EVENT BOUNDARY DETECTION USING WEB-CASTING TEXTS AND AUDIO-VISUAL FEATURES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MÜJDAT BAYAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2011

Approval of the thesis:

EVENT BOUNDARY DETECTION USING WEB-CASTING TEXTS AND AUDIO-VISUAL FEATURES

submitted by **MÜJDAT BAYAR** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Nihan K. Çiçekli _____
Supervisor, **Computer Engineering Dept., METU**

Examining Committee Members:

Assoc. Prof. Dr. Ferda Nur Alpaslan _____
Computer Engineering Dept., METU

Prof. Dr. Nihan K. Çiçekli _____
Computer Engineering Dept., METU

Asst. Prof. Dr. Sinan Kalkan _____
Computer Engineering Dept., METU

Asst. Prof. Dr. İlkay Ulusoy _____
Electrical and Electronics Engineering Dept., METU

Özgür Alan _____
ORBIM Yazılım A.Ş.

Date: 08.09.2011

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Müjdat Bayar

Signature :

ABSTRACT

EVENT BOUNDARY DETECTION USING WEB-CASTING TEXTS AND AUDIO-VISUAL FEATURES

Bayar, Müjdat
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Nihan K. Çiçekli

September 2011, 64 pages

We propose a method to detect events and event boundaries in soccer videos by using web-casting texts and audio-visual features. The events and their inaccurate time information given in web-casting texts need to be aligned with the visual content of the video. Most match reports presented by popular organizations such as uefa.com (the official site of Union of European Football Associations) provide the time information in minutes rather than seconds. We propose a robust method which is able to handle uncertainties in the time points of the events. As a result of our experiments, we claim that our method detects event boundaries satisfactorily for uncertain web-casting texts, and that the use of audio-visual features improves the performance of event boundary detection.

Keywords: Event Boundary Detection, Shot Detection and Classification, Multimedia Mining, Multimodal Fusion, Video Summarization.

ÖZ

GÖRSEL, İŞİTSEL NİTELİKLER VE İNTERNET KAYNAKLARINDAN ÇIKARILMIŞ METİNLER KULLANILARAK OLAY SINIRLARININ BELİRLENMESİ

Bayar, Müjdat
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez yöneticisi: Prof. Dr. Nihan K. Çiçekli

Eylül 2011, 64 sayfa

Bu çalışmada, görsel, işitsel nitelikler ve internet kaynaklarından çıkarılmış metinler kullanılarak futbol videolarında olay ve olay sınırlarının belirlenmesini sağlayacak bir yöntem sunulmaktadır. İnternet metinlerinden elde edilen olay türü ve düşük hassasiyetli olay zamanı bilgisinin videonun görsel öğeleriyle hizalanıp senkronize edilmesi amaçlanmıştır. "uefa.com" (Avrupa Futbol Birliği resmi sitesi) gibi kuruluşların internet sitelerinde sunulan maç raporlarında olay zamanları saniye yerine dakika hassasiyetinde verilmektedir. Olayların tam gerçekleşme zamanlarındaki bu belirsizlikleri ele alan güçlü bir yöntem bu çalışma ile önerilmektedir. Yapılan deneyler, sunulan yöntemin zaman bakımından hassas olmayan internet metinleri kullanılarak olay ve olay sınırlarını tatmin edici düzeyde belirlediğini ve görsel, işitsel öğelerin olay sınırlarını belirlemede büyük rol oynadığını göstermektedir.

Anahtar kelimeler: Olay Sınırlarının Belirlenmesi, Çekim bulma ve Sınıflandırma, Çoklu ortam madenciliği, Multimodal Füzyon, Video Özetleme.

To my family

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Nihan K. Çiçekli for her invaluable supervision, advice, criticism and support that made the study possible.

I would like to thank Assoc. Prof. Dr. Ferda Nur Alpaslan, Özgür Alan, Samet Akpınar and Orkunt Sabuncu for their precious support and feedbacks. I also thank to my friends and colleagues Soner Kara, Doruk Tunaoğlu, Erkin Eryol and Onur Deniz for their support.

This work is partially supported by The Scientific and Technical Council of Turkey Grant TUBITAK EEEAG-107E234 and by TUBITAK TEYDEB-3080231.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ v

ACKNOWLEDGEMENTS vii

TABLE OF CONTENTS viii

LIST OF FIGURES x

LIST OF TABLES xi

LIST OF ALGORITHMS xii

CHAPTERS

1 INTRODUCTION 1

2 BACKGROUND INFORMATION AND RELATED WORK 4

2.1 Multimodal Fusion 4

2.2 Summarization and Event Boundary Detection in Sports Videos 6

2.3 Shot Detection and Shot Classification 10

3 FEATURE EXTRACTION 15

3.1 Visual Analysis 15

3.1.1 Shot Detection 17

3.1.2 Shot Classification 20

3.2 Audio Analysis 24

3.3 Text Analysis 25

4 EVENT BOUNDARY DETECTION 29

4.1 Fusion of features from Text, Audio and Video 30

4.2 Rule Based Method 32

4.3 Classification Based Method 35

4.3.1 What is One Class SVM? 37

4.3.2 Why use One Class SVM? 38

4.3.3 Training and Classification 38

5 EXPERIMENTS AND EVALUATION 41

5.1 Shot Detection 41

5.2 Shot Classification 43

5.3 Event Boundary Detection 44

5.3.1 Rule Based Event Boundary Detection 45

5.3.2 Classification Based Event Boundary Detection 51

6 CONCLUSION AND FUTURE WORK 58

REFERENCES ……...61

LIST OF FIGURES

FIGURES

Figure 1 – Multimodal fusion methods. 6

Figure 2 – Three consecutive shots and two shot boundaries between them [left to right]. 16

Figure 3 - Far view, Medium view and Close-up view [left to right]. 17

Figure 4 – An example transition with a logo between two shots. 20

Figure 5 – Example images of edge pixels. Far view, Medium view, Close-up view [left-to-right]. 22

Figure 6 – Decision tree for shot classification. 23

Figure 7 – An example match report from sporx.com. 26

Figure 8 – An example match report from uefa.com. 27

Figure 9 – Information extraction from text. 28

Figure 10 - An example shot sequence of a goal event. F: Far view, F (SL): Short Length Far view, C: Close-up view, M: Medium view 30

Figure 11 – Rule Based Event Boundary Detection. 35

Figure 12 – Classification Based Event Boundary Detection 36

LIST OF TABLES

TABLES

Table 1 – Results of Automatic Shot Boundary Detection. 42

Table 2 - Shot Classification Accuracy for Three Types of Shots. 43

Table 3 – Event Detection Rates without and with Sound Amplitude. 46

Table 4 – Rule Based Event Boundary Detection Results of Goal Events. ... 47

Table 5 - Rule Based Event Boundary Detection Results of Corner Events. 48

Table 6 - Rule Based Event Boundary Detection Results of Missed Goal Events. 49

Table 7 - Rule Based Event Boundary Detection Results of Red/Yellow Card Events. 50

Table 8 - Rule Based Event Boundary Detection Results of Penalty Events.50

Table 9 – Cross Validation Accuracies. 52

Table 10 - Classification Based Event Boundary Detection Results of Goal Events. 53

Table 11 - Classification Based Event Boundary Detection Results of Corner Events. 53

Table 12 - Classification Based Event Boundary Detection Results of Missed Goal Events..... 54

Table 13 - Classification Based Event Boundary Detection Results of Red/Yellow Card Events. 55

Table 14 – Detection Rate using reference shot and the comparison with the proposed method. 56

LIST OF ALGORITHMS

ALGORITHMS

Algorithm 1 – Shot Detection……19

Algorithm 2 – Shot Classification……24

CHAPTER 1

INTRODUCTION

An extensive amount of multimedia content become available on the internet and broadcast in the last decade. Creating searchable multimedia archives becomes an important requirement for different domains as a result of the increase in the amount of multimedia content. Instead of watching the whole video, most of the people prefer accessing important events. In addition to that, they would like to search specific events in a video. For providing the viewers a feature like this, video analysis is required. Especially, the widespread popularity of soccer broadcasts makes automatic annotation essential for querying the semantic content of soccer videos.

The annotation provides the means for retrieving specific events in the soccer videos, such as goals, fouls, penalties, bookings etc. Events have duration, therefore they cannot be defined by an exact time point; instead, a time interval is required. The problem of event boundary detection is to determine the boundaries of the periods in which events occur. The performance of this detection basically depends on the usage and fusion of different kinds of information sources such as web-casting text and audio-visual features.

The starting point of this study was to create a framework for automatic annotation of Turkish Super League soccer matches. The idea was to fuse information from different sources. These sources are audio-visual features from soccer videos and textual features from web-casting text that is the text of the matches reporting events minute-by-minute. In this way events and event boundaries are detected and additionally videos are annotated event by event. Here the problem is the synchronization of text and the video. Similar fusion methods in the literature use more accurate time information extracted from textual resources. They can find the exact event moment in the video. Web-casting text for Turkish Super League soccer matches are in minute precision (not second) so that it is difficult to synchronize text and the video. The same case exists in European Champions League soccer matches which were added to the scope of the thesis later. A study of us on this topic including Turkish Super League soccer games is published and appears in the proceedings of 2010 IEEE International Conference on Multimedia & Expo (ICME 2010).

In this thesis, we propose a new multi-modal method using web-casting texts and audio-visual features to detect events and event boundaries in soccer videos. The main issue of this method is to align the web-casting text with the visual content, and this is difficult because of the inaccurate time information in the web-casting text. We overcome this issue by utilizing not only the textual and visual information, but also audio features. As mentioned above, the existing methods assume that the time at which an event occurs is given precisely (in seconds). Therefore, they only focus on detecting time boundaries. However most web-casting texts presented by popular organizations such as uefa.com (the official site of Union of

European Football Associations) provide the time information in minutes rather than seconds. We propose a robust method which is able to handle uncertainties in the time points of the events. In this method, audio-visual analysis is done on the soccer video and information is extracted from web-casting texts that are match reports having the events minute by minute. Our aim in this work is to increase the reliability of web-casting text that is not precise and to decrease false detections if text is not well synchronized. As a result of our experiments, we claim that our method detects event boundaries satisfactorily for uncertain web-casting texts. Additionally, we show that employing audio features improves the performance of the event boundary detection.

