

EVENT DETECTION IN SOCCER VIDEO BASED ON
AUDIO/VISUAL KEYWORDS

KANG YU-LIN
(B. Eng. Tsinghua University)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE
SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE
2004

Acknowledgements

First and foremost, I must thank my supervisors, Mr. Lim Joo-Hwee and Dr. Mohan S Kankanhalli, for their patient guidance and supervision during my years at Nation University of Singapore (NUS) attached to Institute for Infocomm Research (I2R). Without their encouragement and help in many aspects of my life in NUS, I would never finish this thesis.

I also want to express my appreciations to School of Computing and I2R for offering me the study opportunities and scholarship here.

I am grateful to the people in our cluster at I2R. Thanks Dr Xu Chang-Sheng, Mr. Wan Kong Wah, Ms. Xu Min, Mr. Namunu Chinthaka Maddage, Mr. Shao Xi, Mr. Wang Yang, Ms. Chen Jia-Yi and all my friends at I2R for giving me many useful advices.

Thanks my lovely wife – Xu Juan for her support and understanding. You make my life here more colorful and more interesting.

Finally, my appreciation goes to my parents and my brother, for their love and support. They keep encouraging me and give me the power to carry on my research.

Table of Contents

Acknowledgements	i
Table of Contents	ii
List of Figures.....	iv
List of Tables	v
Summary	vi
Conference Presentation	viii
 Chapter 1	
Introduction.....	1
1.1 Motivation and Challenge.....	1
1.2 System Overview	4
1.3 Organization of Thesis.....	7
 Chapter 2	
Literature Survey	8
2.1 Feature Extraction	8
2.1.1 Visual Features	9
2.1.2 Audio Features.....	9
2.1.3 Text Caption Features.....	10
2.1.4 Domain-Specific Features	10
2.2 Detection Model.....	11
2.2.1 Rule-Based Model.....	11
2.2.2 Statistical Model	12
2.2.3 Multi-Modal Based Model.....	13
2.3 Discussion	14
 Chapter 3	
AVK: A Mid-Level Abstraction for Event Detection	17
3.1 Visual Keywords for Soccer Video	18
3.2 Audio Keywords for Soccer Video.....	25
3.3 Video Segmentation.....	25
 Chapter 4	
Visual Keyword Labeling.....	29
4.1 Pre-Processing.....	31
4.1.1 Edge Points Extraction	31
4.1.2 Dominant Color Points Extraction	33
4.2 Feature Extraction	34
4.2.1 Color Feature Extraction.....	34

4.2.2 <i>Motion Feature Extraction</i>	39
4.3 Visual Keyword Classification	40
4.3.1 <i>Static Visual Keyword Labeling</i>	40
4.3.2 <i>Dynamic Visual Keyword Labeling</i>	42
4.4 Experimental Results	43
Chapter 5	
Audio Keyword Labeling	47
5.1 Feature Extraction	48
5.2 Audio Keyword Classification.....	50
Chapter 6	
Event Detection	52
6.1 Grammar-Based Event Detector	52
6.1.1 <i>Visual Keyword Definition</i>	53
6.1.2 <i>Event Detection Rules</i>	54
6.1.3 <i>Event Parser</i>	55
6.1.4 <i>Event Detection Grammar</i>	56
6.1.5 <i>Experimental Results</i>	59
6.2 HMM-based Event Detector	60
6.2.1 <i>Exciting Break Portion Extraction</i>	62
6.2.2 <i>Feature Vector</i>	63
6.2.3 <i>Goal and Non-Goal HMM</i>	64
6.2.4 <i>Experimental Results</i>	65
6.3 Discussion	68
6.3.1 <i>Effectiveness</i>	68
6.3.2 <i>Robustness</i>	68
6.3.3 <i>Automation</i>	69
Chapter 7	
Conclusion and Future Work	70
7.1 Contribution	70
7.2 Future Work	71
References	73

