of a Specific Video: the name of the announcer who makes a play-by-play commentary, and the name of the players.

Knowledge of Specific Sports: the kind of plays and key phrases by which the announcers explain each play.

3.2.1.2 Generation of Descriptions

Now that we have extracted the live segments from the CC text, we identify the main players and the plays. As indicated in section 3.2.1, the outline of the play is given in the live segments. Considering this, we can obtain the information about the play and the player from the extracted segments as follows:

[Procedure to determine plays and players]

play: Identify the play associated with the key phrases in Table3.1.

player: Identify the players as shown in Table3.2. In this table, "BY" represents the name which appears after the word "by", "TO" represents the name which appears after the word "to" or "for", "O" represents the name which appears after the key phrases, "S" represents the name which appears in other places, and "–" means that it is impossible to determine the player.

When more than one sentence exists which includes the key phrases, we choose the sentence that appears earlier in the segment, since the announcers tend to talk about the play earlier than other topics in a segment. However, as discussed in the definition of "the play" in Section 3.2.1, an event should occur as a result of some action and should consequently be mentioned after the causing actions. Since an event is more important than an action for the story, we give priority to an event over an action, even if the key phrase for the event exists after that of the action.

Now we summarize the method as follows:

[Procedure to generate descriptions]

For each extracted live segment,

1. Check a sentence to see if it includes the key phrases determined as in Table3.1.

Table 3.2: Procedure to determine players

Play	Player role	how?	
Run	agent		S
Pass	agent	active verbs meaning ”throw”→	S
		(passive)→	BY
		active verbs meaning ”catch”→	–
		(passive)→	–
		(nouns)→	S
	destination	active verbs meaning ”throw”→	TO
		(passive)→	TO
		active verbs meaning ”catch”→	S
		(passive)→	BY
		(nouns) →	TO
Taken Down	patient	active verbs meaning ”stop”→	O
		(passive)→	S
	agent	active verbs meaning ”stop”→	S
		(passive)→	BY
Kick	agent		S
TouchDown	agent		BY or S
ExtraPoint	agent		BY or S
FieldGoal	agent		BY or S
Return	agent		BY or S
Punt	agent		BY or S
KickOff	agent		BY or S
Intercept	agent	active verbs→	S
		(passive or noun) →	BY
Fumble	agent		BY or S

2. If so, identify the plays and the players as shown in [Procedure to determine plays and players]. Otherwise, return to step 1. and check the next sentence.

3. If both the play and the players have been identified, use both to determine the descriptions for the live segment. If either the play or the players have not been identified, return to step 1. and check the remaining sentences to see if the missing information can be obtained.

4. If the event can be identified after the action is identified, re-determine the play as the identified event.

5. If all of the sentences of the live segment have been checked, determine the information obtained so far as descriptions for the live segment.

3.2.2 Image Stream Analysis

The "Timed-Stamp" in the CC text represents the rough time when each sentence was spoken. Therefore, these stamps enable us to estimate the time when the sentences were spoken on the video. However, since a human CC translator types what is said just after he/she hears the utterance in sports videos, the CC text lags behind the audio [1]. Moreover, since the text stream is merely segmented according to speaker changes, the boundaries are not necessarily the semantic borders. Therefore, simply calculating from the "Timed-Stamp" can not attach the descriptions to the proper segment of the video.

To solve this problem, we now focus on the image stream. 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. If we consider the segment from these stationary images to any kinds

[1] This is the characteristic of sports videos. The time difference between the CC text and audio depends on the kind of videos.

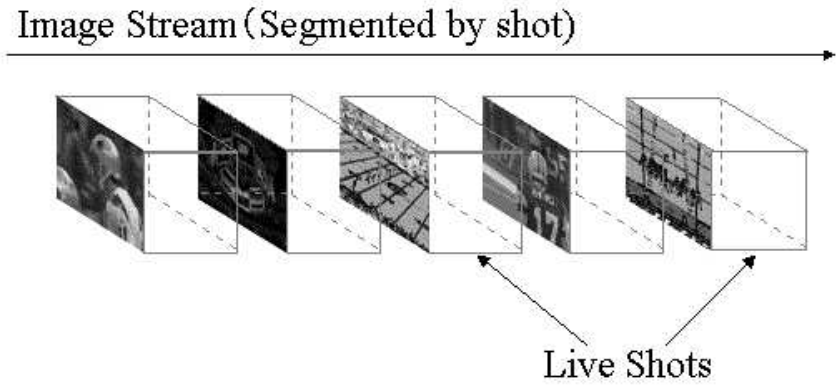


Figure 3.4: Characteristics of the image stream

of shot changes as a live scene, then these segments are randomly located in the video as shown in Figure3.4. These characteristics in the image stream help us distinguish the live scenes in the video.

For American football, we found three kinds of stationary images based on the analysis of the actual broadcasted videos. All the plays acquired from the CC text start with any of these three kinds of images. “ExtraPoint” and “FieldGoal” start with the players lined up before the goal line. Furthermore, the scene is always taken horizontally to the lines (Figure3.5-(b)), and “KickOff” and “Punt” start with the players lined up at the end of the field (Figure3.5-(c),(d)) followed by the shot showing the ball flying (Figure3.5-(e)). Other plays always start with a formation called “Scrimmage” in which players of each team line up face-to-face at a standstill for a while (Figure3.5-(a)). Table3.3 shows the kind of images each play starts with. Since a single camera is used to take one play, every play ends with a shot change. Therefore, a shot change after the frames as shown in Figure3.5 should indicate the end of a live scene. Taking advantage of these features, we segment the video and extract the live scenes as shown below.

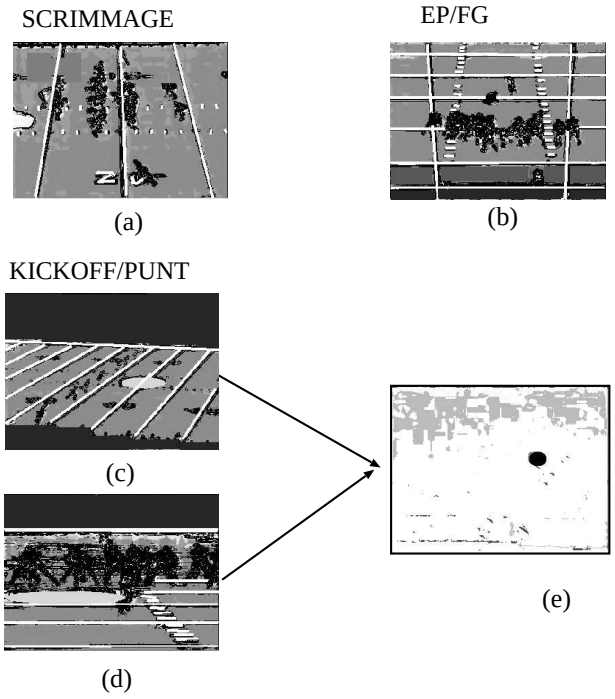


Figure 3.5: Examples of beginning images

[Procedure to extract live scenes from the image stream]

1. Detect shot changes to segment the image stream.
2. Compare the first f_1 frames (add the middle f_2 frames for the ball flying image) of each shot with the example beginning images, such as in Figure3.5, and if there are more than th similar frames, determine the shot as a live scene.

Note that we check both the first and the middle frames of a shot with the ball flying image for "KickOff/Punt". The "KickOff/Punt" scene sometimes consists of two shots: one shot starting with the beginning image of "KickOff/Punt", and the other shot starting with the ball flying image. In this case, we determine these two shots together as a live scene.

Moreover, we calculate the similarity between the two image frames as shown below:

Table 3.3: Relationship between plays and beginning images

Play	Beginning Image
Run	Scrimmage, KickOff/Punt
Pass	Scrimmage
Taken Down	Scrimmage, KickOff/Punt
Kick	KickOff/Punt, EP/FG
TouchDown	Scrimmage
ExtraPoint	EP/FG
FieldGoal	EP/FG
Return	Scrimmage, KickOff/Punt
Punt	KickOff/Punt
KickOff	KickOff/Punt
Intercept	Scrimmage
Fumble	Scrimmage

[Calculation of Similarity]

For the example image I_a and the image I_b to be compared,

1. Divide both images into $M \times N$ rectangular blocks. Taking color distribution as a feature parameter in the vector form such as

$$(R_{m,n}, G_{m,n}, B_{m,n}),$$

where $R_{m,n}$,$G_{m,n}$,$B_{m,n}$ are RGB color histograms in the mth $\times$ nth block, we measure the distance between the vector of x_i in I_a and the vectors of all the same-sized blocks in the adjacent block to the corresponding block in I_b (See Figure3.6).

2. If any of the distances measured in Step 1. are smaller than some threshold, we determine that there is a block in I_b which is similar to x_i. We repeat Step 1. for all the $M \times N$ blocks in I_a, count the number of x_i which has a similar block in I_b, and let $Total(a, b)$ be the total number.