The rest of the thesis is organized as follows. In Chapter 2, the background information is given and the related work is explained. The extraction of features from different modalities is described in Chapter 3. In Chapter 4, our approach for Event Boundary Detection is presented. The results of the experiments are shown in Chapter 5. In Chapter 6, conclusions and the future work are presented.

CHAPTER 2

BACKGROUND INFORMATION AND RELATED WORK

Multimedia analysis has been a popular research area in recent years. Researchers extract as much information as possible from multimedia data. They have several purposes in studying the multimedia content, such as classification, summarization, annotation of videos, pictures etc. Summarization of videos is very important for the sports industry, especially for soccer games. Because of this, summarization of soccer videos is a good research topic to work on to meet the needs of this industry.

Multimedia analysis for videos is done by extracting several forms of data (i.e. text, image and sound) and fusing them. It is called multimodal fusion. In this chapter, multimodal fusion for multimedia analysis is described, and then the literature on summarization of sports videos is presented and finally the related work about basic video analysis is described.

2.1 Multimodal Fusion

The very first step in the analysis of multimedia data is to extract useful data from different media that composes the multimedia data. Each media provides us different clues about the content of the multimedia. The fusion of these clues to generate the desired information is called multimodal fusion. The fusion of different modalities increases the accuracy of decision

making process. The first important point is to extract right features from different modalities. A survey [1], which is published as a state-of-the-art overview of fusion strategies, classifies the features into five categories: visual features, motion features, text features, audio features, metadata. They can be shortly summarized as:

- *Visual and motion features.* Visual features can be color based features like color histogram. It can be based on shape, segmented images, texture. These features are extracted from images. Motion features are similarly extracted using consecutive images rather than a single image. These are pixel variation within a shot, motion direction, and optical flows.
- *Text features.* These are the texts extracted from video or can be extracted using a speech recognizer. It can be the closed caption text from videos too. Web resources may also be used to get text data.
- *Audio features.* These features can be directly recognized speech or can be the extracted data using some audio processing methods. Simply the pitch may be extracted. Non-silence ratio is one of the features that can be extracted from audio data.
- *Metadata.* Metadata features are the information created during production. The name, the time stamp, the duration are some of these features.

As stated above, the first step is to decide what to fuse. The next step is how to fuse the extracted data. Atrey *et al.* [1] summarizes the existing methods as shown in Figure 1.

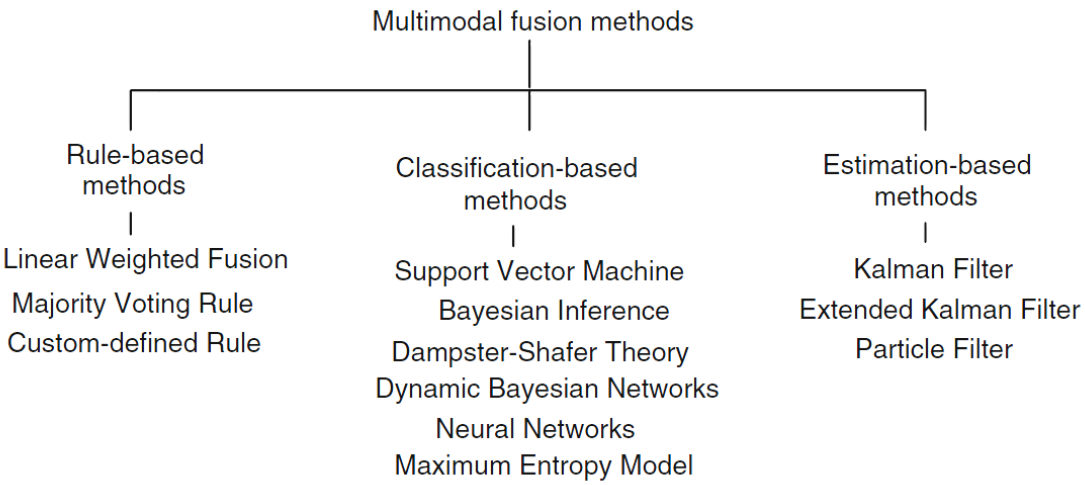

Figure 1 – Multimodal fusion methods.

There are mainly rule-based methods [26, 27], classification-based methods [28, 29] and estimation-based methods [30]. Rule-based methods combine the features using some pre-defined rules. Classification-based methods use classification techniques to classify features extracted from multimedia to one of the predefined classes. Estimation-based methods are used for estimation by fusing multimodal features. For example it can be used in object tracking. These methods can be used together to make a hybrid method. In our study, two separate methods, rule-based and classification-based, are proposed and used on the same experimental data, and the results are compared.

2.2 Summarization and Event Boundary Detection in Sports Videos

The broadcasting of sports events emerged the needs for summarizing these broadcasts. The people following the sports events want to see only important minutes of the event or they want to search specific events in the

video. It is time-consuming to summarize sports videos manually. The idea of automatic summarization of sports videos attracted the interest of many researchers. Extracting highlights of sports events basically includes finding exciting and important minutes of the video. At first, researchers have studied only the extraction of highlights. Later on, detailed summaries and annotation of videos are needed. These needs are labeling the events on video and detecting the boundaries of each event correctly.

A common approach to event boundary detection in soccer games is to utilize visual features only [2, 3, 4]. In these approaches very detailed visual information is extracted. Since videos contain a great amount of visual information, it is a very costly process to extract all this information. Additionally, it is also hard to detect semantic events and event boundaries in this context. The methods using only visual features are suitable for sports whose videos have simple events and few camera views. This approach is applied to tennis videos in [2]. First of all, the video is divided into shots. In these shots view recognition is done to catch the moments of "serve" in tennis games. Motion features and text features are also extracted to find the commercials. Then these frames and shots are merged to a game because there is a serve in each game. Games are merged into the set and sets are merged to the match. The summary is produced through this hierarchical tree.

Another work for summarization focuses on replay scenes [3]. Replay scenes are detected at first, and then the summary is constructed around these scenes. Highlights are constructed merging these replay scenes or the live scenes around the replay scenes. The most cited study about this

approach is the work of A. Ekin and M. Tekalp [4]. The study is about event detection in soccer videos. They decompose the videos into shots at the beginning. As the second step, they classify these shots into some classes like global view, medium view and close up view. The study focuses on goal event detection. Some defined rules are used for detection. The sequential shots are considered. The class of the shots and some patterns are evaluated using some rules. Shots extracted using the rules are merged to construct the summary of the soccer video. In this work shot boundaries are also the boundaries of the events.

Assfalg *et al.* [5] model the highlights in their work. They model each event so that the summarized video is annotated. Visual features extracted in their study are ball motion tracking, playfield detection using grass color and detection of players. Modeling is done with finite state machine. In the finite state machine edges are considered as events. It is a good method to annotate videos but computation cost may be high.

These approaches are suitable for extracting good summaries but getting more information is not easy. For example, in a soccer game, a goal event is extracted but it is impossible to extract the name of the scorer. As a result, the extracted features are insufficient to meet user expectations for rich semantics. These approaches do not handle the issues such as event semantics like the goal scorer and how it is scored, exact event boundaries and accurate annotation. Fusing more data from different sources is needed to get more detailed information.

The alternative approach is to fuse the information extracted from textual and audio-visual sources. Textual resources may be the text extracted from the video or text from web resources. Web-casting texts are textual resources that describe sports events minute-by-minute. In these multimodal fusion methods, texts are used to extract high level (semantic) events and their time while visual sources are used to detect boundaries of these events.

Fusion of several modalities is applied to soccer domain in [6, 7]. Xu *et al.* [6] propose a framework for the alignment of web-casting text and the video. Firstly, shots are extracted and classified. Texts are extracted from web-casting text or closed caption text. Text is used to determine the moment of events and the other semantic content. The type of the event is also extracted from text (e.g. Goal). The moment of the event is marked in the video and the shots around the marked point are merged using a Hidden Markov Model (HMM) and finally event boundaries are found. For each event a model is constructed and the event type extracted from text is used to know which model to use. Similarly, Xu *et al.* [7] apply the same approach to live broadcasts with closed caption texts. In both of these studies, the textual sources include exact time points of events with an accuracy measured in seconds. Thus, the synchronization of text with video becomes easier. However, most textual sources do not include high precision information in terms of time.

Additionally some studies use audio features beside the visual features to find exciting events in a sports video [8, 9, 10]. Chen *et al.* [8] uses the basic visual features like nearly all studies do. They extract shots and classify

shots. They also execute an excitement descriptor and find the exciting moments. They combine visual features and audio features to summarize the video. Radhakrishnan *et al.* [9] and Rui *et al.* [10] focus on audio features. Both studies extract several audio features like energy, noisy speech recognition, background noise recognition. Again the two studies use these features with visual features to generate highlights.

The proposed method in this thesis is a multimodal approach using audio, video and web-casting text features to detect event boundaries in videos of soccer games. It uses three modalities and differs from the existing multimodal approaches in terms of the precision and reliability of web-casting texts. In existing methods, the time of events is assumed to be given precisely and this time is extracted from the text in seconds. On the other hand, in this thesis, the time of the events is extracted from web casting texts in minutes and this imprecise time information is used to detect the exact event boundaries. The proposed method enables to use match reports presented minute by minute in most of web sources for the detection of event boundaries.

2.3 Shot Detection and Shot Classification

In video analysis, the first and the main step is to divide the video into shots. In the next step, these shots are classified according to the purposes of the application. In this section the related work on these two steps (i.e. shot detection and shot classification) is summarized.

Shot is the smallest meaningful piece of video including only one camera view. When the camera view is changed, the existing shot ends and a new shot begins. There are many methods proposed for shot detection. All shot detection algorithms use differences between consecutive frames. During a shot, frames do not change rapidly. Since the camera is steady in a shot, only moving objects change the scene. This corresponds to a lower frame difference. However, the frame structure changes when the shot changes.

Main shot detection approaches are Pixel-wise comparison [11, 12], Histogram Comparison [13, 14], Edge Tracking [15], Motion Vector [16] and combinations of these methods [17]. In pixel-wise comparison, the number of pixels that change in value for a defined threshold in two consecutive frames is calculated. Although it is powerful to detect hard cuts, it is influenced negatively by the motions in the video since the pixels change rapidly in videos having moving objects. As a result of this, pixel-wise comparison methods are not suitable for soccer videos.