List of Figures

Fig. 1-1 AVK sequence generation in first level	5
Fig. 1-2 Two approaches for event detection in second level.....	6
Fig. 3-1 Far view (left) mid range view (middle) close-up view (right).....	19
Fig. 3-2 Far view of whole field (left) and far view of half field (right)	21
Fig. 3-3 Two examples for mid range view (whole body is visible)	21
Fig. 3-4 Edge of the field	22
Fig. 3-5 Out of the field	22
Fig. 3-6 Inside the field	23
Fig. 3-7 Examples for dynamic visual keywords.....	24
still (left) moving(middle) fast moving(right)	24
Fig. 3-8 Different semantic meaning within one same video shot	26
Fig. 3-9 Different semantic meaning within one same video shot	27
Fig. 3-10 Gradual transition effect between two consecutive shots	27
Fig. 4-1 Five steps of processing	30
Fig. 4-2 I-Frame (left) and its edge-based map (right)	33
Fig. 4-3 I-Frame (left) and its color-based map (right).....	34
Fig. 4-4 Template for ROI shape classification	38
Fig. 4-5 Nine regions for motion vectors.....	39
Fig. 4-7 Rules for dynamic visual keyword labeling.....	42
Fig. 4-8 Tool implemented for feature extraction.....	44
Fig. 4-9 Tool implemented for ground truth labeling.....	44
Fig. 4-10 “MW” segment which is labeled as “EF” wrongly.....	46
Fig. 5-1 Framework for audio keyword labeling	48
Fig. 6-1 Grammar tree for corner-kick	58
Fig. 6-2 Grammar tree for goal	59
Fig. 6-3 Special pattern that follows the goal event.....	61
Fig. 6-4 Break portions extractions.....	63
Fig. 6-5 Goal and non-goal HMMs.....	65
Fig. 7-1 Relation between syntactical approach and statistical approach.....	72

List of Tables

Table 1-1 Precision and recall reported by other publications	4
Table 3-1 Static visual keywords defined for soccer videos.....	19
Table 3-2 Dynamic visual keywords defined for soccer videos.....	24
Table 4-1 Rules to classify the ROI shape.....	38
Table 4-2 Experimental Results.....	45
Table 4-3 Precision and Recall	46
Table 6-1 Visual keywords used by grammar-based approach	53
Table 6-2 Grammar for corner-kick detection	57
Table 6-3 Grammar for goal detection.....	58
Table 6-4 Result for corner-kick detection	60
Table 6-5 Result for goal detection.....	60
Table 6-6 Result for goal detection ($T_{Ratio}=0.4$, $T_{Excitement}=9$)	66
Table 6-7 Result for goal detection ($T_{Ratio}=0.3$, $T_{Excitement}=7$)	67

Summary

Video indexing is one of the most active research topics in image processing and pattern recognition. Its purpose is to build indices for the video database by attaching text-formed annotation to the video document. For a specific domain such as sports videos, an increasing number of structure analysis and event detection algorithms are being developed in recent years. In this thesis, we propose a multi-modal two-level framework that uses Audio and Visual Keywords (AVKs) to analyze high-level structures and to detect useful events from sports video. Both audio and visual low-level features are used in our system to facilitate event detection.

Instead of modeling the high-level events directly on low-level features, our system first label the video segments with AVK which is a mid-level representation with semantic meaning to summarize the video segments in text form. Audio keywords are created from low-level features by using twice-iterated Fourier Transform. Visual keywords are created by detecting Region of Interest (ROI) inside playing field region, motion vectors and support vector machine learning.

In the second level of our system, we have studied and experimented with two approaches. One is statistical approach and the other is syntactical approach. For syntactical approach, a unique event detection grammar is applied to the visual keyword sequence to detect the goal and corner-kick from soccer videos. For statistical approach, we use HMMs to model different structured “break” portions of the soccer video and

detect the “break” portions with goal event anchored. We also analyze the strengths and weaknesses of these two approaches and discuss some potential improvements for our future research work.

A goal detection system has been developed based on our multi-model two-level framework for soccer video. Compared to recent research works in content-based sports video domain, our system produces advantages in two aspects. First, our system fuses the semantic meaning of AVKs by applying HMM in the second-level to the AVKs which are well aligned to the video segments. This makes our system very easy to extend to other sports video. Second, the usage of ROIs and SVM achieves good result for visual keywords labeling. Our experimental results show that the multi-modal two-level framework is a very effective method for achieve a better result for content-based sports video analysis.