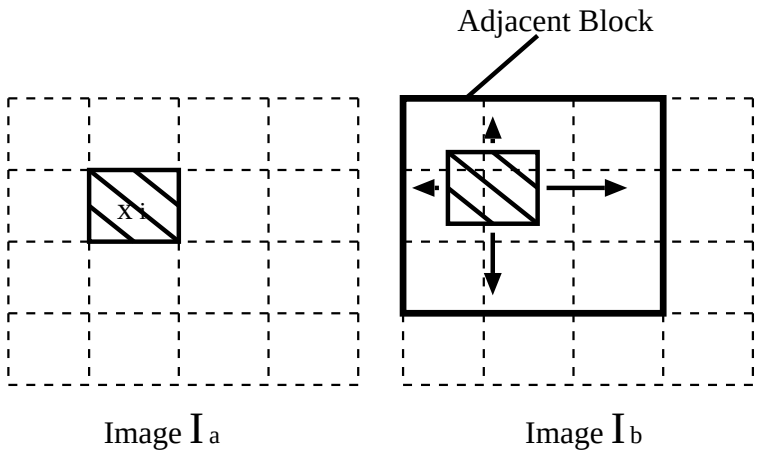


Figure 3.6: Block matching

3. Obtain the similarity $s_{a,b}$ between I_a and I_b defined as

$$s_{a,b} = \frac{Total(a,b)}{M \times N},$$

normalizing the total number of similar blocks between I_a and I_b.

Further, we classify the play category of each live scene into one of the three kinds of play. If there are more than two play categories to which the frame is similar, however, we select one of these two categories in the following condition.

- Since we check the middle frames as well as the beginning frames for "KickOff/Punt", we place more confidence on the extraction of this play category and give priority to it.

- For "Scrimmage" and "EP/FG", calculate the average similarity S(c) as

$$S(c) = \frac{\sum_{i=1}^{f_1} s_{c,i}}{f_1},$$

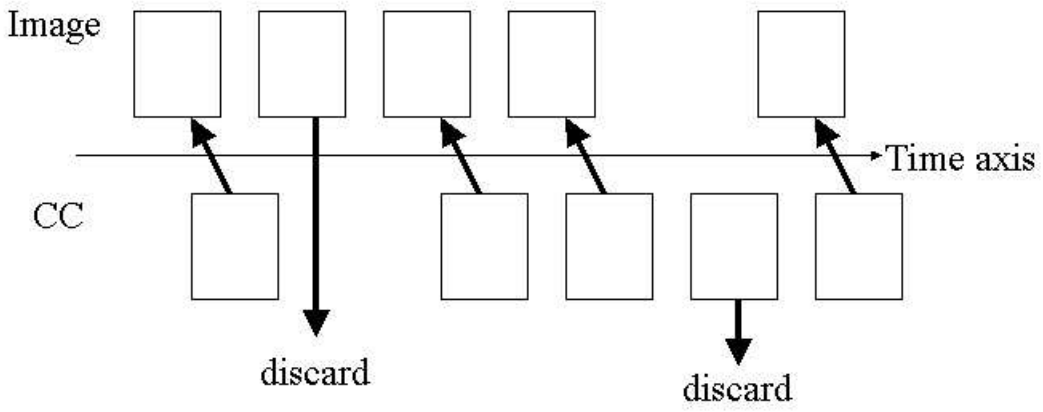


Figure 3.7: Text-image streams association

where c represents the play category and i represents the frame number from the beginning of a shot. Next, determine the play category as c such that $S(c)$ is larger.

3.2.3 Text-Image Streams Association

Finally, we associate the descriptions generated from the CC text with the live scenes extracted from the image stream. We first calculate the start time of the CC live segment with the "Timed-Stamp". As shown in Figure3.7, if the video live segment starts just before the calculated time, we determine the segment as the corresponding scene. Otherwise, we discard the CC segment. Furthermore, we also discard those video live segments which can not be associated with the CC segments.

Play information is determined as shown in Table3.4. The columns in this table respectively represent the play acquired from the CC text, the play category acquired from the image stream, and the final play determined after the association. In the third column, "CC" means that the play from the CC text takes priority over the play category from the image stream. Basically, we consider the CC text as the more reliable information stream

Table 3.4: How to determine plays integrating text and image streams

CC	Image	After Association
Taken Down, Run	KickOff /Punt	KickOff/Punt &Return
Other Plays	All	CC
None	All	Image

for acquiring the desired content, and give priority to the CC text. However, we obtain the "KickOff/Punt&Return" play by combining the "Run" play acquired from the CC text and the "KickOff/Punt" play category acquired from the image stream. This combination is needed because "Run" usually follows "KickOff/Punt" as "Return" and the "Run" tends to be mentioned more often than the "KickOff/Punt" in the CC text.

In summary, we have 14 kinds of plays as the play descriptions: 10 plays from the CC text, 3 plays from the image stream, and 1 play from the association. "Scrimmage" represents all the plays which start with the "Scrimmage" image, in other words, plays other than "KickOff", "Punt", "ExtraPoint", and "FieldGoal".

3.3 Experimental Results

We implemented our method and tested for parts of 2 broadcasted American football games (about 1 hour each). We present the results of the CC text analysis, of the image stream analysis, and of the association of the text and image streams below, respectively.

Here, as the evaluation factor, we use the precision and the recall rate defined as:

$$Recall = \frac{\sharp\, of\ correct\ extractions}{\sharp\, of\ segments\ to\ be\ extracted}$$

$$Precision = \frac{\sharp\, of\ correct\ extractions}{\sharp\, of\ all\ extractions}$$

Table 3.5: Results of a live scene extraction from CC text

	Video1	Video2	Total
Recall	95.5%(44/46)	88%(37/42)	92%(81/88)
Precision	91.5%(44/48)	95%(37/39)	93%(81/87)

3.3.1 Results of Text Stream Analysis

We chose about 100 key phrases by considering sample broadcasted videos (including the test videos) beforehand, and by learning the frequently used phrases for each kind of play. We first show the results of the extraction of the live segments from the CC text in Table3.5. After the inspection-based check of the actual live segments in the CC text, we performed the experiments and examined whether or not the checked segments were successfully extracted.

Table3.6 shows the results of the details of the generated scene descriptions. In the "Correct" column, (X/Y/Z) represents the X descriptions which missed the play, the Y descriptions which missed the player, and the Z descriptions which missed neither. In addition, the number adjacent to the (X/Y/Z) is the total of X, Y, and Z. The "Incorrect" column shows the number of descriptions which missed both the play and the player. Note that when multiple plays are taking place in a live scene, we say we acquired the information successfully if we were able to obtain one of these plays. The "Recall" and the "Precision" columns show the recall and the precision rate of generation of the scene descriptions, and the "Play Recall" and the "Play Precision" columns show those of the generation of the play descriptions. Finally, the "Player Recall" and the "Player Precision" columns show those of the generation of the player descriptions.

The results of this experiment indicate that

- Since the significant plays such as the score events tend to be mentioned repeatedly, those plays can be mentioned before the actual play in a live scene. As a result, the

Table 3.6: Details of generated descriptions

	Video1	Video2	Total
Correct	43(3/5/35)	35(2/5/28)	78(5/10/63)
Incorrect	1	2	3
Recall	93%(43/46)	83%(35/42)	89%(78/88)
Precision	89%(43/48)	90%(35/39)	90%(78/87)
Play Recall	87%(40/46)	79%(33/42)	83%(73/88)
Play Precision	83%(40/48)	85%(33/39)	84%(73/87)
Player Recall	83%(38/46)	71%(30/42)	79%(68/88)
Player Precision	79%(38/48)	77%(30/39)	77%(68/87)

play information can be erroneously acquired as significant plays, since we give the events priority over the action.

- Since the player is considered to be less important than the play for the story, the announcer puts emphasis in explaining the play, and as a consequence, we find it relatively difficult to acquire information about the player.

- When two people are involved in a play such as “Pass” and “Taken Down”, only one person tends to be mentioned and, therefore, only the mentioned person can successfully be acquired. Moreover, although we used simple rules for identifying players’ names and their roles, we made no mistakes in determining their roles in the experiments.

- There are some miswritten or missing words in the CC text, and sometimes the human CC translators also change the words to summarize the long sentences. The rate of the miswritten or missing words varies among the videos. While the mistakes in the announcers’ names, the players’ names, and the words corresponding to the key phrases should affect the performance, only one error, which was the failure of the player acquisition, occurred as a result of the mis-spelled words in the CC text.

Table 3.7: Results of extracting live scenes from the image stream

	Video1	Video2	Total
Recall	98%(45/46)	90%(38/42)	94%(83/88)
Precision	92%(45/49)	61%(38/62)	75%(83/111)

In addition, since the key phrases play the most important role in this method, we have to further examine how to determine them systematically.

3.3.2 Results of Image Stream Analysis

We next show the results of video live scene extraction with the image stream analysis in Table3.7. Table3.8 presents the details of the acquisition of a play category.

Abrupt shot changes, such as the cuts, were detected by our original algorithm based on a histogram analysis in the sub-regions of an image frame. The gradual shot changes such as dissolves, wipes, and digital video effects(DVEs) were detected manually, generating the video data with shot changes perfectly detected. The sampling rate for the video was six frames per second. We compared the images by setting the parameters as $f_1 = 6$, and $f_2 = 14$ (each first frame of one second from 2 to 15 seconds after the beginning of the shot), and $th = 5$. We gave 1 Scrimmage, 1 ball flying, 2 KickOff/Punt and 2 EP/FG example images for each video stream. For EP/FG, we used images of both sides of the field. We selected the example images taken up front of the field as the general camera direction. The experiments were conducted on the "O2" machine of the SGI with a CPU of "MIPS R5000" and a 180MHz clock. The machine was able to process the image with 0.17 seconds/frame, that is, approximately 6 seconds/shot.

Since we define a live scene as a shot which starts with specific images, the shots corresponding to the live scene are the ground truth for the experiments. Therefore, the evaluation was based on a shot.