In edge tracking, the ratio of the edges going in and out of the frame is calculated. Again motions affect the method negatively. This effect results in doing more calculations which can lead to slow processing in real time video analysis.

The techniques using motion vectors [16] are utilized to detect zooms and gradual transitions. They are used to help other techniques to detect whether a shot includes motion, zoom, or gradual transition.

The most common shot detection technique is histogram comparison and the mostly used histogram is the color histogram. In contrast to pixel-wise comparison, the colors entering and exiting the frame are subject to comparison and the color content does not change rapidly during the shot. This technique is more motion independent since the color histogram is more stable than the pixel values despite the motions.

All of these techniques are compared in the studies of Lienhart *et al.* [13] and Boreczky *et al.* [14]. Comparisons show that the color histogram difference technique is the simplest and the most accurate one to detect shots. It has high detection rates on all sorts of videos. There are some combined methods having slightly higher detection rates. But, this technique is easy to implement and has low computational complexity. Therefore, we adopted color histogram based technique mentioned in [13] and improved it for soccer videos.

Most of the soccer shot classification methods [4, 6, 15, 18, 19] use similar features and combine them. These features are mainly color distribution, edge pixels, and object segmentation. Color distribution is used for finding the colors that are dominant in the image. Field ratio is calculated in soccer shots. Edge pixels are the pixels in the edges of the objects in the image. The edges provide clues about the size of the objects. Object segmentation is used to detect objects in an image. Object segmentation is used to determine the location, the color and the size of the objects in an image. These features can be combined to classify shots.

The main shot classes are far view (global view), medium view and close-up view [4, 6] in soccer videos. The most popular classification feature is the grass ratio [4, 6] in soccer shot classification. Since the green color is dominant in soccer videos and shots, the distribution of colors is important in shot classification. For instance, green color is everywhere in a far view; and the color of a player's kits covers half of the frame in a close-up view. Therefore, the color distribution gives clues about a shot class.

Some approaches use object segmentation for utilizing this fact [15]. In object segmentation, green color is used for field extraction and then the objects in the field area are segmented. As a result, the size of the segmented object determines the shot class. In far view, shot objects are too small. In close-up they are so big, and in medium view, objects are mid-sized. This method has a negative aspect that all soccer fields may not be of the same color. As the color changes according to the angle of light and weather conditions, false detections can be obtained.

Edge pixel distribution is another feature used to classify shots [6]. This method is color independent. In this approach edge pixels are counted and edge pixel count varies according to the camera view. In close-up view, the number of edge pixels decreases. In medium view, edge pixel count increases because several players are included in the scene. Using edge pixel count is not preferable for far view because the entire field is sometimes included in this view and edge pixel count decreases. Moreover, the appearance of supporters in the scene may increase the edge pixel count and divert the method.

The study of Tong *et al.* [18] extracts features like field-ratio, head (of players) detection, object sizes. Object sizes provide info about the camera position in terms of far, medium or close. Zhou *et al.* [19] uses a combination of color distribution, edge distribution and shot length. These features are given to an SVM (Support Vector Machine) as inputs and shots are classified.

The color histogram is used to calculate the dominant color in our shot classification approach. A color histogram difference value is also calculated. Additionally the edge pixel count is used to classify shots. In the shot classification method presented in this thesis, all these three main features are combined.

CHAPTER 3

FEATURE EXTRACTION

In this chapter, we explain how the features are extracted from different modalities. In our approach, three modalities are used. These are video, audio and text. Video and audio are the multimedia content of the soccer match which is recorded and broadcasted. Text source is the match report published on the Internet. The features extracted from these sources are described in the following sections. We can group the analyses performed into three categories which are Visual Analysis, Audio Analysis and Text Analysis.

3.1 Visual Analysis

Video analysis is done by applying several image processing methods on the video. The smallest piece of the video is called *frame* that is an image containing the visual information of a moment in the video. A video is constructed with consecutive images (usually 25/30 frame per second in a broadcast). A frame keeps lots of visual information but only a single frame does not mean anything to us. In a video, a frame gains meaning with the preceding frames. A continuous video piece that runs for an uninterrupted period of time is called a shot. A shot is composed of a series of similar frames. We can say that a shot is the smallest meaningful piece of video

recording, including only one camera view. When the camera view is changed, the existing shot ends and a new shot begins.

In video analysis, the first process is *shot detection* where hard cuts on the video are detected and shot boundaries are labeled. A few example shot boundaries are illustrated in Figure 2. Consecutive frames are shown left-to-right in the figure and red lines represent the shot boundaries (hard cuts in the video).

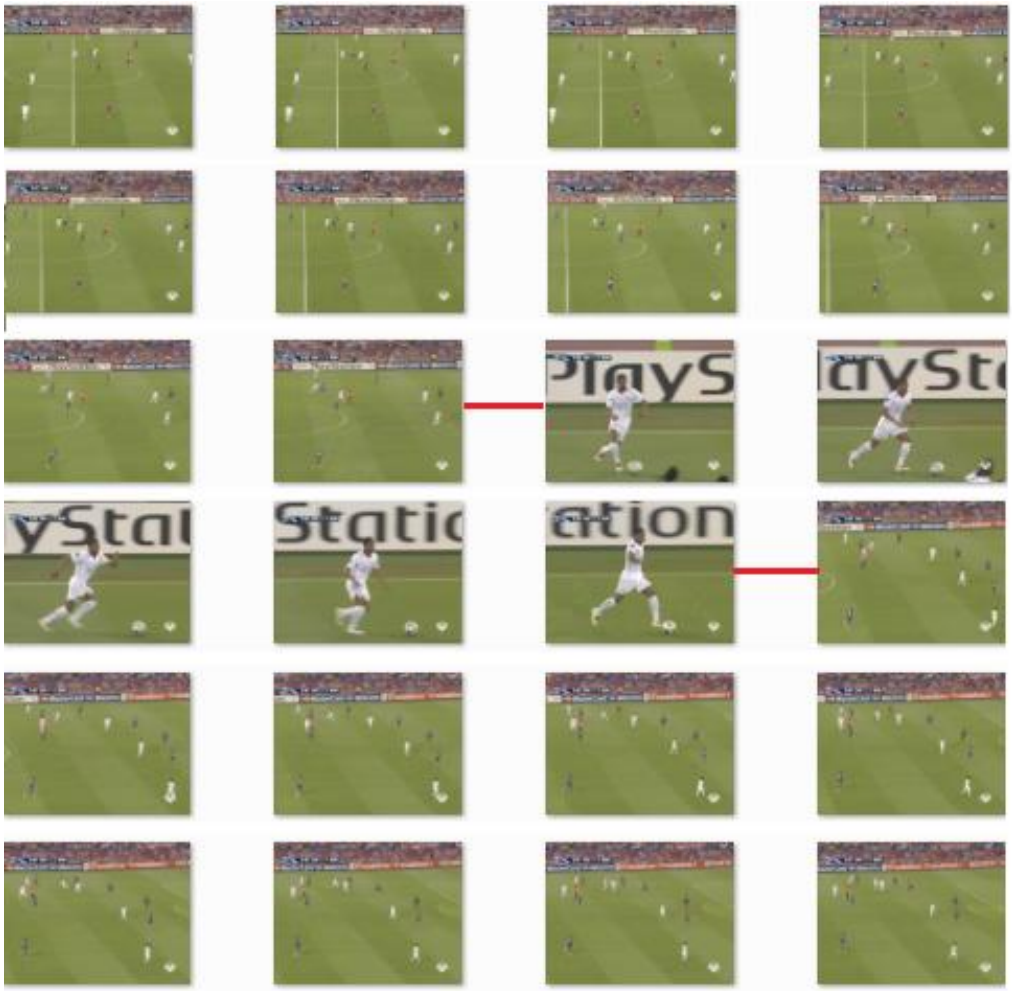

Figure 2 – Three consecutive shots and two shot boundaries between them [left to right].

The second step after shot detection is *shot classification*. Since a single shot includes a single camera view, we can classify the shots according to the camera views. In a soccer broadcast, the main camera views are 'far view', 'medium view', and 'close-up view' [6]. Far views are recorded by main (global) cameras from which you can see nearly all the pitch, while close-up views are products of cameras focusing on players. The medium views have characteristics between the far views and close-up views, and these scenes include several players in action. Each camera view is illustrated in Figure 3.

Figure 3 - Far view, Medium view and Close-up view [left to right].

3.1.1 Shot Detection

In shot detection, the key point is to catch hard cuts in the video where the camera view is completely changed. During a shot, the camera records the scene continuously so that similar consecutive frames are produced. Two consecutive frames in a shot are pretty much alike and they have similar color distribution, similar objects and similar visual characteristics. When a

shot ends and a new shot begins, the frame in the new shot differs a lot from the frame in the preceding shot. We can find hard cuts in the video which are shot boundaries, by comparing consecutive frames visually. As it is mentioned before, there are several comparison methods like pixel-wise comparison, edge tracking and histogram comparison. Histogram comparison is a simple and effective method to find shot boundaries.

We use histogram difference to compare frames in our shot detection method. The color histogram of each frame is calculated in RGB color space. In order to obtain more generalized results, R, G, B values are discretized. Each R, G, B value is down sampled to 3 bits from 8 bits. Then, we have 2^3x2^3x2^3 = 512 color samples. Each occurrence of a color is counted in a frame by traversing all the pixels. Eventually, the histogram of the current frame is obtained. This method is applied to each frame we are working on. The next step is to calculate the histogram difference. All color count values of two consecutive frames are subtracted and the sum of all differences is calculated. This result is divided by the number of total pixels, *N*, and the color histogram difference value is obtained.

Let $F_i(r, g, b)$ be the number of pixels having color vector (r, g, b) in the i^{th} frame of video having *N* pixels. *D* denotes our color histogram difference value, which is calculated by the following formula:

$$D = \frac{1}{N} * \sum_{r=0}^{7} \sum_{g=0}^{7} \sum_{b=0}^{7} |F_i(r,g,b) - F_{i-1}(r,g,b)| \tag{3.1}$$

When the color histogram difference value, D, is above a certain threshold for soccer videos, the current frame (F_i) is meant to be the first frame of the new shot. The previous frame (F_{i-1}) is labeled as the last frame of the previous shot. Here the video time of the shot boundaries are also saved. The shot detection algorithm is summarized in Algorithm 1.