Conference Presentation

[1] Yu-Lin Kang, Joo-Hwee Lim, Qi Tian and Mohan S. Kankanhalli. Soccer video event detection with visual keywords. *IEEE Pacific-Rim Conference on Multimedia, Dec 15-18 2003.*

(Oral Presentation)

[2] Yu-Lin Kang, Joo-Hwee Lim, Qi Tian, Mohan S. Kankanhalli and Chang-Sheng Xu. “Visual keywords labeling in soccer video”. To be presented at *IEEE International Conference on Pattern Recognition, Cambridge, United Kingdom, Aug22-26, 2004.*

[3] Yu-Lin Kang, Joo-Hwee Lim, Mohan S. Kankanhalli, Chang-Sheng Xu and Qi Tian. “Goal detection in soccer video using audio/video keywords”. To be presented at *IEEE International Conference on Image Processing, Singapore, Oct 24-27, 2004.*

Chapter 1

Introduction

1.1 Motivation and Challenge

The rapid development of technologies in computer and telecommunications industries have brought larger and larger amount of accessible multimedia information to the users. Users can access high-speed network connection via cable modem and DSL at home. Larger data storage devices and new multimedia compression standards make it possible for users to store much more audio and video data in their local hard-disk than before. Meanwhile, people quickly get lost in myriad of video data and it becomes more and more difficult to locate a relevant video segment linearly because of the time consuming task of annotation to the video data manually. All these problems call for the tools and technologies which could index, query, and browse the video data efficiently. Recently, many approaches have been proposed to address these problems. These approaches mainly focus on video indexing [1-5] and video skimming [6-8]. Video indexing aims at building indices for the video database so that user can browse the video efficiently. Research in video skimming area focuses on creating a summarized version of the video content by eliminating the un-important part. Research topics in these two areas include shot boundary

detection [9,10], shot classification [11], key frame extraction [12,13], scene classification [14,15], etc.

Besides the general areas like video indexing and video skimming, some researchers target their objectives to specific domains such as musical video [16,17], news video [18-22], sports video, etc. Especially for sports video, due to its well-formed structure, an increasing number of structure analysis and event detection algorithms are being developed in this domain recently.

We choose event detection in sports video as our research topic and use one of the most complicated structured sports videos – soccer video as our test data due to following two reasons:

1. Event detection systems are very useful.

The amount of accessible sports video data is growing very fast. It is quite time consuming to watch all these video. In particular, some people might not want to watch the whole sports video. Instead, they might just want to download or watch the exciting part of the sports video such as goal segments in soccer videos, touchdown segments in football videos, etc. Hence, a robust event detection system in sports video becomes very useful.

2. Although many approaches have been presented for event detection in sports video, there is still room for improvement from system modeling and experimental result point of views.

In the beginning, most of the event detection systems share two common features. First the modeling of high-level events such as play-break, corner kicks, goals etc are anchored directly on low-level features such as motion and colors leaving a large semantic gap between computable features and content meaning as understood by humans. Second some of these systems tend to engineer the analysis process with very specific domain knowledge to achieve more accurate object or/and event recognition. This kind of highly domain-dependent approach makes the development process and resulting system very much ad-hoc and not reusable.

Recently, more and more approaches divide the framework into two levels by using mid-level feature extraction to facilitate high level event detection. Overall, these systems show better performance in analyzing the content meaning of sports video. However, these approaches also share two features: First, most of these approaches need to create some heuristic rules in advance and the performance of the system greatly depends on those heuristic rules which make their system not flexible. Second, some approaches use statistical approaches such as HMM to model the temporal patterns of video shots but can only detect relatively simple structured event such as play and break.

From the experimental result point of view, Table 1-1 shows the precision, recall, testing data set, and important assumption of the goal detection systems for soccer videos reported by some of the relevant publications presented recently. As we can see, both approaches proposed in [24] and [26] are based on some important assumptions which make their system not applicable to the soccer videos that do not satisfy the assumptions. The testing data set in [23] is weak, only 1 hour of videos is tested. Moreover, the testing data is extracted from 15 European competitions manually. A generic approach is proposed for goal detection in proposed in [25]. This approach is developed without any important assumption and the authors use 3 hours of videos as their testing data set. However, their precision is relatively low which leaves rooms for improvement.

Table 1-1 Precision and recall reported by other publications

Reference	Precision	Recall	Testing Data Set	Important Assumption
[23]	77.8%	93.3%	1 hour of videos, separated in 80 sequences, selected from 15 European competitions manually	No
[24]	80.0%	95.0%	17 video clips (800 minutes) of broadcast soccer video	Slow motion replay segments must be highlighted by adding special editing effects before and after by the producers.
[25]	50%	100%	3 soccer clips (180 minutes)	No
[26]	100%	100%	17 soccer segments, the length of the game segments range from 5 seconds to 23 seconds	The tracked temporal position information of the players and ball during a soccer game segment must be acquired.