Table 3.8: Results of acquisition of plays from the image stream

Play Category	Recall	Precision
Scrimmage	75%(52/69)	80%(52/65)
KickOff/Punt	92%(12/13)	37.5%(12/32)
EP/FG	50%(3/6)	21%(3/14)
Total	76%(67/88)	60%(67/111)

These results imply that:

- The live scenes sometimes include other kinds of shots in the midst of the beginning images. In the case of additional shots in the middle of the beginning images, we extracted both shots on either side of the insignificant shot, and considered the first shot a false detection. As a consequence of the integration with the text stream, however, we were able to remove the unnecessary extraction.

- Although the beginning images succeeded in extracting the live scenes correctly, we were unable to distinguish the play category as accurately as the live scene extraction, because the color distribution of the beginning images resembles each other. To improve identification, we need to find a method other than the comparison of color distribution, such as line detection.

3.3.3 Results of Text-Image Streams Association

We attached the descriptions to the video by associating the text and the image streams. Examples of the attached descriptions are shown in Table3.9. In this table, each column represents a scene category, a particular time in the video, parts of the CC text used for generating the descriptions (the underlined parts), the descriptions generated from the CC text, the play category acquired from the image stream and the attached descriptions after associating the text and the image streams. We show the descriptions in the form of $<$Play,

Table 3.9: Results of generating descriptions

Scene Category	Time	CC	Result from CC	Image	Final Result
Player Closed-up	0:19:20– 0:19:33	PAT: THERE WOULD BE SOMETHING WRONG...			
Live	0:19:33– 0:19:47	PAT: GRAHAM IS TAKEN DOWN AROUND THE 20.	<Taken Down, Graham>	KickOff /Punt	<K.O./Punt& Return,Graham>
Player Introduction	0:19:47– 0:20:20	AND LET'S LOOK AT THAT NEW ENGLAND PATRIOT...			
Live	0:20:20– 0:20:29	PAT: MARTIN. THE BALL CARRY YEAR STOPPED BY BRIAN WILLIAMS.	<Run, Martin>	Scrimmage	<Run, Martin>
Player Introduction	0:20:29– 0:20:50	LET'S LOOK AT THAT STRONG PACKER DEFFENSE.			
Live	0:20:50– 0:21:00	PAT: JEFFERSON, FIRST DOWN FOR THE NEW ENGLAND PATRIOTS.	<,Jefferson>	Scrimmage	<Scrimmage, Jefferson>
Player Closed-up	0:21:00– 0:21:05	JOHN: THAT'S WHAT THEY WERE TALKING ABOUT IS ...			

Player> in this table. The times in the video were calculated from the first and the last image frame number of each extracted live scene. The actual description for the first live scene in the table is shown in Figure3.8 as an example.

Notice that the play information of the descriptions were sometimes compensated by the play category from the image stream. In the first live scene in the table, the play information was obtained as a combined result of the text and the image streams. In the second scene, the play information was obtained only from the CC text, and in the third scene, the information was obtained only from the image stream.

```
<AudioVisualSegment id=''ko''>
  <MediaTime>
    <MediaRelTimePoint>
         T0:19:33
    </MediaRelTimePoint>
    <MediaDuration> PT14S </MediaDuration>
  </MediaTime>
  <Semantics>
    <SemanticsBase xsi:type=''EventType''
              id=''ko/punt&return1_ins-eve ''>
    <Relation xsi:type=''ObjectEventRelationType''
       name=''hasAgentOf'' target=''#graham_ins-obj''/>
    </SemanticsBase>
  </Semantics>

<Semantics>
<SemanticsBase xsi:type=''AgentObjectType''
                        id=''graham_ins-obj ''>
  <Agent xsi:type=''PersonType'' id=''graham-per''>
    <Name>
      <GivenName>Graham</GivenName>
      <FamilyName>Hason</FamilyName>
    </Name>
  </Agent>
</SemanticsBase>
</Semantics>
```

Figure 3.8: Example of final description results

We present the final results of attaching the descriptions to the video in Table3.10. Due to the manner of the association, the extracted segments unable to find their associated segments canceled out each other, and as a consequence, excessive extraction was rated lower, while insufficient extraction was rated higher. Since we extracted most of the live segments from both the text and the image streams, this phenomenon was not so prominent for Video 1. However, because more segments could not be extracted from either the text or the image stream, the recall rate deteriorated more remarkably for Video 2. Moreover, since more live video segments were extracted by mistake from Video2, 2 out of 9 missing video segments were actually attached to the video segments which temporally shifted from the

Table 3.10: Results of attachment of descriptions to videos

	Video1	Video2	Total
Recall	93%(43/46)	79%(33/42)	86%(76/88)
Precision	98%(43/44)	92%(33/36)	95%(76/80)

correct segments.

Only 3 kinds of play categories can be acquired from the image stream, while 12 kinds of plays can be acquired from the CC text. However, the image stream helped us obtain "KickOff/Punt&Return" play which was hard to acquire with only the CC text analysis. The recall rate for this play rose from 29%(4/14), which was the result of the CC text analysis, to 71%(10/14) after association with the image stream.

Moreover, while the analysis of the CC text and the association of the CC text and the image stream, respectively, took only a few seconds for a single 2-hours sports program, the image processing took approximately half the length of the program (6 seconds/shot), making up the greatest portion of the entire time of our proposed method. Obviously, more complex image processing, such as object and motion detection, will require more time. The experiments showed that the quicker and more effective use of the CC text effectively provided us with the semantic content, thus reducing the burden of image processing on the entire system.

3.3.4 Experiments with Baseball Videos

In order to test the generality of our method, we discuss the results of a preliminary experiment with a baseball video (about 1 hour). Making the system work with baseball required changes such as:

the kinds of play

We defined the plays of baseball as four actions: "At Bat", "Strike", "Ball", and

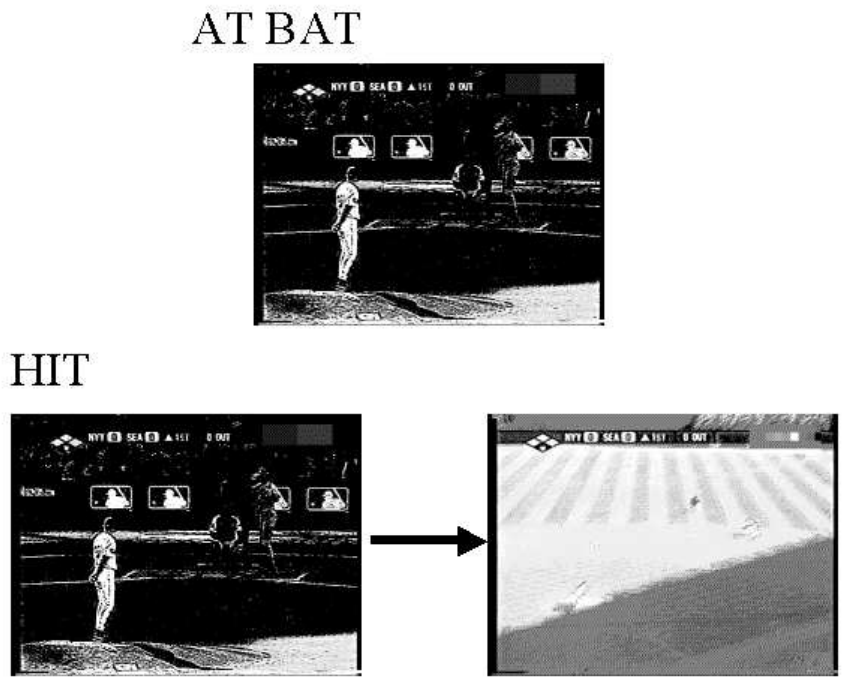


Figure 3.9: Examples of beginning images for baseball

"Foul", and four events: "Ball Four", "Hit", "Out", and "Home Run". For simplicity, we restricted the player to the batter.

the key phrases for each play

We determined about 60 key phrases by consulting some sample broadcasted videos beforehand.

the example images

Figure3.9 shows example images for baseball. The play categories for the images of baseball are "AT BAT" and "HIT". Each of "At Bat", "Strike", "Ball", and "Ball Four" starts with the "AT BAT" image, and "Foul", "Hit", "Out", and "Home Run"

consist of two shots each one of which starts with the images of the "HIT" in Figure3.9. The "HIT" images can be used in the same way as the "KickOff/Punt", which consists of the two consecutive shots of American football.

The recall rate for the extraction of the live segments from the CC text was 71%(67/94), and the precision rate was 89%(67/75). Among the correctly extracted live segments, we were able to acquire both the play and the player information from 49 segments, only the play information from 11 segments, only the player information from 6 segments, and neither the play nor the player information from 1 segment. The results indicate that

- In the baseball video, the announcers sometimes do not talk about the current situation even if it is a live scene, and especially if the play is an action which seldom affects the story of the game such as "At Bat", "Strike", "Ball", and "Foul". In the 19 out of 27 live scenes we failed to extract, the topics were not the current situation, and we therefore consider it impossible to extract those segments from the CC text.

- We failed to extract 5 live scenes because of the errors in the CC text, which is the lack of the information about changes of the speakers.

We segmented the video with image analysis with a recall rate of 95%(89/94) and a precision rate of 77%(89/115). For the acquisition of the play categories, we made only three errors: 1 "AT BAT" was acquired as a "HIT", and 2 "HITs" were acquired as "AT BATs".

Similarly to American football, the live scenes sometimes include other kinds of shots in the midst of the beginning images. In this case, we extracted both shots on either side of the insignificant shots, with the first shot resulting in the false detection. 6 out of 26 excessive extractions are due to this phenomenon. Moreover, the beginning images of the baseball video consist of a green background with the players in the middle. Consequently,

our method erroneously extracted some close-up or full-figure shots with the green background.