Algorithm 1 Shot Detection

```
1: T ⇐ ThresholdForShotChange
2: for i = 1 to TotalFrames do
3:     CH[i] ⇐ GetColorHistogram(Frame[i])
4:     D ⇐ CalculateColorHistogramDifference(CH[i], CH[i−1])
5:     if D > T then
6:         Label(Frame[i − 1], "EndOfShot")
7:         Label(Frame[i], "BeginningOfnewShot")
8:     end if
9: end for
```

In the difference calculation, we choose consecutive frames with k-distance (i.e. every fifth frame where k is 5) in order to avoid false detections resulting from gradual transitions. Gradual transition is a smooth transition between two shots where a hard cut does not exist. A transition with a logo animation is also applied in soccer broadcasts. An example of transition with a logo is shown in Figure 4. In gradual transition cases, it is possible to miss shot transitions using a small *k*. It is also possible to make false detections with a higher *k*. The optimal value that we decided on for this parameter is 5. Despite these optimizations, false detections have occurred in some cases. To overcome this issue, a rule is defined. This rule says that a shot cannot last less than half second. If a shot lasts less than half second, it

is a false detection. In this case, we ignore the detected shot boundary and continue with the current shot.

Figure 4 – An example transition with a logo between two shots.

In shot detection, the dominant color and edge pixel count are also calculated in addition to the color histogram for each frame. After the shots are detected, the mean of these features are calculated for each shot in order to be used in shot classification.

3.1.2 Shot Classification

There are mainly three camera views in soccer videos as illustrated in Figure 3. We classify the detected shots according to these camera views.

We propose a new approach for shot classification. It is not fully color independent, but ignores small field color differences. First, we classify shots as field and non-field. During shot detection, total histogram, average

color histogram difference, and the average number of edge pixels are calculated. If a new shot is detected, previous shot's total histogram is evaluated. The dominant color in the shot is obtained using this histogram. The dominant color classifies a shot to be field or non-field. If grass color ratio is high, it is field, otherwise it is non-field. Average color histogram difference is used to get information about the focus of the camera. Edge detection is done for each frame and the edge pixels are counted. The number of edge pixels for each frame is averaged within a shot and average edge pixel count is obtained. *Canny Edge Detector* is used for edge detection.

Canny Edge Detector is a very popular and effective edge detector that is used in many computer vision algorithms as a pre-processing step. The Canny edge detection was developed by John F. Canny in 1986 [20]. Although his work was done in the early days of computer vision, the canny edge detector is still a state-of-the-art edge detector. It is a multi-stage detector which performs smoothing and filtering, non-maxima suppression, followed by a connected-component analysis stage to detect edges, while suppressing non edge filter responses. The effect of the canny operator is determined by three parameters: the width of the Gaussian kernel used in the smoothing phase, and the upper and lower thresholds used in connected component analysis stage. These three parameters are defined in our implementation as 16 for Gaussian kernel width, 5 and 2.5 for the thresholds. Example outputs of our canny edge detector can be seen in Figure 5. An image for each shot class is shown respectively.

Figure 5 – Example images of edge pixels. Far view, Medium view, Close-up view [left-to-right].

For field shots, we observed that far view shots have nearly zero color histogram difference value. In far view, moving objects are the players who are too small in the scene. Thus, it does not cause a radical change in the histogram. In contrast, in medium view and close-up view, players are bigger in the scene and histogram changes increase a little more because of the movement of players. Therefore, we used color histogram difference to distinguish far view shots and the others. Medium views and close-up views are distinguished by edge pixel count with a threshold. If the edge pixel count is below the threshold, we label the shot as close-up view. If the edge pixel count is above the threshold, it is labeled as medium view. The camera focuses on an object in close-up view shot. One player in the frame makes the shot to have lower edge pixel count. In medium view shots, there are several objects in the scene so the number of edge pixels increases.

In non-field shots, it is impossible to have a far view shot because the far view camera always displays the entire play which includes most of the field. Therefore, non-field shots are classified as medium view or close-up view only. We use edge pixel count for this operation too, but we use a different threshold value. Non-field shots have a higher threshold because non-field objects are included in the frames. If the edge pixel count is below this higher threshold, we label the shot as close-up view. If the edge pixel

count is above the threshold, it is labeled as medium view. These threshold values for edge pixel count are determined experimentally. Figure 6 depicts the overall method and a summary of shot classification process is given in Algorithm 2.

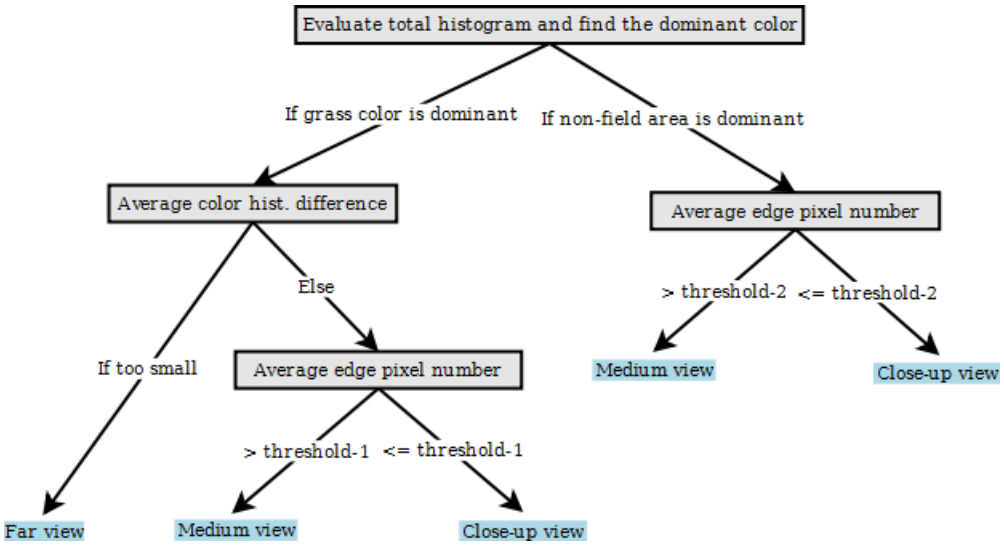

Figure 6 – Decision tree for shot classification.

Algorithm 2 Shot Classification

```
1: TF ⇐ ThresholdForFarViewIdentification
2: T1 ⇐ ThresholdOfEdgePixelNumber
   ForGrassDominantShots
3: T2 ⇐ ThresholdOfEdgePixelNumberForNon −
   grassDominantShots
4: for i = 1 to TotalShots do
5:     TCH[i] ⇐ GetTotalColorHistogram(Shot[i])
6:     EPav ⇐ GetAverageEdgePixelNumber(Shot[i])
7:     DC ⇐ FindDominantColor(TCH[i])
8:     if DC = GrassColor then
9:         CHDav ⇐
           GetAverageColorHistogramDifference(Shot[i])
10:        if CHDav < TF then
11:            Label(Shot[i], "FarView")
12:        else
13:            if EPav > T1 then
14:                Label(Shot[i], "MiddleView")
15:            else
16:                Label(Shot[i], "Close − upView")
17:            end if
18:        end if
19:    else
20:        if EPav > T2 then
21:            Label(Shot[i], "MiddleView")
22:        else
23:            Label(Shot[i], "Close − upView")
24:        end if
25:    end if
26: end for
```

3.2 Audio Analysis

Audio is an important part of soccer broadcasts. Commentary voice and the noise of supporters are included in the broadcasts. The commentary voice and the noise of supporters increase according to the exciting events in the match.

The proposed method for audio analysis aims to get clues about the major events that supporters and the speaker get excited. For this purpose, we

extract sound amplitude value for each second of the soccer video. To calculate this value, we use *rms* (root mean square) of the audio samples in a second. This value is calculated for each second of the audio. Below is the formula to calculate the sound amplitude value:

$$rms = \sqrt{\frac{\sum_{i=0}^{n-1} x_i^2}{n}} \tag{3.2}$$

Here x_i indicates the i^{th} sample in the data-set and n indicates the number of samples in a second. All samples in a second are squared and then summed up. The result is divided by the number of samples in a second. Finally the root of this value is calculated and *rms* of the audio samples is obtained.

After the shots are detected, an average sound value is calculated for each shot by averaging the sound amplitude value of each second of the shot. Audio amplitude increases during exciting events such as goals and missed goals so that we use the shots having high sound amplitude to detect event and event boundaries of exciting events. Audio amplitude is a simple calculation but this feature is so important when combined with the other features.

3.3 Text Analysis

Textual information describing the high level events in a video is an important source for assisting video semantic analysis. In soccer domain,

textual information is mostly extracted from web-casting texts given in the form of match reports. Popular sports web sites (uefa.com, fifa.com, sporx.com etc.) and internet media provide adequate amount of soccer match reports. These match reports focus on the events of soccer games and give detailed information such as the time of the event, the type of the event, actors of the event. Web-casting texts are presented in different languages. We use English and Turkish match reports (sources are from uefa.com, sporx.com) in our work. Example of web pages for match reports of sporx.com and uefa.com are shown in Figure 7 and Figure 8 respectively.

Figure 7 – An example match report from sporx.com.