1.2 System Overview

We propose a multi-modal two-level event detection framework and demonstrate it on soccer videos. Our goal is to make our system flexible so that it could be adapted to various events in different domains without much modification. To achieve our goal, we use a mid-level representation called Audio and Visual Keyword (AVK) that can be learned and detected in video segments. AVKs are intended to summarize the video segment in the text form and each of them has its semantic meaning. In our thesis, nine visual keywords and three audio keywords are defined and classified to facilitate highlight detection in soccer videos. Based on AVK, a computational system that realizes the framework comprises two levels of processing:

1. The first level focuses on video segmentation as well as AVK classification. The video stream is partitioned into visual stream and audio stream first. Then, based on the visual information, video stream is segmented into video segments and each segment is labeled with some visual keywords. At the same time, we divided audio stream into audio segments of same lengths. Generally, the duration of the audio segments is much shorter than the average duration of the video segments and one video segment might contain several audio segments. For each video segment, we compute the overall excitement intensity and label each video segment with one audio keyword. In the end, for each video segment, we label two visual keywords and one audio keyword. In other words, the first level analyzes the video stream and outputs a sequence of AVK (Fig. 1-1).

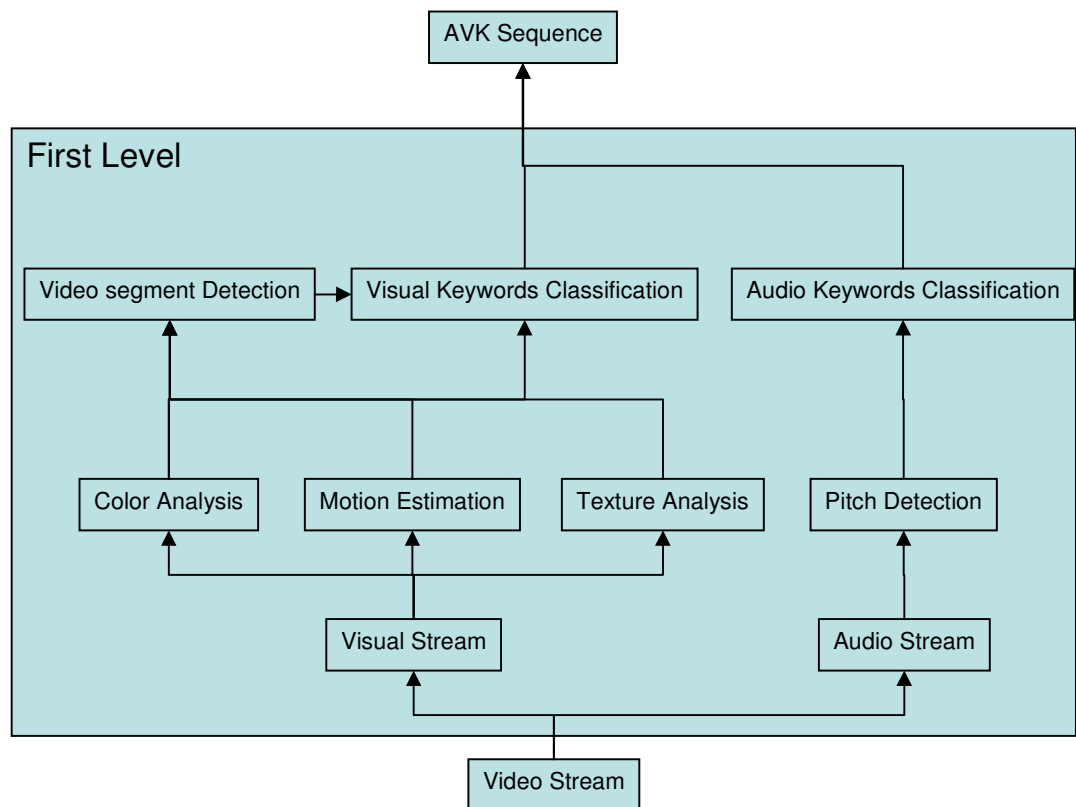


Fig. 1-1 AVK sequence generation in first level

2. Based on the AVK sequence, the second level performs event detection. In this level, according to the semantic meaning of the AVK sequence, we detect the portions of the AVK sequence within which the events we are interested with anchor. At the same time, we also remove the portions of AVK sequence within which no interested event anchors.