Finally, as a result of the association of the CC text and the image stream, we were able to attach the descriptions to the exact live video scene with a recall rate of 63%(59/94) and a precision rate of 87%(59/68). We failed to remove 1 of the 6 segments which were excessively extracted through the image analysis because of the inserted shots in the midst of the beginning images of a live scene. Additionally, 5 of 35 missing extractions were shifted by 1 shot from the actual boundaries. If we allow this slightly shifted association, the recall rate was 69%(65/94) and the precision rate, 96%(65/68).

3.4 Discussion

In this section, we discuss the generality of our method. Our proposed method is based on the general structure of a sports game and a sports TV program, and tries to extract and tag the live scenes which semantically correspond to the story units. As the characteristics of the live scenes, the CC text has distinctive key phrases for each kind of play and the image stream has the characteristic stationary images at the beginning of each play.

Although we used American football videos as the main examples of sports videos, the characteristics of the football game hold true for other kinds of sports, such as baseball and tennis, games considered a repetition of a unit and structured in a tree form with these units, provided that the key phrases and the beginning images are determined for each kind of sports. We provided the results of the experiment using a baseball video in Section 3.3.4 as a basis for this assumption.

For the videos taken of the same sports, the style of the program sometimes differs among production companies. However, the composition of the beginning images, such as

the field distribution and the spatial geometric structures induced by the players, is quite similar for the same sports, and even if the colors vary among the videos because of the players' uniforms or the weather, the style will be the same throughout a video. Although some programs use the characteristic beginning images in other scenes such as the replays, the live scenes rarely start with other kinds of images. Therefore, as long as the live scenes start with a characteristic image, our method can cope with the several kinds of production styles by obtaining the example images from the corresponding video.

Furthermore, the descriptions discussed in this chapter are also based on the content of a general sports game, and therefore should be applicable to other kinds of sports. In other words, the plays and the players are important elements for all types of sports, considering the story of the sports video, and the kinds of plays can be determined for each type of sport. On the basis of these two facts, we can say that our method can be commonly used for many other kinds of sports video, too.

Here, let us compare our work with related work. As we discussed in section 3.1, other studies exist which aim at content analysis for sports videos. Lazarescu et al. [Lazarescu99] proposed a model to describe the play and succeeded in acquiring detailed information about the play, such as the actions of each player and the formations of a team, but their method is more costly, since they focus on measurement of the players' positions on the image frame, while we only need to use template matching to identify specific scenes focusing more on the text stream to acquire semantic information. Moreover, the other authors do not consider where to index in the video.

Chang et al. [YChang96] used the audio stream instead of a text stream and tried to detect important events such as score events, by keyword spotting and cheering detection on the audio stream, as well as line detection on the image stream. Babaguchi et al. [Babaguchi01] tried to detect score events with the image and the text streams, also by

keyword spotting on the text stream and template matching on the image stream. They both limited the search space of the image stream by pre-processing other information streams. Unlike the just-mentioned methods, our method uses the text stream and the image stream for different purposes supplying the deficiency of each other. Moreover, both methods put focus on only specific events, and did not take other plays or the players into consideration.

Xu et al. [Xu01] and Zhong et al. [Zhong01] also considered the semantic structure of a sports program and the characteristic images of the beginning of the live scene, by calling these images the “canonical views”. Although these authors tried to detect the start time of the live scene as the event boundary from color features and object detection on the image stream, they mentioned neither the sports game structure nor the semantic descriptions of each segment.

Considering the structure of a sports game and a sports TV program, we focused more on where to attach the descriptions and the contents of the descriptions themselves. In addition, integrating the text stream and the image stream, we were able to attach descriptions about the players and the plays, which provide more detailed information than the events, to all the live scenes with good success.

3.5 Conclusion

This chapter proposed a method for generating part of the semantic descriptions proposed in Chapter 2, which can be useful for an effective retrieval system by integrating the text and the image streams. We implemented the method and acquired at least either the play, or the player, and attached the descriptions to the proper segments in the video with a precision rate of 86% and a recall rate of 95% for American football videos. As we discussed above, insufficient extraction was rated higher as a result of integration. To prevent this, we should

contrive some way of association through considerations of the context and the structure of the game. Generated descriptions can easily be applicable to standardized MPEG-7 description tools. We also considered it possible to apply our method to other sports videos, which can be structured likewise, and demonstrated the experiment with a baseball video. Overall, the experimental results verified the effectiveness of integrating the text and the image stream, thereby obtaining detailed semantic content in a short period of time: half the length time of the target video, which was remarkably shortened compared to the few days necessary for the manual descriptions, by limiting the use of the time-consuming image processing and putting more value on the text stream analysis.

Note that our method needs to be extended to cover the limitations listed below.

- How to decide appropriate key phrases for each kind of sport. Here, the term *appropriate* means sufficient but not redundant to obtain *all* but *only* the "Live" scenes for every video of the same kind of sports.

- This method uses only a limited part of the closed-caption text, that is, the segments including the sentences with the key phrases. The other parts of the closed-caption text, such as the segments corresponding to the "Replay" scenes should also be used as an information source for semantic content.

- This method only obtains information about a single play and at most two related players. Since there can be several plays in a "Live" scene, and there is other semantic content information to be acquired such as game situations and the score, descriptions including more semantic content are expected.

Chapter 4

Story Segmentation

4.1 Introduction

In Chapter 3, we proposed a method for acquiring the specific information of plays and players for the story units. However, the proposed method can obtain only a limited amount of information: one play and at most two players. Moreover, only a small part of the closed-caption(CC) text, which includes the sentences including the key phrases, has been utilized. Furthermore, the proposed method needs predefined key phrases for each kind of sport. Although the method can be made applicable to several kinds of sports with further work, deciding the key phrases suitable to extract all but only the live scenes is not an easy task. In order to use the CC text more effectively, segmenting it into the scene units and story units will help us decide where we should look for the needed information.

Many researchers have attempted *topic segmentation* or *text categorization* of the automatically transcribed or the CC text of news videos [Shahraray95, Takao01, Zhu01, Greiff01, Mulbregt99, Ponte97]. For example, Shahraray et al. [Shahraray95] proposed an automatic authoring method of hypermedia documents for news videos by segmenting the

CC text correspondingly to the video units. Takao et al. [Takao01] 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. [Greiff01] used the 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 topic boundary. However, due to the relative uniformity of the topics for the sports videos, few researchers have succeeded in semantic topic segmentation of the CC text of sports videos.

In this chapter, we propose a method of segmenting the CC text of sports videos into both the scene and the story units by recognizing certain patterns in the CC text for each scene, without using many domain-dependent key phrases, and of associating the segmented story units to the corresponding video segments. The results of the method will ease the acquisition of important semantic information for each story unit and extend the limitation of the information to be included in the descriptions.

4.2 Video Story Segmentation for Semantic Content Acquisition

Figure4.1 shows the outline of our proposed method. First, in order to make applying the 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. The video stream is segmented into the story units with template matching in the image stream, the same way as proposed in Chapter 3, and then is associated with the segmented story units in the CC text. The details of each step are discussed below.

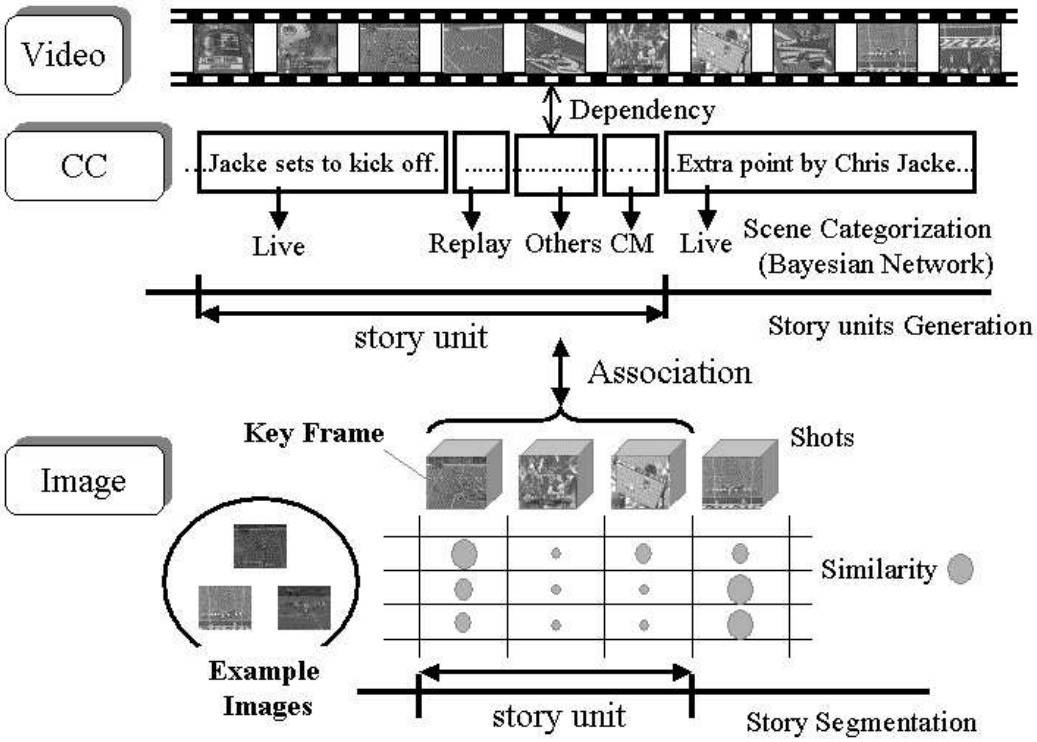


Figure 4.1: Outline of the proposed method

4.2.1 Segmentation of Text Stream

As discussed in Chapter 3, the CC text usually lags behind the actually spoken words in sports videos. This time-lag varies from approximately 0 to 20 seconds, and moreover, the announcers do not necessarily talk about the present scene. Therefore, synchronizing the CC text and video stream is difficult [Hauptmann98]. This method tries to segment the CC text separately from the video stream so that the CC text segments correspond to the video scenes semantically.