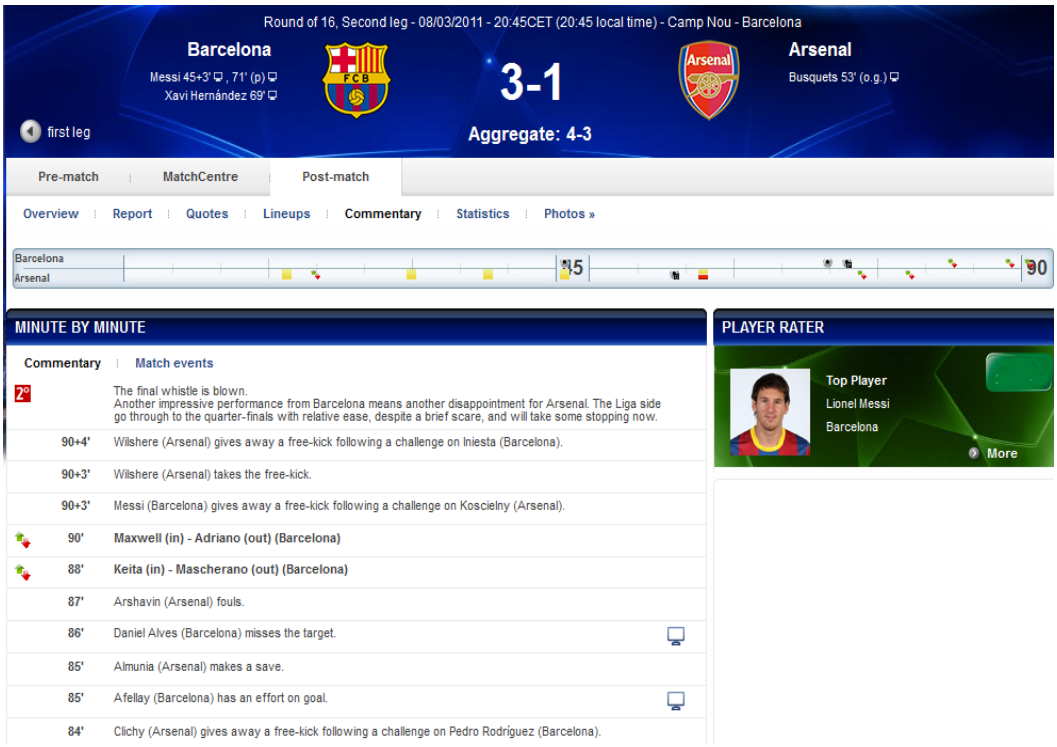

Figure 8 – An example match report from uefa.com.

Tunaoglu *et al.* [21] propose a method that is used for extracting the textual information in both languages. In [21], a domain specific information extraction approach is presented. Manually formed templates are used to extract information from unstructured text. This method has three steps. First, available web-casting texts are fetched by a web crawler to an intermediate file. Then, the named entities are tagged such as teams and players in the narrations for each match. Finally, two level lexical analyzer extracts events by analyzing the narrations for each event separately. As a result the type of the event, actors in the event and the time of the event are extracted from the web-casting text. We use these extracted data as input to our method. This method is not implemented in this study. The information from web-casting text is used as it is extracted. A summary of information extraction from text can be seen in Figure 9.

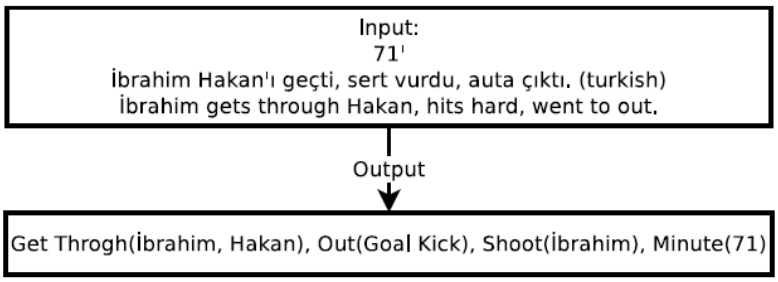

Figure 9 – Information extraction from text.

CHAPTER 4

EVENT BOUNDARY DETECTION

In this chapter, we explain how to fuse the extracted data to detect events and event boundaries in soccer videos. All features extracted from different modalities are used to find events on the video. By this way the match is summarized and it is also annotated semantically. Now that we have all the features, the next step is to combine these features. After feature extraction, the data at hand includes:

- Shots (start time , end time, length).
- Shot classes (Far view (F), Medium view (M), Close-up view(C)).
- Average Sound Amplitude (for each shot).
- Extracted data from match reports on the web (sporx.com and uefa.com):
    I. Type of the event
    II. Time of the event in minutes.

In multimodal systems like this, it is important to align the features extracted from video, audio and text. A sort of synchronization should be done to combine these features. Now we explain how the features are

aligned and fused. Then two approaches, namely, rule based and classification based, are presented to detect event boundaries.

4.1 Fusion of features from Text, Audio and Video

In soccer broadcasts, the video has a general structure. The match is recorded from the main camera (far view). The camera view is fixed until an event occurs. If an event occurs, camera view changes and it is switched to medium view, close-up view, or replay mode. Replays are medium view shots or short length far view shots. When the event ends, camera switches to the main camera view (far view) again. This structure enables us to look for events between two far view shots that are long enough. We are able to detect the events and their boundaries according to the shot structure between two far view shots. Events do not have the same shot structure, but they have similar camera switches as mentioned above. An example of a shot sequence for a goal event is given in Figure 10.

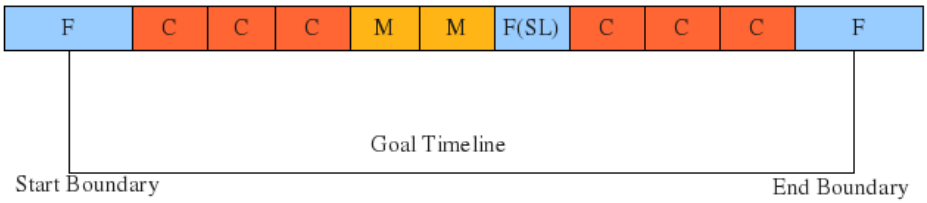

Figure 10 - An example shot sequence of a goal event. F: Far view, F (SL): Short Length Far view, C: Close-up view, M: Medium view

As we remarked before, if we know the exact time of the event, it is easier to find the event and its boundaries. However, in our case, we do not have

the exact moment of the event and web-casting text gives us inaccurate time. Match reports mark events minute by minute, and on the average 2 or 3 events occur in a minute. For example, the time points 14:11, 14:30, and 14:45 are all labeled as the 15th minute, or sometimes they may even be labeled as the 16th minute if the video is not well synchronized. An event can start in 15th minute and can end in 16th minute as well. As a result we cannot find the event by going to the 15th minute of the video. There are 2 or 3 candidate events around this minute. We have to determine the correct event among the possible events. We know that shot sequences between two far view shots are important to define the boundaries of an event. For this purpose, the following rules have been defined:

- Between two far view shots, only a single event can occur.

- Given the approximate time of the event, the search range for the event boundary starts three far view shots before the event time and ends three far view shots after the event time.

These rules give us five possible time intervals which an event could belong to. The next step is to choose one of these time intervals correctly. The shot sequence of the search range looks like the following:

$$F_1 \ldots\ldots Int_1 \ldots\ldots F_2 \ldots\ldots Int_2 \ldots\ldots F_3 \ldots\ldots Int_r \ldots\ldots F_4 \ldots\ldots Int_4 \ldots\ldots F_5 \ldots\ldots Int_5 \ldots\ldots F_6$$

Here F_i stands for far view shots. Int_i is the shot sequence (Interval) between far view shots F_i and F_{i+1}. Int_r indicates the shot sequence (Interval) including the 'reference shot'. The reference shot is the shot which includes

the related event time. The search range can be narrowed down or widened up according to the reliability of the web-casting text. Currently we use five 'far-view to far-view' intervals (shot sequences). The one that looks like the event we are looking for is chosen to determine the event boundaries. Event boundaries start somewhere in the far view shot and finish in the next far view shot. Shot sequence between these two far view shots is related with this event.

To sum up, according to the rules above we look for the event between two far view shots. Because of the inaccuracy of the time of data extracted from the web-casting text, we define a search range composed of five "far view to far view" shot sequences. The event is in one of these five shot sequences. Now the problem is how to select the correct one of these intervals. Two approaches are presented in this thesis. The first one is a rule based approach and the other one is a classification based approach.

4.2 Rule Based Method

We have found five candidate shot sequences as a result of the synchronization of text and video using the defined rules. To select the target event from one of these candidates, we developed a rule-based method. This method is presented in our previous work [22] with few experimental data. The experiments are extended in this thesis.

In this approach, after a search range is constructed with five intervals, each interval is voted by the rules defined for the event type (there are different rules for different event types). The interval that gets the highest vote is

chosen and the event boundary is determined using this interval (shot sequence).

Let us give an example to make the process clear. First, the event type and event minute is extracted from the web-casting text. Let us assume that the event is 'a goal event' and the time is 'the 30th minute'. Then we find the shot (reference shot) which corresponds to the "30:00" of the video. After the reference shot is found, five 'far view to far view' intervals are determined by going backwards and forwards through the neighboring shot sequence. Each interval is voted by the rules defined specifically for the goal event and the interval with the highest vote is determined to be the interval containing the goal event. Finally, the winning interval is used to extract the boundaries of the goal event.

Boundary extraction from the winning interval is a simple procedure. An event starts in the first far view shot, continues during the interval until the second far view shot. So we assume that the event boundary starts in 20 seconds before the end of the first far view shot and ends with the beginning of the second far view shot of the interval.

A voting mechanism is defined for each event type. The considered features include the number of shots in the interval, the length of the interval, the number of close-up shots in the interval, and the total length of medium view shots. We also reward the interval including the reference shot. For exciting events (goals and missed goals), the interval with the highest sound amplitude is also rewarded. We can summarize the voting mechanism item by item as follows:

- A rule set for each event is constructed by observation. Each rule has a weight.
- For each candidate interval, satisfying rules' weights are added.
- The interval getting the highest point is selected.

The constructed rules differ according to the event type. The rules are determined by observation. Because of the broadcast standards and event specifications, each event has a particular shot count, class and duration on the average. For example, the rule for a goal event says:

1) The interval having the goal event must contain more than 6 shots.
2) Total close-up view shot duration must be longer than 15 seconds.
3) The interval must include minimum 2 medium views or short length far view shots.
4) The interval cannot be shorter than 25 seconds
5) It must have a higher sound amplitude than the threshold.

All these items have different weights between 1 and 2 with respect to the importance. The interval getting the highest vote by these items is chosen as the goal event. Another example rule is for the corner event:

1) The interval having the corner event must contain shot count between 2 and 8.
2) The duration of the interval need to be between 8 and 25 sec.
3) Close-up views must last shorter than 15 sec.
4) The corner event must include at least one medium view shot and total medium view shot duration need to be at most 12 sec.