In general, the probabilistic mapping between the keyword sequence and the events can be modeled either statistically (e.g. HMM) or syntactically (e.g. grammar). In this thesis, both statistical and syntactical modeling approaches are used to see their performance on event detection in soccer video respectively. More precisely, we develop a unique event detection grammar to parse the goal and corner-kick events from visual keyword sequence; we also apply a HMM classifier to both the visual and audio keyword sequence for goal event detection. For both two approaches, satisfactory results are achieved. In the end, we compare the two approaches by analyzing the advantages and disadvantages of these two approaches.

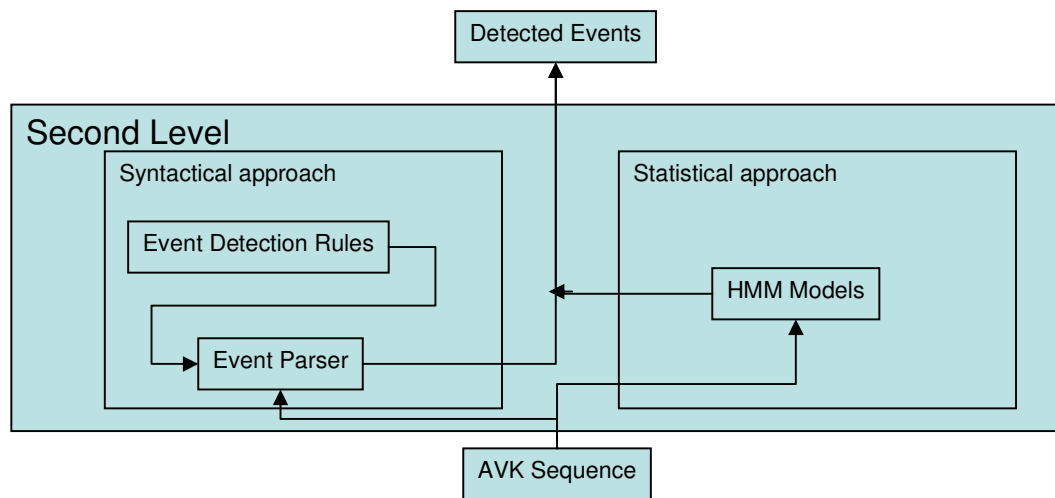


Fig. 1-2 Two approaches for event detection in second level

The two-level design makes our system reconfigurable. It can detect different events by adapting the event detection grammar or re-train the HMM models in the second level. It can also be applied to different domains by adapting the vocabulary of visual and audio keywords and its classifiers or defining new kind of keywords such as text keywords, etc.

1.3 Organization of Thesis

In chapter 2, we survey some related works, and then, discuss the strengths and weaknesses of other event detection systems.

In Chapter 3, we first introduce how we segment video stream into video segments and the different semantic meanings of different classes of video segments. Then, we give the definition of the AVKs and explain why we define them.

In Chapter 4, we first explain how we extract low-level features to segment visual images into Regions of Interest (ROIs). Then, we introduce how we use the ROI information and Support Vector Machines (SVM) to label the video segment with visual keywords. We also present the satisfactory experimental results on visual keywords labeling at the end of this chapter.

In Chapter 5, we first briefly explain how we get the excitement intensity of the audio signal based on twice-iterated Fourier Transform. Then, we introduce how we label the audio segment with audio keywords.

In Chapter 6, we explain how we detect the goal event in soccer videos with the help of AVK sequence. We use two sections to present how we use statistical approach and syntactical approach respectively to detect the goal event in soccer videos. At the end part of each section, experimental results are presented. At the end of chapter 6, we compare these two approaches and analyze the strengths and weaknesses.

Finally, we summarize our work and discuss the possible ways to refine our work and extend our methods to other event detections in Chapter 7.

Chapter 2

Literature Survey

Recent years, an increasing number of event detection algorithms are being developed for sports video [23-26]. In the case of the soccer game that attracts a global viewer-ship, research effort has been focused on extracting high-level structures and detecting highlights to facilitate annotation and browsing. To our knowledge, most of the methods can be divided into two stages: feature extraction stage and event detection stage. In this chapter, we will survey related work in sports video analysis from the feature extraction and detection model point of views respectively. We will also discuss the strengths and weakness of some event detection systems in this chapter.