Here, let us consider the semantic element, which corresponds to the “shot” in the video stream, in the CC text. The CC text is just a sequence of words and does not have

Table 4.1: 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
Players' names	likely	likely	cannot be determined	rarely
Situational Phrases	highly likely	less likely	probably	rarely

prominent indicators of scene changes. The change of the speaker can be a boundary of the topic. However, the same speaker sometimes keeps talking throughout several kinds of scenes, and in that case, there is no pronounced marking. Therefore, by calculating the time interval between the sentences, 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 sports TV program, the scene categories we try to categorize are the "Live" ("Live"+"Pre-Live" in Figure2.2), "Replay"("Post-Live"), "CM", and "Others" ("Report", "Studio", etc.) scenes.

Table4.1 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 "♯ 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" and "Situational Phrases" respectively, show the likeliness of the appearance of the players' names and *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.

Since the characteristics discussed above are highly ambiguous in categorizing the CC segments to the corresponding scenes, a more precise pattern 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 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.) in order to categorize the CC segments into the four kinds of scenes. Moreover, the structure of the sports TV program shows that the scenes have some rules in how they line up. Therefore, 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 Bayesian Network is a powerful tool for knowledge representation and inference under conditions of uncertainty encoding the conditional dependencies among the various elements [Duda]. The Bayesian Network is a directed acyclical graph(DAG) in which each node represents one of the domain variables, each arc describes a direct relationship between the linked variables, and the strength of these links is given by conditional probability distributions (CP table). The probability of a particular value x for a node X can be calculated as the product of the two factors, one of which depends on the set of children nodes C, and the other, which depends on the set of parent nodes P, as follows:

$$P(x|e) \propto \prod_{j=1}^{|C|} P(e_{C_j}|x)[\sum_{allP_{mn}} P(x|P_{mn}) \prod_{i=1}^{|P|} P(P_i|e_{P_i})], \tag{4.1}$$

where e represents the values of variables on nodes other than X, $C_j(P_i)$, the $j(i)$th child(parent) node, and $e_{C_j}(e_{P_i})$ the values of its state, and P_{mn} denotes a particular value

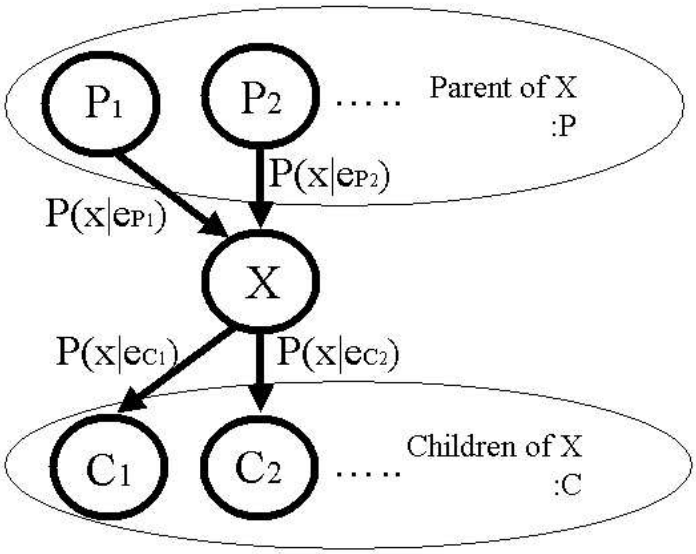


Figure 4.2: Bayesian Network

for state n on parent node P_m. Figure4.2 shows the relationship of the nodes.

The BN in Figure4.3 shows the relationship between the scene category of the present CC segment and other factors. Node **B** represents the scene category of the previous CC segment and is the parent of node **X**, which represents the scene category of the present CC segment. Nodes **Fj**, which are the **j**th children nodes of **X**, represent the features of the present CC 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.

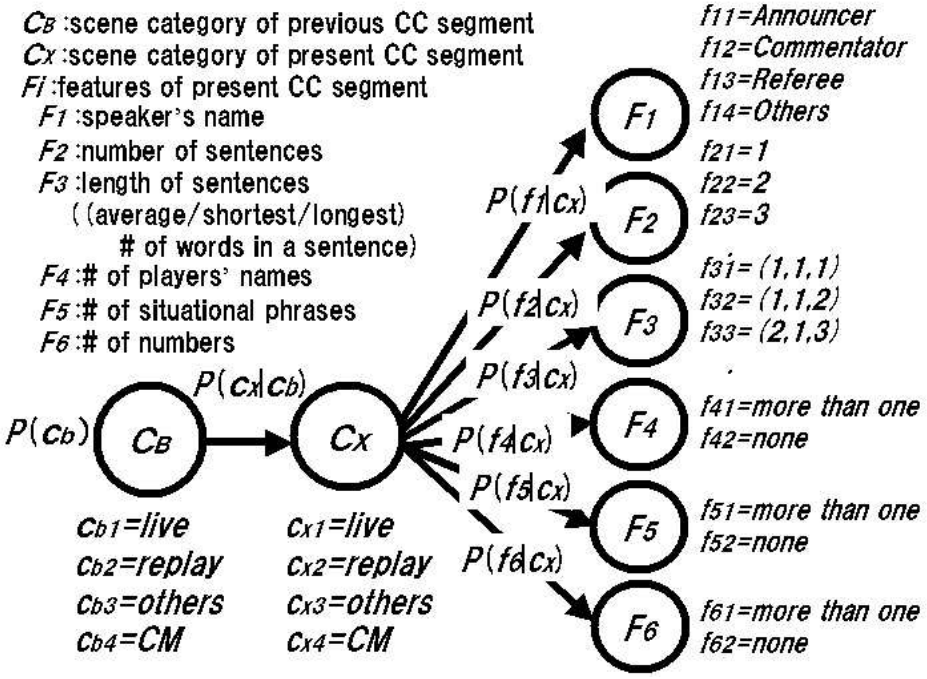


Figure 4.3: Bayesian Network for categorizing CC segments

Based on this BN, we can calculate the probability for the present scene categories as

$$P(x|e) \quad = \quad [\sum_{all\ b} P(x|b)P(b)] \prod_{j=1}^{|F|} P(f_j|x), \tag{4.2}$$

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

$$P(live|e) \quad = \quad [P(live|live)P(live) + P(live|replay)P(replay) + \cdots]$$

$$\times P(announcer|live) \times \cdots. \tag{4.3}$$

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

[Procedure to categorize CC segment]

1) Generate the CP table for every arc 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) Normalize the calculated probability of each scene category, update the $P(b)$ with these values and repeat 2) and 3) for the rest of the CC segments.

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 there are consecutive live segments, we consider them as consecutive separate live scenes if they have an interval more than a threshold between themselves. Otherwise, we determine these live segments as being included in the same live scene.

4.2.2 Association of Text and Video Streams

After segmenting the CC text, we need to attach each CC story unit to the corresponding video segments. However, the boundaries in the CC text are extracted only from speaker changes and blank portions of the talk and do not necessarily indicate the semantic boundaries. Since the video stream is a sequence of camera shots, each of which represents a continuous action in time and space, the boundaries of the shots which can be extracted from the image stream are considered to be more suitable for story-based segmentation than the boundaries extracted from the CC text. Therefore, we also segment the video into

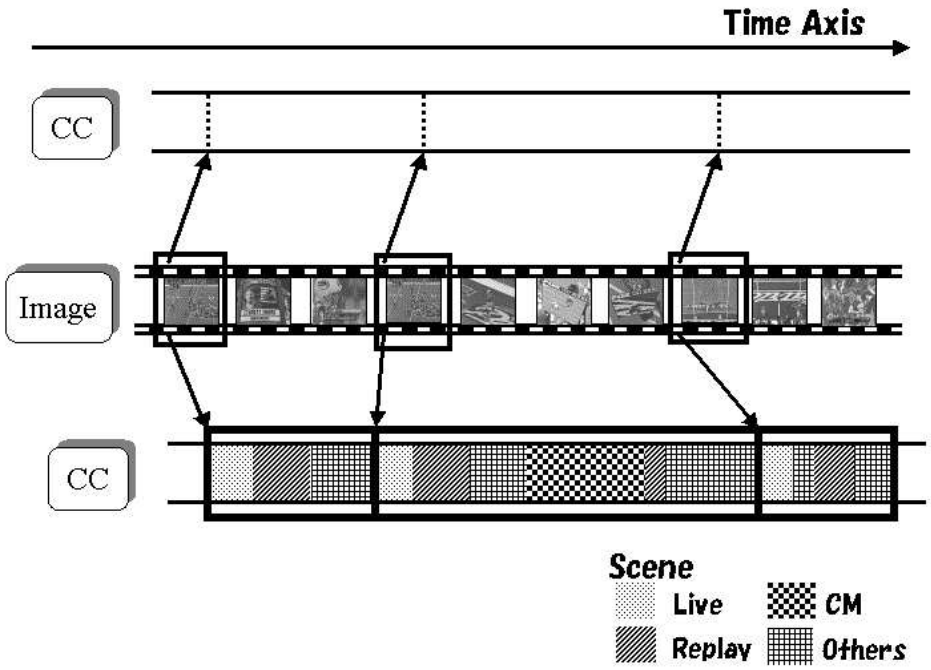


Figure 4.4: Association of CC and video: the CC text in the lower part shows the actual boundaries of the corresponding CC text. Finding the boundaries at the position with the same time lag doesn't necessarily provide the correct boundaries (the CC text in the upper part).

video story units by finding the beginning of the "Live" scene using the method proposed in Chapter 3.