The rules for the other events are created in a similar way and some events use the same rules. A summary of the rule based event boundary detection is shown in Figure 11.

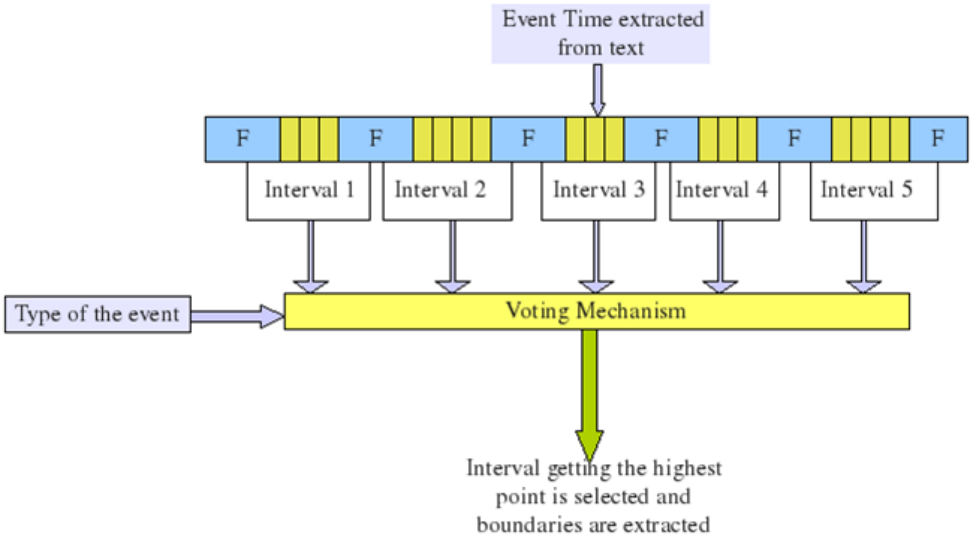

Figure 11 – Rule Based Event Boundary Detection.

4.3 Classification Based Method

Defining rules for each event type is a hard process and it is restricted to broadcasters and may be to the domain. A flexible and more generalized method is needed for event boundary detection. We took our study one step forward and replaced rule based method with a classification approach to select one of the candidate intervals. In this new approach feature extraction process and the construction of the search range (five candidate intervals) are the same as before. After this point the selection of the correct interval from five candidate intervals is done with a classification approach.

Then extracting the boundaries of the selected shot sequence is the same as it is in the rule based method. Figure 12 shows the event boundary detection process with the classification based approach.

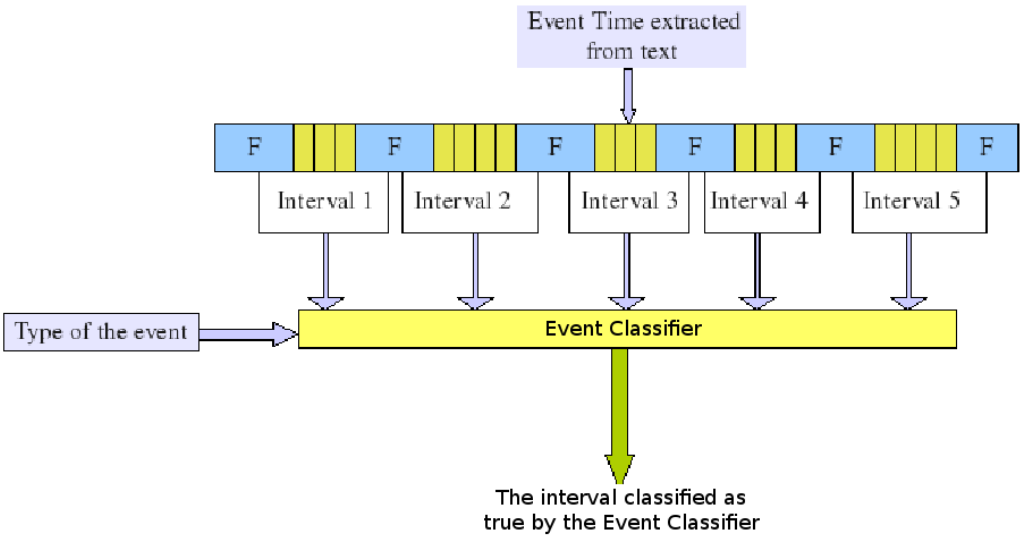

Figure 12 – Classification Based Event Boundary Detection

Event classifier is formed by constructing a model for each event type. Because we know the type of the event, event classifier uses the model of the known event type. For instance, if it is a goal event, event classifier chooses the goal event from five candidate intervals.

One Class SVM (Support Vector Machine) is used for classification. First, One Class SVM is explained.

4.3.1 What is One Class SVM?

Support Vector Machines are a group of supervised learning methods that are used in machine learning and pattern recognition for classification and regression. A support vector machine performs binary classification. Support Vector Machines non-linearly map their n-dimensional input space into a high dimensional feature space. In this high dimensional feature space a linear classifier is constructed. Multiclass SVM and one class SVM are the extensions to the original SVM. SVM takes a set of inputs (vector) and predicts the class of the input. SVM builds a model using the given training examples. Each of the training examples belongs to one of the categories (classes). SVM can model complex, real-world problems such as text and image classification, hand-writing recognition, and bioinformatics and biosequence analysis. SVM performs well on data sets that have many attributes, even if there is few data to train the model.

One class SVM first introduced by Schölkopf [23, 24]. One-Class SVM distinguishes one class of data from the rest of the feature space given only a positive data set. Based on a strong mathematical foundation, One Class SVM draws a nonlinear boundary of the positive data set in the feature space using two parameters; one to control the noise in the training data and the other to control the "smoothness" of the boundary. One Class SVMs have the same advantages as SVM, such as efficient handling of high dimensional spaces and systematic nonlinear classification using advanced kernel functions. In this study one class SVM is implemented using LIBSVM library [25].

4.3.2 Why use One Class SVM?

In our event boundary detection mechanism, we know the type of the event that is extracted from text. Since we know the type of the events that we are trying to detect, it is enough to check if a candidate interval belongs to that event class or not. A model for each event type is constructed. There is a model for a "Goal" event, a model for a "Corner" event etc. Then candidate sequences are compared with the trained model and the one that is classified as the event is selected. As a result, it is a one-class classification. The candidate interval belongs to that class or not.

We also do not have negative samples. There is only the event sequence which is labeled. Negative samples are the complement of the labeled data. It is also not correct to define that negative sample of a goal event is a corner event. It is possible that two events can have similar features.

Another issue is the lack of the training data. We have not got much data. Training data is extracted from 15 soccer games. We know that SVM performs well with small amount of data. Because of all these reasons, it is decided that one class SVM is appropriate for this study.

4.3.3 Training and Classification

Training of each event type is made on its own. Training data is collected for each event type with the feature extraction which is explained in the previous chapters.

Feature vector is constructed using the training data extracted in feature extraction. While feature vector is being constructed, all the features are normalized. Features are subject to change by experimenting but we can itemize the main ones as:

1- Total number of shots in a sequence (interval).

2- Total duration of a sequence.

3- Number of close-up shots in a sequence.

4- Total close-up duration in a sequence.

5- Number of medium shots in a sequence.

6- Total medium duration of a sequence.

7- The average of peak sound amplitude for a sequence.

Feature vectors are constructed for training data as explained above for the true sequence of the event. The extracted feature vector is used to construct the event's model. The model is used in the detection of this event. For instance, suppose a goal event in the 16^{th} minute will be detected. The reference shot in the 16:00 of the video is found. Five candidate sequences of shots are determined and for each sequence, feature extraction is done. Then feature vector of each one is given to the event classifier of "goal" event. Event classifier returns if the candidate interval is a goal event or not. If it is a goal event, the interval is labeled as a goal. There is a priority between five candidate shot sequences. The interval having the reference shot is checked first. If it is the desired event, the rest is not processed. The priority of the intervals is:

Interval 3 -> Interval 2 -> Interval 1 -> Interval 4 -> Interval 5

This is because match reports transfers the events after the occurrence of the events. The event probably occurs before it is reported. Because of this, the priority is through backward from the reference shot and then through the following sequences.

CHAPTER 5

EXPERIMENTS AND EVALUATION

A sufficient number of soccer match videos is needed to extract enough data for training and testing of each event type. At the same time, it is hard to access media sources from soccer broadcasts. Labeling the true data is also a difficult process. In this thesis, shots, shot classes and events are all labeled manually to construct the test and the training data. Experiments are conducted on 15 soccer games. These are Turkish Super League matches and UEFA Champions League matches. Web-casting texts for these games are obtained from "uefa.com", and "sporx.com" which is the most popular website for live match reports in Turkey. Experiments are held in three parts. These are shot detection, shot classification and event boundary detection. The results of the experiments on each step are shown in the following sections.

5.1 Shot Detection

In the evaluation, 439 shot boundaries in 5 matches were labeled manually. Automatically detected shot boundaries are compared empirically with the ones labeled manually. Table 1 shows the results of automatic shot boundary detection. 420 out of 439 shot boundaries were detected correctly. The 'Missed' column represents the missed boundaries which were

supposed to be detected. 18 boundaries were missed because the video had gradual transitions between shots whose color distributions were almost the same. Wrong detections are shown under the column 'False'. They were caused by the rapid motion in close-up views, which leads a remarkable change in histogram.

Table 1 – Results of Automatic Shot Boundary Detection.

Match Id	# of Shots	Correct	Missed	False	Precision (%)	Recall (%)
1	98	96	2	9	98	91
2	144	133	10	7	92	95
3	73	72	1	7	99	91
4	61	58	3	3	95	95
5	63	61	2	5	97	92
TOTAL	439	420	18	31	96	93

Since we do not have the same data set, it would be misleading to compare results with the related works. However, [14] compared different shot detection algorithms. They used histogram differences method and they obtained results with different thresholds. They reached 90% detection rate on the average. With 96% precision and 93% recall rates; we have satisfactory results compared to them. C. Xu [6], which has good results on event boundary detection, used a commercial tool for shot detection. Unfortunately we were not able to compare our results with theirs.