2.1 Feature Extraction

As we know, sports video data is composed of temporally synchronized multimodal streams such as visual, auditory and text streams. Most of the approaches proposed recently extract some features from the information in the above mentioned three streams. Based on the kind of features

used, we divide the recent proposed approaches into four classes: visual features, audio features, text caption features and domain-specific features.

2.1.1 Visual Features

The most popular features used by researchers are visual features such as color, texture and motion, etc [27-36]. In [36], Xie et al. extract dominant color ratio and motion intensity from the video stream for structure analysis in soccer video. In [32], Huang et al. extract the color histogram, motion direction, motion magnitude distribution, texture directions of sub-image, etc to classify the baseball video shot into one of the fifteen predefined shot classes. In [33], Pan et al extract color histogram and pixel-wise mean square difference of the intensity of every two subsequent fields to detect the slow-motion reply segments in sports video. In [34], Lazarescu et al. describe an application of camera motion estimation to index cricket games by using the motion parameters (pan, tilt, zoom and roll) extracted from each frame.

2.1.2 Audio Features

Some researchers use audio features [37-40], and from the experimental results reported in recent publications, audio features can also contribute significantly in video indexing and event detection. In [37], Xiong et al. employ a general sound recognition framework based on Hidden Markov Models (HMM) using Mel Frequency Cepstral Coefficients (MFCC) to classify and recognize the audio signals such as: applause, cheering, music, speech and speech with music. In [38], the authors use a simple, template-matching based approach to spot important keywords spoken by commentator such as “touchdown” and “fumble”, etc. They also detect the crowd cheering using audio stream to facilitate video indexing. In [39], Rui et al. focus on excited/non-

excited commentary classification for TV baseball programs highlights detection. In [41], Wan et al. describe a novel way to characterize dominant speech by its sine cardinal response density profile in a twice-iterated Fourier transform domain. Good result has been achieved for automatic highlight detection in soccer audio.

2.1.3 Text Caption Features

The text caption features include two types of text information: closed text caption and extracted text caption. For broadcast video, the closed text caption is the text form of the words being spoken in the video and they can be acquired directly from video stream. Extracted text caption is the text that is added to the video stream during editing process. In sports videos, extracted text caption is the text in the caption box which provides important information such as score, foul statistics, etc. Compared to closed text caption, extracted text caption cannot be acquired directly from video stream. It has to be recognized from image frames of the video stream. In [42], Babaguchi et al. make use of closed text caption for video indexing of events such as touchdown (TD) and field goal (FG). In [43], Zhang et al. use extracted text caption to recognize domain-specific characters, such as ball counts and game score of baseball videos.

2.1.4 Domain-Specific Features

Apart from the above mentioned three kinds of general features, some researchers use domain-specific features in order to obtain better performance. Some researchers extract the properties such as the line marks, goal post, etc from image frames or extract the trajectory of the players and ball in the game for further analysis. There are some attempts to detect the slow-motion segments by extracting the shot boundary with flashing transition effect. In [38], the authors make

use of line marks, players' numbers, goal post, etc to improve the accuracy for the touchdown detection. In [44], the authors use players' uniform colors, edges, etc to build up semantic descriptor for indexing of TV soccer videos. In [23], the authors extract five basic playfield descriptors from the playfield lines and the playfield shape and then use a Naive Bayes classifier to classify the image into one of the twelve pre-defined playfield zones to facilitate highlight detection in soccer videos. Players' positions are also used to further improve the system accuracy. In [45], Yow et al. propose a method to detect and track soccer ball, goal post and players. In [46,47], Yu et al. propose a novel framework for accurately detecting the ball for broadcast soccer video by inferring the ball size range from the player size, removing non-ball objects and a Kalman filter-based procedure.

2.2 Detection Model

After the feature extraction, most of the methods either apply some classifiers to the features or use some decision rules to perform further analysis. According to the model adopted by these methods, we divide them into three classes: rule-based model, statistical model and multi-modal based model.