Since the "Timed-Stamp" in the CC text represents the rough time when the sentences were spoken, these stamps enable us to estimate the time when the sentences were spoken on the video. Figure4.4 shows that since the time lag between the CC text and the image stream varies approximately from 0 to 20 seconds, finding the most appropriate CC story unit from all the generated ones (the CC text in the lower part in Figure4.4) should work better than finding the most appropriate boundary from the whole CC text, which is just a sequence of words (the CC text in the upper part in Figure4.4 shows boundaries with the

same time lag to the image stream).

Moreover, it should be noted that the “Live” scene usually occurs during a certain interval. Therefore, the 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) Search the CC story unit, which starts around *Th* seconds after the beginning of a video story unit.
2) If there is more than one CC story unit before the beginning of the following video story unit, associate the present video story unit with the closest CC story unit. Otherwise, divide the previously associated CC story unit at the boundary of the CC segment which is the closest to *Th* seconds after the beginning of the video story unit, and change the scene category of the first scene unit of the divided CC story unit to “Live”.
3) If the following video story unit starts within *Th2* seconds,
   - If the beginning of the following video story unit is closer to *Th* seconds before the CC story unit which was associated with the previous video story unit than the beginning of the previous video story unit, associate the CC story unit with the following video story unit.
   - Otherwise, discard the following video story unit.
4) After repeating 1), 2) and 3) for all the extracted video story units, change the scene category of the first scene unit of each remaining CC story unit to

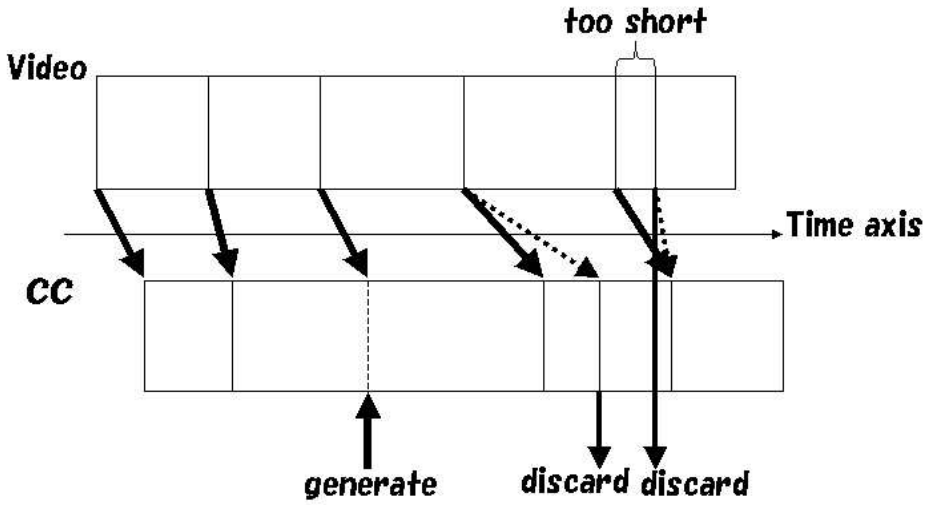


Figure 4.5: Association of CC and video

"Others", and merge the CC story unit with its previous CC story unit.

Figure4.5 illustrates the behavior of this procedure.

4.3 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. 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 is shown below.

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.

$$CC\ Live\ Scene\ Recall = \frac{\sharp\, of\ correctly\ categorized\ \text{“}Live\text{”}\ CC\ segments}{\sharp\, of\ the\ actual\ \text{“}Live\text{”}\ CC\ segments}$$

$$CC\ Live\ Scene\ Precision = \frac{\sharp\, of\ correctly\ categorized\ \text{“}Live\text{”}\ CC\ segments}{\sharp\, of\ the\ CC\ segments\ categorized\ as\ a\text{“}Live\text{”}\ scene}$$

We also evaluated 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:

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

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

4.3.1 Experiments with American Football Videos

Table4.2 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 a blank portion that lasts more than 3 seconds. The situational phrase for American football is "num down (and $num2$)", where num represents either of "1st", "2nd", "3rd" or "4th", and $num2$ represents any number.

Table 4.2: 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

We show the results of the CC scene categorization for each video in Table4.3. 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).

These results indicate that

- The accuracy was 59% on average, and ranged 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 seen in the CC text was the omission of the speaker changes at the time when the program scene changes to the CM scene. As a consequence, “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

Table 4.3: 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)

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 4.4 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 4.3 and Table4.4 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. While 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 could be correctly generated.

Table 4.4: 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)

We next experimented with the video segmentation method. We provided four kinds of images shown in Figure4.6 (1 Scrimmage, 2 KickOff/Punts, and 1 EP/FG which are the same ones used in the experiments in Chapter 3) as the beginning images for each video stream. For EP/FG, we used the images of both sides of the field. We selected the example images considered to be taken from up front of the field as the general camera direction. Table4.5 shows the results of the segmentation.

Table4.6 shows the results of the association of the CC text and the video stream with the parameters $Th = 10$ and $Th2 = 10$. 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).

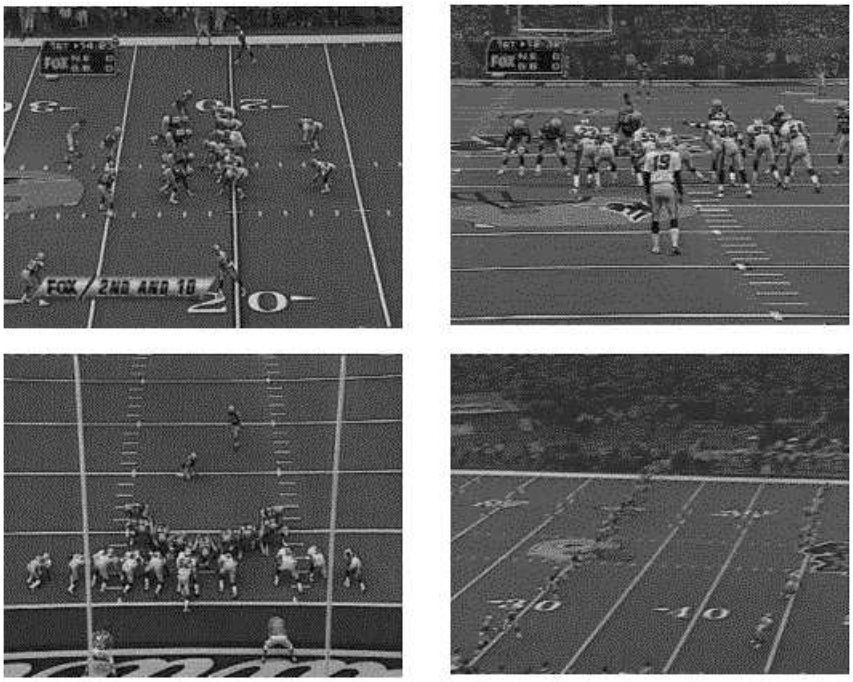


Figure 4.6: Examples of beginning images (American football)

Table 4.5: 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)

These results imply that:

- 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 Table4.6). 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" column 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 can

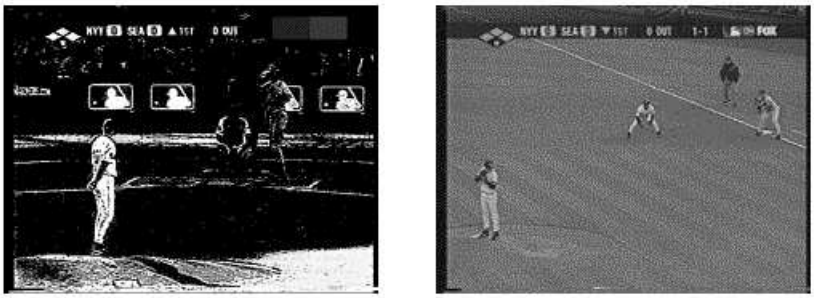
Table 4.6: Results of CC and video integration (American football)

	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

not be recovered with the association. Therefore, the recall rate in the video segmentation should be emphasized more than the precision rate.

4.3.2 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 ball(s) and $Num2$ strike(s)", "full count", and "$Num3$ away(out)", where Num represents the integer from 0 to 4, $Num2$ the integer from 0 to 3, $Num3$ 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 pither posing before throwing the ball was taken from (1) the back of the pither, and (2) the front of the pither so that the image shows both the pither and the player at the base (See Figure4.7). Note that the image(2) is related to the play "Steal Base", which was not



(1) (2)

Figure 4.7: Examples of beginning images (baseball)

considered in the experiments in Chapter 3, and was added as the beginning images.

Table4.7 shows the results of CC scene categorization for each video using learned data from the other 4 video streams. As we have just discussed, 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 data for the CC scene categorization. Of these three, changing the sample data requires much more work than the other two. However, since the sports videos generally consist of four kinds of scenes, which are “Live”, “Replay”, “Others”, and “CM”, and the characteristics of these four scenes discussed in Section 4.2.1 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 the assumption, we have also tested the CC scene categorization for baseball videos using learned data from American football videos. The results are shown in Table4.8.

Table 4.7: 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 4.8: 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)

Comparing Table4.7 and Table4.8 shows us that there is no significant difference between these two experiments. From the comparison, we can infer that since the method uses few domain-dependent features, the sample data for a kind of sports 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 using learned data from the baseball videos for the following experiments.