5.2 Shot Classification

Since it is a time consuming task to label all shots in all test videos, we chose part of some games for evaluation of shot classification. Parts of 5 UEFA Champions League games were analyzed automatically by our shot detection tool and they were classified using the technique proposed in this thesis. Shots are also classified manually to use in evaluation. The shot classification accuracy is evaluated by comparing the automatically classified shots with the ones classified manually. The results are given in Table 2.

Table 2 - Shot Classification Accuracy for Three Types of Shots.

Match Id	Shot Class	# of Shots	Correct	Incorrect	Accuracy (%)
1	Far view	41	39	2	95
	Medium view	26	20	6	77
	Close-up view	31	25	6	81
2	Far view	46	39	7	85
	Medium view	58	44	14	76
	Close-up view	23	18	5	78
3	Far view	27	24	3	89
	Medium view	20	15	5	75
	Close-up view	20	17	3	85
4	Far view	22	22	0	100
	Medium view	17	8	9	47
	Close-up view	18	17	1	94
5	Far view	25	22	3	88
	Medium view	15	8	7	53
	Close-up view	18	18	0	100
TOTAL	Far view	161	146	15	91
	Medium view	136	95	41	70
	Close-up view	110	95	15	86

Far view shots have a high classification rate of 91%. True classification of far view shots is very important for us to detect event boundaries correctly. Medium and close-up view shots have lower classification rates; they are more prone to be confused since it is hard to separate medium and close-up view shots, even by observation. Close-up view shots have a classification rate of 86% while medium view shots have 70%. As we described, it is difficult to discriminate medium view shots from the other views.

The results are not as good as the work of Xu [6] but the results are satisfactory to use in event boundary detection. They have a classification rate of 98% in far view shots and 92% in close-up view shots. Our results are comparable with this classification rates. Medium view shots are classified with a rate of 89% in their study. We have a worse classification rate of medium view shots but this is not critical. In the worst case, they are classified as close-up view. Additionally, our evaluation data is different from the compared study. As a result, we can say that the evaluation results are satisfactory and we can use our classification approach in feature extraction safely.

5.3 Event Boundary Detection

15 soccer games are used for evaluation of event boundary detection. 5 of them are Turkish Super League's matches and the rest is UEFA Champions League's matches. 5 types of events are extracted from these games. These events are goal, corner, missed goal, red/yellow cards and penalty. Web-

casting text sources are “sporx.com” for Turkish Super League and “uefa.com” for UEFA Champions League. The events of 15 soccer games are labeled using text sources and the boundaries of events are set manually. The evaluation data includes 55 goal events, 113 corner events, 81 missed goal events, 52 red/yellow cards events and 3 penalty events. All these events are labeled and have boundaries. Our proposed event boundary detection method is expected to determine these boundaries correctly. Small shifts in the boundaries are ignored. The accuracy of the proposed event boundary detection method is calculated with the formula below:

$$Precision = Number\ of\ correctly\ detected\ events / Total\ number\ of\ events \quad (5.1)$$

The evaluation of rule based method and classification based method is separated and presented in the following sections.

5.3.1 Rule Based Event Boundary Detection

Rule based event boundary detection algorithm is applied on the evaluation data. As it is mentioned before, different rules are used for different events. The rules were constructed for Turkish Super League games and they are revised to be used with UEFA Champions League games too. The audio features are utilized in exciting events such as goals and missed goals. The effect of audio features on exciting events is presented in our previous work [22]. Audio features have a great positive

effect on missed goals which do not have a definite structure. Table 3 shows the experiment results conducted in our previous work with 25 goal and 18 missed goal events. The results show the detection rates with and without the sound amplitude. It is obvious that audio feature has a great effect.

Table 3 – Event Detection Rates without and with Sound Amplitude.

Event Type	Accuracy without Sound (%)	Accuracy with Sound (%)
Goal	80	88
Missed Goal	39	67

The experimental results of rule based event boundary detection are explained for each event respectively. The first event is a goal event. 55 goal events are used for evaluation. 45 of 55 event boundaries are found correctly. We obtained a detection rate of 82% in goal events. The detailed results can be seen in Table 4. Wrong detections are produced because of several reasons. One reason is the case where an event does not show the specific characteristics of that type of event. Another reason is that there may be other intervals having similar characteristics of the event in the search range. And in some other cases, there may be two events of the same type in the search range which means that it is possible to select one of them instead of the other.

Table 4 – Rule Based Event Boundary Detection Results of Goal Events.

Match Id	# of Goal Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	5	5	0	100
2	2	2	0	100
3	2	2	0	100
4	2	2	0	100
5	3	3	0	100
6	6	5	1	83
7	9	7	2	78
8	1	1	0	100
9	7	3	4	43
10	3	3	0	100
11	3	2	1	67
12	4	4	0	100
13	4	3	1	75
14	4	3	1	75
TOTAL	55	45	10	82

The results of corner event detection are shown in Table 5. The overall accuracy of corner event boundary detection is 56%. 63 corner events out of 113 are detected correctly.

Table 5 - Rule Based Event Boundary Detection Results of Corner Events.

Match Id	# of Corner Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	13	7	6	54
2	9	8	1	89
3	8	4	4	50
4	12	6	6	50
5	6	4	2	67
6	4	2	2	50
7	2	1	1	50
8	12	5	7	42
9	3	2	1	67
10	11	7	4	64
11	6	3	3	50
12	10	4	6	40
13	7	3	4	43
14	10	7	3	70
TOTAL	113	63	50	56

Table 6 presents the results of missed goals. 81 missed goal events are used in evaluation. 50 of them are marked correctly with an accuracy of 62%.

Table 6 - Rule Based Event Boundary Detection Results of Missed Goal Events.

Match Id	# of Missed Goal Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	4	3	1	75
2	6	4	2	67
3	6	4	2	67
4	4	4	0	100
5	5	3	2	60
6	5	3	2	60
7	5	3	2	60
8	6	3	3	50
9	5	4	1	80
10	6	3	3	50
11	5	4	1	80
12	9	5	4	56
13	7	4	3	57
14	8	3	5	38
TOTAL	81	50	31	62

The results of red/yellow card events are shown in Table 7. The precision of red/yellow card event boundary detection is 52%. 27 of 52 experiment data are labeled correctly and boundaries are extracted successfully. Table 8 presents the results of penalty events. Only 3 penalty events in 15 games are used and they are all marked correctly. The accuracy of penalty event boundary detection is 100%.

Table 7 - Rule Based Event Boundary Detection Results of Red/Yellow Card Events.

Match Id	# of Red/Yellow Card Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	4	2	2	50
2	4	2	2	50
3	5	2	3	40
4	2	2	0	100
5	1	0	1	0
6	4	2	2	50
7	2	1	1	50
8	3	0	3	0
9	3	2	1	67
10	3	2	1	67
11	6	3	3	50
12	3	0	3	0
13	3	3	0	100
14	5	2	3	40
15	4	4	0	100
TOTAL	52	27	25	52

Table 8 - Rule Based Event Boundary Detection Results of Penalty Events.

Match Id	# of Penalty Events	# of Correct Detection	# of Outliers	Accuracy (%)
TOTAL	3	3	0	100

The results are not directly comparable with any of the related work. The study of Xu [6] is the closest one. They have goal, corner and red/yellow card events in common. Their event boundary accuracy is 98% for goal, 86% for corner and above 90% for card events. They have good accuracies. The results of our rule based event boundary detection method are not as

good except goal events. Goal events reached an accuracy of 82% which is acceptable.

However, it is not correct to compare our results with the study of Xu [6]. First of all, the same data set is not used. Secondly, the most important difference is the reliability of the web-casting text. Xu *et al.* used the exact moment of the event and constructed the boundaries of that event around that moment so that it is impossible to have the event outside of these boundaries. The critical work is to find boundaries correctly. However, we do not know the exact moment of the event since we have web-casting text in minute precision. We choose the correct event from 5 possible intervals and then find the boundaries. As a result, the scopes of the studies are different. This new approach brings a new perspective to use unreliable text in multimodal fusion. The results are encouraging. It is predicted that adding new features specific to events will increase the accuracy and robustness of this new approach.

5.3.2 Classification Based Event Boundary Detection

The test data used in rule based method is also used to evaluate the classification based event boundary detection. Because of the lack of data, each match data is excluded from the rest to be used as test data and the rest is used to make a model. They are trained and a model is constructed. Data of each event is used separately to make a model for each event type. With this method, all matches are used as test data in turn.

The same procedure is applied in validation of training. Cross Validation is performed. A sort of Leave One Out Cross Validation (LOOVC) is done by using features of 1 soccer game as validation data and using the remaining 14 games as training data. Cross validation results for each event type is given in Table 9.

Table 9 – Cross Validation Accuracies.

Event Type	Cross Validation Accuracy (%)
Goal	80
Corner	78
Missed Goal	78
Yellow/Red Card	79

Cross validation accuracy of the goal events is 80% while corner events and missed goal events are both having 78% accuracy. Similarly, Yellow/red card events have a cross validation accuracy of 79%.

Table 10 displays the results of the classification based event boundary detection method on goal events. The accuracy of the method on goal events is 80%. This result is a bit worse than the rule based method results but we got nearly the same accuracy value.

Table 10 - Classification Based Event Boundary Detection Results of Goal Events.

Match Id	# of Goal Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	5	4	1	80
2	2	1	1	50
3	2	2	0	100
4	2	2	0	100
5	3	3	0	100
6	6	5	1	83
7	9	5	4	56
8	1	0	1	0
9	7	6	1	86
10	3	3	0	100
11	3	3	0	100
12	4	3	1	75
13	4	3	1	75
14	4	4	0	100
TOTAL	55	44	11	80

Table 11 - Classification Based Event Boundary Detection Results of Corner Events.

Match Id	# of Corner Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	13	5	8	38
2	9	8	1	89
3	8	5	3	63
4	12	12	0	100
5	6	5	1	83
6	4	3	1	75
7	2	0	2	0
8	12	6	6	50
9	3	2	1	67
10	11	8	3	73
11	6	5	1	83
12	10	6	4	60
13	7	6	1	86
14	10	4	6	40
TOTAL	113	75	38	66

Table 11 shows the results of corner events. The detection of corner events reaches an accuracy of 66%. This is a good result that beats the rule based method.