Table4.9, Table4.10, and Table4.11 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

- 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.

Table 4.9: 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 4.10: 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)

In baseball videos, the announcers often 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, the actual story units which are rather short 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.

Table 4.11: 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

4.4 Discussion

Here, we discuss the potentiality of our method for semantic content acquisition. Figure4.8 illustrates an example of the 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. Table4.12 shows the usability of the CC segments attached to each story unit. In Table4.12, “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 Table4.12, segmentation of the CC text and its association with the video stream allows us to acquire semantic content with the extraction of the simple key words. Note that the key words used here are much more general to the same kind of sports. In

```
<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>
```

Figure 4.8: Example of the final description result

addition, more information can be acquired compared to the method proposed in Chapter 3. Moreover, since the CC story units are segmented into scene units, taking out the "Others" and "CM" CC scene segments filters out the unnecessary information from the CC text and makes the summary of each story unit. These CC segments themselves can be used as the semantic text descriptions for each story unit when used in the same way as the text retrieval [Oard97], and work as a rich information source for a retrieval system.

We next discuss the generality of our method. Most of the sports videos are constructed with four kinds of scenes as discussed in Section 4.2.1: Live, Replay, Others, and CM, and the characteristics of each scene in the CC text, which were discussed also in Section 4.2.1, hold true for most of the sports videos. Our text segmentation method tried to find the story units based on this general structure of a sports TV program and the general characteristics of the CC text. Although the method proposed in Chapter 3 also tried to extract the "Live" segments from the CC text using some key phrases determined for each play, we have to

Table 4.12: 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

pre-define the kind of plays and the key phrases beforehand for each kind of sports. On the contrary, the segmentation method proposed in this chapter used only the players' names and game situation phrases as the domain-dependent information, and as a result, improved applicability to several kinds of sports.

4.5 Conclusion

This chapter proposed a method for segmenting the closed-caption text into scene and story units, and 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 to obtain better results than would be achieved with just 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.

Chapter 5

Conclusion

Effective retrieval, summarization, or filtering systems require semantic representation of the videos. Although the MPEG-7 has been standardized to describe multimedia content in a textual form, which is easily and efficiently usable for the video retrieval system, *what* and *how* to effectively describe the videos need to be defined. In order to answer this *what* question, we also need to consider *where* in the videos the descriptions should be attached. Chapter 2 discussed the common structures throughout several kinds of sports videos: that of the sports program and the sports game, and defined the *story units*, which are the fundamental logical units for the sports videos, and proposed a semantic description model summarizing these semantic structures.

Even though what to describe are defined, manually acquiring the information needed for the descriptions is hugely time-consuming and a method for automatically or semi-automatically generating these descriptions to aid the manual descriptions is mandatory. This method, which answers the *how* question, consists of two steps: 1)Temporal Video Segmentation, and 2)Semantic Content Acquisition. Many researchers have been aware that the videos are generally composed of several information streams: image, audio, and

text streams. Since it is easily predictable that the speech of the announcers can be a rich semantic information source in sports videos, this thesis is mainly focused on the closed-caption text, which is the text transcript of the speech.

Chapter 3 proposed a method for automatically attaching the descriptions about plays and players for each story unit. This method used the closed-caption text for semantic content acquisition and the image stream for video segmentation. The key phrases were used to simplify the text analysis and to obtain only the aimed at information. Due to the results of the experiments with two American football videos, we were able to acquire at least either the play or player and attach them to the correct “Live” scenes with a recall rate of 86% and a precision rate of 95% on average. We also conducted the experiments with a baseball video to examine the generality of our method and obtained reasonable results: the recall rate and the precision rate respectively were 69% and 96%.

These results indicated that the closed-caption text were able to provide us the detailed semantic information, which is hard to acquire from the image stream. The overall time our method takes to generate the descriptions of the video is highly dependent on the time to process the image stream, inferring the effectiveness of the use of the closed-caption text, which can be analyzed with much lower cost, reducing the burden of the image processing on our system. Moreover, the time lags between the closed-caption text and the audio/image stream can be accounted for with the approximate temporal association of the segments acquired from each stream without the precise word-to-word association of audio and text streams. As a result, the acquired information from the closed-caption text can be correctly attached to the semantically corresponding video segments with the proposed simple text-image association method. Convinced of the advantage of the integration of the closed-caption text and the image stream, we now consider making better use of the closed-caption text to acquire more semantic content.

One of the ways to achieve this goal is to attach the semantically corresponding closed-caption segment as a document that includes more detailed information to each video segment. In order to effectively synchronize the closed-caption text and the video stream, Chapter 4 proposed to segment these two streams separately. The proposed method tried to automatically segment the closed-caption text probabilistically with Bayesian Networks by learning the characteristic patterns of each scene in the closed-caption text without using much domain-dependent information, and to associate the results with the video segments attained with image stream analysis. We also conducted the experiments with 10 American football videos, segmenting the video stream into story units and attaching the correct closed-caption story units with a recall rate of 92% and a precision rate of 89% on average. The experiments with 5 baseball videos also resulted in a recall rate of 93% and a precision rate of 94%.

The results showed that the proposed method required less labor to make itself applicable to several kinds of sports because of the lesser use of the domain-dependent information. We can also infer from the results that the corresponding closed-caption story units, which surely include more detailed information as semantic descriptions, help us acquire more information about semantic content. Moreover, segmenting the closed-caption text into the scene units as well as the story units provides us with the location to look for the significant information for each story unit, filtering out the insignificant parts of the closed-caption text.

Next, we consider future work. Although the whole closed-caption segments corresponding to the story units will work as semantic descriptions in the same way as the text retrieval system, more simple descriptions including only the significant information are desirable. One of the ways to fulfill the goal would be 1) keyword extraction, and 2) parsing the structure of the story units with knowledge about the flows of the game. This

challenge will be our foremost task in the future. Moreover, the speaker identification from the audio stream will improve the CC scene categorization, and another way to associate the CC text and the image stream should be examined to make the best of each result.

Although there remains a lot to be done to accomplish the ultimate goal, on the whole, this thesis pointed out the importance of the text stream for semantic content analysis of the sports videos, and the experimental results verified that

- The closed-caption text worked well for semantic content acquisition, and provided us with more detailed information than by just using the image stream.

- Although the proposed method conducted rather simple text analysis, without more complicated processes such as natural language processing, we were able to easily and effectively acquire the significant semantic content from the closed-caption text.

- The image stream provided us with the precise story boundaries, which were difficult to extract from the text stream.

- Integration of the multimodal information streams compensated for the limitations of each stream. Moreover, this integration successfully realized what was considered impossible with the analysis of a single stream.

The results of our proposed method should help to produce MPEG-7 descriptions easily. Due to both the diversity of the semantic content of the videos and the limitation of the video processing technology, we believe that a method which always acquires the correct semantic content is almost impossible to develop. However, even though the proposed basic semantic description model for sports videos has room for improvement, and the proposed method did not achieve perfect results; nonetheless, the generated descriptions saved tremendous time over manually making the descriptions, from a few days to a few hours

for a single sports program. In addition, we showed that the use of the text stream played an important role in reducing time by limiting the labor of image processing throughout the semantic analysis of the video. We hope our work contributes to the development of the field of semantic content analysis of multimedia data in some way.

Bibliography

[Alatan01] 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.

[Babaguchi01] 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.

[Benitez01] A.B.Benitez, D.Zhong, S-F.Chang, and J.R.Smith, “MPEG-7 MDS Content Description Tools and Applications”, *Proc. the International Conference on Computer Analysis of Images and Patterns (CAIP’01)*, 2001.

[Blumin] 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*.

[SFChang01] S-F.Chang, T.Sikora, and A.Puri, “Overview of the MPEG-7 Standard”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol.11 no.6, pp.688–695, June 2001.

[SFChang99] S-F.Chang, Q.Huang, T.Huang, A.Puri, and B.Shahraray, "Multimedia Search and Retrieval", book chapter in *Advances in Multimedia: Systems, Standards, and Networks*, New York: Marcel Dekker, 1999.

[YChang96] Y.Chang, W.Zeng, I.Kamel, and R.Alonso, "Integrated Image and Speech Analysis for Content-Based Video Indexing," *Proc. IEEE International Conference on Multimedia Computing and Systems (ICMCS'96)*, pp.306–313, 1996.

[Dimitrova99] N.Dimitrova, "Multimedia Content Analysis and Indexing for Filtering and Retrieval Applications", *Informing Science Special Issue on Multimedia Informing Technologies-Part 1*, vol.2, No.4, pp.87–100, 1999.

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

[Eickeker99] 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.

[Gao00] 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.

[Gong95] 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.

[Greiff01] 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.

[Hanjalic99] 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.

[Hauptmann98] A.G.Hauptmann and M.J.Witbrock, “Story Segmentation and Detection of Commercials in Broadcast News Video”, *Proc. Advances in Digital Libraries (ADL’98)*, 1998.

[Huang99] Q.Huang, Z.Liu, A.Rosenberg, D.Gibbon and B.Shahraray, ”Automated Generation of News Content Hierarchy by Integrating Audio, Video, and Text Information”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’99)*, vol.6, pp.3025–3028, 1999.

[Jasinschi01] 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”, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’01)*, 2001.

[Javed01] 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.

[Kittler01] J.Kittler, K.Messer, W.J.Christmas, B.Levienaise-Obadia, and D.Koubaroulis, “Generation of Semantic Cues for Sports Video Annotation”, *Proc. International Workshop on Information Retrieval*, pp.171–178, 2001.

[Kwon00] 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.

[Lazarescu99] M.Lazarescu, S.Venkatesh, G.West, and T.Caelli, “On the Automated Interpretation and Indexing of American Football,” *Proc. IEEE International Conference on Multimedia Computing and Systems (ICMCS’99)*, vol.1, pp.802–806, 1999.

[BLi01] B.Li and M.I.Sezan, “Event Detection and Summarization in Sports Video”, *Proc. IEEE Computer Vision and Pattern Recognition (CVPR’01)*, Demos, pp.29–30, 2001.

[YLi01] 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.

[Li00] F.C.Li, A.Gupta, E.Sanocki, L.He, and Y.Rui, “Browsing Digital video”, *Proc. ACM Conference on Human Factors in Computing Systems (CHI’00)*, pp.169–176, 2000.

[Lienhart97] R.Lienhart, S.Pfeiffer, and W.Effelsberg, ”Video Abstracting,” *Communications of the ACM*, Vol.40, No.12, pp.55–62, Dec. 1997.

[Mani97] I.Mani, D.House, M.T.Maybury, and M.Green: ”Towards Content-Based Browsing of Broadcast” *in Intelligent Multimedia Information retrieval*, The MIT Press, pp.241–258, 1997.

[MDS01] 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.

[Miyauchi02] S.Miyauchi, N.Babaguchi, and T.Kitahashi, “Highlight Detection and Indexing in Broadcasted Sports Video by Collaboration Processing of Text, Audio, and Image”, *Transactions of The Institute of Electronics, Information and Communication Engineers of Japan*, Vol.J85-D-II, No.11, pp.1692–1700， 2002 (in Japanese).

[Mulbregt99] P.van Mulbregt, I.Carp, L.Gillick, S.Lowe, and J.Yamron, “Segmentation of Automatically Transcribed Broadcast News Text”, *Proc. DARPA Broadcast News Workshop*, pp.77–80, 1999.

[Nakamura97] Y.Nakamura and T.Kanade, ”Semantic Analysis for Video Contents Extraction – Spotting by Association in News Video.”, *Proc. of 5th ACM International Multimedia Conference (MM’97)*, pp.393–402, Nov 1997.

[Nitta02a] N.Nitta and N.Babaguchi, “Acquisition of Semantic Content of Broadcasted Sports Video based on the Structure Analysis”, *Forum on Information Technology Letters (FIT2002)*, LI-19, pp.161–162, 2002 (in Japanese).

[Nitta02b] N.Nitta and N.Babaguchi, “Automatic Story Segmentation of Closed-Caption Text for Semantic Content Analysis of Broadcasted Sports Video”, *Proc. 8th International Workshop on Multimedia Information Systems (MIS’02)*, pp.110–116, 2002.

[Nitta02c] N.Nitta, N.Babaguchi, and T.Kitahashi, “Story Based Representation for Broadcasted Sports Video And Automatic Story Segmentation”, *Proc. IEEE International Conference on Multimedia and Expo (ICME’02)*, pp.813–816, 2002.

[Nitta02d] N.Nitta and N.Babaguchi, “Semantic Content Representation Model for Sports Video and Automatic Story Segmentation”, *Technical Report of The Institute of Electronics, Information and Communication Engineers of Japan, Pattern Recognition and Media Understanding(PRMU) 2002-20*, Vol.102, No.155, pp.1–8, 2002 (in Japanese).

[Nitta01] N.Nitta, N.Babaguchi, and T.Kitahashi, “Automated Annotation to Significant Scenes Based on Structure of Broadcasted Sports Video”, *Transactions of The Institute of Electronics, Information and Communication Engineers of Japan*, Vol.J84-D-II, No.8, pp.1838–1847, 2001 (in Japanese).

[Nitta00a] N.Nitta, N.Babaguchi, and T.Kitahashi, “Extracting Actors, Actions and Events from Sports Video – A Fundamental Approach to Story Tracking –”,*Proc. International Conference on Pattern Recognition (ICPR’00)*, pp.718–721, 2000.

[Nitta00b] N.Nitta, N.Babaguchi, and T.Kitahashi, “Annotation to Sports Video by Integrating Linguistic and Image Information”, *Proc. of Meeting on Image Recognition and Understanding (MIRU’00)*, pp.I-319–I-324, 2000 (in Japanese).

[Nitta00c] N.Nitta, N.Babaguchi, and T.Kitahashi, “Extraction of Actors, Actions and Events from Sports Video by Integrating Linguistic and Image Information”, *Technical Report of The Institute of Electronics, Information and Communication Engineers of Japan, Pattern Recognition and Media Understanding(PRMU) 99-256*, Vol.99, No.709, pp.75–82, 2000 (in Japanese).

[Nitta99] N.Nitta, N.Babaguchi, and T.Kitahashi, “Extraction of Actors and Actions from Continuous Media by Linguistic Information”, *Proc. of 1999 Institute of Electronics, Information and Communication Engineers, General Conference*, vol.2, p.365, 1999 (in Japanese).

[Oard97] D.W.Oard, “The State of the Art in Text Filtering”, *User Modeling and User-Adapted Interaction*, vol.7, no.3, pp.141–178, 1997.

[Ponte97] J.M.Ponte, W.B.Croft, “Text Segmentation by Topic”, *Proc. 1st European Conference on Research and Advanced Technology for Digital Libraries (ECDL’97)*, pp.113–125, 1997.

[Roach02] M.Roach, J.Mason, N.Evans, L-Q.Xu, and F.Stentiford, “Recent trends in video analysis: A taxonomy of video classification problems”, *Proc. 6th IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA’02)*, pp.12–14, 2002.

[Rui00] Y.Rui, A.Gupta, and A.Acero, “Automatically Extracting Highlights for TV Baseball Programs”, *Proc. ACM Multimedia 2000*, pp.105–115, 2000.

[Satoh99] S.Satoh, Y.Nakamura and T.Kanade, ”Name-It: Naming and Detecting Faces in News Videos”, *IEEE Multimedia*, pp.22–35, 1999.

[Shahraray95] 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.

[Shearer00] 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.

[Shearer99] K.Shearer and S.Venkatesh, “Detection of Setting and Subject Information in Documentary Video”, *Proc. IEEE International Conference on Multimedia Computing and Systems (ICMCS’99)*, vol.1, pp.797–801, 1999.

[Smith97] M. A. Smith and T. Kanade, “Video Skimming and Characterization Through the Combination of Image and Language Understanding Techniques,” *Proc. IEEE Computer Vision and Pattern Recognition (CVPR’97)*, pp.775–781, 1997.

[Sudhir98] 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.

[Takao01] 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.

[Toklu00] C.Toklu, S-P.Liou, and M.Das, “Videoabstract: A Hybrid Approach to Generate Semantically Meaningful Video Summaries”, *Proc. IEEE International Conference on Multimedia and Expo (ICME’00)*, vol.1, pp.57–60, 2000.

[Xu01] 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 International Conference on Multimedia and Expo (ICME’01)*,pp.928–931, Aug. 2001.

[Zhong01] D.Zhong and S.F.Chang, “Structure Analysis of Sports Video Using Domain Models”, *Proc. IEEE International Conference on Multimedia and Expo (ICME’01)*,pp.920–923, Aug. 2001.

[Zhou00] 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.

[Zhu01] 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.

List of Publications

A. Journal papers

1. 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).

2. N.Nitta, N.Babaguchi, and T.Kitahashi, “Automated Annotation to Significant Scenes Based on Structure of Broadcasted Sports Video”, Transactions of The Institute of Electronics, Information and Communication Engineers of Japan, Vol.J84-D-II, No.8, pp.1838-1847, 2001 (in Japanese).

B. Conference papers (refereed)

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

2. N.Nitta and N.Babaguchi, “Acquisition of Semantic Content of Broadcasted Sports Video based on the Structure Analysis, Forum on Information Technology Letters (FIT2002), LI-19, pp.161-162, 2002 (in Japanese).

3. N.Nitta, N.Babaguchi, and T.Kitahashi, “Story Based Representation for Broadcasted Sports Video and Automatic Story Segmentation”, Proc. IEEE International Conference on Multimedia and Expo (ICME2002), pp.813-816, 2002.

4. N.Nitta, N.Babaguchi, and T.Kitahashi, “Extracting Actors, Actions and Events from Sports Video – A Fundamental Approach to Story Tracking –”, Proc. International Conference on Pattern Recognition (ICPR2000), pp.718-721, 2000.

5. N.Nitta, N.Babaguchi, and T.Kitahashi, “Annotation to Sports Video by Integrating Linguistic and Image Information”, Proc. of Meeting on Image Recognition and Understanding (MIRU2000), pp.I-319–I-324, 2000 (in Japanese).

C. Technical papers

1. N.Nitta and N.Babaguchi, “Semantic Content Representation Model for Sports Video and Automatic Story Segmentation”, Technical Report of The Institute of Electronics, Information and Communication Engineers of Japan, Pattern Recognition and Media Understanding(PRMU) 2002-20, Vol.102, No.155, pp.1-8, 2002 (in Japanese).

2. N.Nitta, N.Babaguchi, and T.Kitahashi, “Extraction of Actors, Actions and Events from Sports Video by Integrating Linguistic and Image Information”, Technical Report of The Institute of Electronics, Information and Communication Engineers of Japan, Pattern Recognition and Media Understanding(PRMU) 99-256, Vol.99, No.709, pp.75 - 82, 2000 (in Japanese).

3. N.Nitta, N.Babaguchi, and T.Kitahashi, “Extraction of Actors and Actions from Continuous Media by Linguistic Information”, Proc. of the 1999 Institute of Electronics, Information and Communication Engineers, General Conference, vol.2, p.365, 1999 (in Japanese).