The results of missed goals are presented in Table 12. Missed goals are detected with an accuracy of 62%.

Table 12 - Classification Based Event Boundary Detection Results of Missed Goal Events.

Match Id	# of Missed Goal Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	4	2	2	50
2	6	4	2	67
3	6	4	2	67
4	4	2	2	50
5	5	4	1	80
6	5	2	3	40
7	5	3	2	60
8	6	2	4	33
9	5	4	1	80
10	6	1	5	17
11	5	3	2	60
12	9	7	2	78
13	7	7	0	100
14	8	5	3	63
TOTAL	81	50	31	62

Table 13 demonstrates the results of red/yellow card events. Red/yellow card events are detected with a rate of 62%. Penalty events are not included

in classification based method because we lack the training data for penalty events. The structure of the penalty events is very close to the goal events so that it is predicted to have high detection rates in penalty events too.

Despite the fact that the results are not bad, improvement on the method is needed. It is obvious that with some enhancement, the classification based method will perform better than the rule based one. Goal events performed a bit worse because the significance of audio amplitude decreased in the feature vector whereas it was very important for goal events in the rule-based method. It is good to get an improvement on the other events and it is a great motivation for the future work.

Table 13 - Classification Based Event Boundary Detection Results of Red/Yellow Card Events.

Match Id	# of Red/Yellow Card Events	# of Correct Detection	# of Outliers	Accuracy (%)
1	4	2	2	50
2	4	4	0	100
3	5	2	3	40
4	2	1	1	50
5	1	1	0	100
6	4	2	2	50
7	2	0	2	0
8	3	1	2	33
9	3	3	0	100
10	3	3	0	100
11	6	4	2	67
12	3	2	1	67
13	3	1	2	33
14	5	3	2	60
15	4	3	1	75
TOTAL	52	32	20	62

The results of the classification based event boundary detection are not worse than the rule based method. Moreover classification based method performs better in some events. This is important because we aimed to replace rule based method. The replacement is needed because it is hard to define new rules when new features are added. This will increase the flexibility of the overall event boundary detection and this will simplify the addition of new features. We can say that the classification based method performed well and made it possible to increase the detection rate by adding new features easily.

We also conducted an experiment to show the effect of event boundary detection using an unreliable web-casting text. The correction rate of our method on unreliable web-casting text is calculated. First, event boundaries are extracted from the interval having the reference shot and then the results are compared with the results of the proposed classification based method. Table 14 shows the results and the difference value that shows the correction rate.

Table 14 – Detection Rate using reference shot and the comparison with the proposed method.

Event Type	Detection Accuracy Using Only The Reference Shot (%)	Detection Accuracy of The Proposed Method (%)	The Difference Value (%)
Goal	61	80	19
Corner	43	66	23
Missed Goal	38	62	24
Red/Yellow Card	46	62	16

The results show that the proposed method corrects the uncertainty of web-casting text by 20% in average. We see that the method presented in this thesis finds the event boundaries in soccer games by correcting the information extracted from unreliable and uncertain web-casting texts.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, we present a new multi-modal method for event and event boundary detection by aligning web-casting texts with videos. The event boundary detection technique is based on web-casting texts which are not precise. Web-casting texts describe the events minute-by-minute but these minutes are not the exact time points of the event. As a result, it is necessary to deal with the problem of synchronization in seconds.

We aimed to solve this problem and increase the robustness of asynchronous web-casting texts. Similar approaches use extracted data from web-casting texts with the exact event time points and determine boundaries accordingly. When we consider inexact data, it is necessary to look for the events in a wider time period which causes poor event detection rates. Our experiment results are not excellent but they are encouraging for this newly proposed framework.

The experimental results show that our method has satisfying rates for the event and event boundary detection. It enables us to detect the events and event boundaries that are asynchronous with the web-casting text data. Our experiments also demonstrate that audio-visual features become more important when the text sources are inaccurate. We show that the audio features have remarkable effects on event and event boundary detection in

important and exciting events. We proposed methods for shot detection and shot classification too. Shot detection and shot classification methods are specialized for soccer videos. Experiments on shot detection and shot classification show that methods perform well to be used in feature extraction.

Although we have obtained satisfactory results, detailed audio-visual features are necessary to increase the detection rate. As a future work, we plan to add event specific audiovisual features such as time detection (time digits from the embedded text on the image), referee detection, ball detection, goal and goal keeper detection and yellow/red card detection. These new features are planned to be used with the basic features like shot durations and classes. For instance, new features can be constructed like the proportion of the frames having the ball in a shot, proportion of the frames having the referee in a shot, the number of medium shots having goal in an interval etc. The features can be extended when new detailed audio-visual analyses are done. We expect to reach higher detection rates by adding new features on the base method we presented here.

Another improvement considered is to make different models for different broadcasters. Each broadcaster has its own production standards. They all have similar production structures but they differ in some points. They may have different shot transition structures, different logos. Constructing models for each broadcaster will definitely improve the quality of the model and event boundary detection.

To conclude, our new approach for event boundary detection using audio visual features and web-casting texts brings a new perspective to use unreliable text in multimodal fusion. The results are encouraging. It is predicted that adding new features specific to events will increase the accuracy and robustness of this new approach.

REFERENCES

1. Pradeep K Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik and Mohan S. Kankanhalli. Multimodal Fusion for Multimedia Analysis: A Survey, *In: Multimedia Systems*, Vol. 16, Nr. 6 (2010), S. 345-379, 2010.

2. D. Zhong and S.-F. Chang. Real-time view recognition and event detection for sports video. *Journal of Visual Communication and Image Representation*, 15(3):330–347, 2004.

3. J. quan Ouyang, J. tao Li, and Y. dong Zhang. Replay scene based sports video abstraction. *Lecture Notes In Computer Science*, 3614:689–697, 2005.

4. A. Ekin and A. M. Tekalp. Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing*, 12:796–807, 2003.

5. Jrgen Assfalg, Marco Bertini, Carlo Colombo, Alberto Del Bimbo, andWalter Nunziati, Semantic annotation of soccer videos: Automatic highlights identification, *Computer Vision and Image Understanding*, vol. 92, pp. 285–305, 2003.

6. C. Xu, J. Wang, H. Lu, and Y. Zhang. A novel framework for semantic annotation and personalized retrieval of sports video. *IEEE Transactions on Multimedia*, 10:421–436, 2008.

7. C. Xu, J. Wang, K. Wan, Y. Li, and L. Duan. Live sports event detection based on broadcast video and web-casting text. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 221–230, 2006.

8. Shu-Ching Chen, Min Chen, Chengcui Zhang and Mei-Ling Shyu, Exciting Event Detection Using Multi-level Multimodal Descriptors and Data Classification, *International Symposium on Multimedia ISM'06*,pp. 193 – 200, 2006.

9. Radhakrishnan, R, A Divakaran, and T S Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. *International Conference on Multimedia and Expo ICME 03*, 2003.

10. Yong Rui, Anoop Gupta, and Alex Acero. Automatically extracting highlights for TV Baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia* (MULTIMEDIA '00), 2000.

11. C. Zhang, S.-C. Chen, and M.-L. Shyu. Pixso: a system for video shot detection. *Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, 3:1320–1324 vol.3, Dec. 2003.

12. M. Luo, D. DeMenthon, and D. Doermann. Shot boundary detection using pixel-to-neighbour image di erences in video. *TRECVID 2004 Workshop Notebook Papers*, 2004.

13. R. Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases*, number SPIE 3656, pages 290–301, January 1999.

14. J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. In *Storage and Retrieval for Still Image and Video Databases IV*, number SPIE 2664, Los Angeles, California, January 1996.

15. J. Wang, E. Chng, and C. Xu. Soccer replay detection using scene transition structure analysis. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2:ii/433–ii/436 Vol. 2, March 2005.

16. Y. Hu, B. Han, G. Wang, and X. Lin. Enhanced shot change detection using motion features for soccer video analysis. *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1555–1558, July 2007.

17. A. Jacobs, A. Miene, G. T. Ioannidis, and O. Herzog. Automatic shot boundary detection combining color, edge, and motion features of adjacent frames. *TRECVID 2004 Workshop Notebook Papers*, pages 197–207, 2004.

18. X. Tong, Q. Liu, and H. Lu. Shot classification in broadcast soccer video. *ELCVIA*, 1:16–25, 2008.

19. Y.-H. Zhou, Y.-D. Cao, L.-F. Zhang, and H.-X. Zhang. An svm-based soccer video shot classification. *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, 9:5398–5403 Vol. 9, Aug. 2005.

20. J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, pages 679-714, 1986.

21. D. Tunaoglu, O. Alan, O. Sabuncu, S. Akpinar, N. Cicekli, and F. Alpaslan. "Event extraction from turkish football web-casting texts using hand-crafted templates," Proc. Of Third IEEE International Conference on Semantic Computing (ICSC '09), September 2009.

22. Mujdat Bayar, Özgür Alan, Samet Akpinar, Orkunt Sabuncu, Nihan K. Çiçekli, Ferda Nur Alpaslan. Event boundary detection using audio-visual features and web-casting texts with imprecise time information. *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME 2010)*, 578-583, 2010.

23. B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 1207-1245, 2000.

24. B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 1443-1471, 2001.

25. http://www.csie.ntu.edu.tw/~cjlin/libsvm/, retrieved by September 1th, 2011.

26. J. Wang, M.S. Kankanhalli, W.Q Yan and R. Jain. Experiential sampling for video surveillance. *ACM Workshop on Video Surveillance,* 2003.

27. M.T. Yang, S.C. Wang and Y.Y. Lin. A multimodal fusion system for people detection and tracking. *International Journal of Imaging Systems and Technology,* 131–142, 2005.

28. Y. Wang, Z. Liu and J.C. Huang. Multimedia content analysis: using both audio and visual clues. *IEEE Signal Processing Magazine,* pp. 12–36, 2000.

29. Y. Ding and G. Fan. Segmental hidden markov models for viewbased sport video analysis. *International Workshop on Semantic Learning Applications in Multimedia,* 2007.

30. Q. Zhou and J. Aggarwal. Object tracking in an outdoor environment using fusion of features and cameras. *Image and Vision Computing,* 1244-1255, 2006.