2.2.1 Rule-Based Model

Given the extracted features, some researchers apply decision rules on the features to perform further analysis. Generally, approaches based on domain-specific features and system using two-level frameworks tend to use rule-based model. In [44], Gong et al. apply an inference engine to the line marks, play movement, position and motion vector of the ball, etc to categorize the soccer video shot into one of the nine pre-defined classes. In [23], the authors use Finite State Machine (FSM) to detect the goal, turnover, etc based on some specific features such as players' position

and playfield zone, etc. This approach shows very promising result by achieving 93.3% recall in goal event detection. But it uses too much domain-specific features which makes it very difficult to be applied to other sports video. In [26], Tovinkere et al. propose a rule-based algorithm for goal event based on the temporal position information of the players and ball during a soccer game segment and achieve promising result. But, the temporal position information of the players and ball is labeled manually in their experiments. In [48], Zhou et al. describe a supervised rule-based video classification system as applied to basketball video. The if-then rules are applied to a set of low-level feature-matching functions to classify the key frame image into one of the several pre-defined categories. Their system can be applied to applications such as on-line video indexing, filtering and video summaries. In [49], Hanjalic et al. extract overall motion activity, density of cuts and energy contained in the audio track from video stream, and then, use some heuristic rules to extract highlight portions from sports video. In [50], the authors introduce a two-level framework for play and break segmentation detection. In the first level, three views are defined and the dominant color ratio is used as a unique feature for view classification. Some heuristic rules are applied to the view label sequence in the second level. In [24], Ekin et al. propose a two-level framework to detect the goal event by four heuristic rules such as: the existence of slow motion replay shot, the existence of *before* relation between the replay shot and the close-up shot, etc. This approach greatly depends on the detection of the slow motion replay shot which is spotted by detecting the special editing effect before and after the slow motion replay segment. Unfortunately, for some soccer videos, such special editing effect does not exist.

2.2.2 Statistical Model

Apart from the rule-based models, some researchers aim to provide more generic solutions for sports video analysis [51-53]. Some of them use statistical models. In [32] [33], the authors input the low-level features extracted from video stream to Hidden Markov Models for shot

classification and slow motion shot detection. In [54], Gibert et al. address the problem of sports video classification using Hidden Markov Models. For each sports genre, the authors construct two HMMs to represent motion and color features respectively and achieve an overall classification accuracy of 93%. In [36], the authors use Hidden Markov Models for the play and break segments detection in soccer games. Low-level features such as dominant-color ratio, motion intensity, etc is directly sent to HMM and six HMM topologies are trained to model the play and break respectively. In [55], Xu et al. present a two-level system based on HMMs for sports video event detection. First, the low-level features are sent to HMMs in the bottom layer to get the basic hypotheses. Then, the compositional HMMs in the upper layers add constraints on those hypotheses of the lower layer to detect the predefined events. The system is applied to basketball and volleyball videos and achieves promising result.

2.2.3 Multi-Modal Based Model

Recent years, multi-modal approaches become more and more popular for content analysis in news video and sports video domain. In [38], Chang et al. develop a prototype system for automatic indexing of sports video. The audio processing module is first applied to locate candidates in the whole data. This information is passed to the video processing module which further analyzes the video. Some rules are defined to model the shot transition for touchdown detection. Their model covers most but not all the possible touchdown sequences. However, their simple model provides very satisfactory results. In [56], Xiong et al. make an attempt to combine the motion activity with audio features to automatically generate highlights for golf, baseball and soccer games. In [57], Leonardi et al. propose a two-level system to detect goal in soccer video. The video signal is processed first by extracting low-level visual descriptor from the MPEG compressed bit-stream. A controlled markov model is used to model the temporal evolution of the visual descriptors and find a list of candidates. Then, the audio information such as the audio

loudness transition between the consecutive candidates shot pairs is used to refine the result by ranking the candidate video segments. According to their experiments, all the goal event segments are enclosed in the top twenty-two candidate segments. Since the average number of the goals in the experiment is 2.16, we can say that the precision of this method is not high. The reason for that might is because the authors do not use any color information in their method. In [25], a mid-level representation framework is proposed by Duan et al. to detect highlight events such as free-kick, corner-kick, goal, etc. They create some heuristic rules such as the existence of persistent excited commentator speech and excited audience, long duration within the OPS segment, etc to detect the goal event in soccer video. Although the experimental result shows that their approach is very effective, the decision rules and heuristic model has to be defined manually before detection procedure can be applied. For the events with more complex structure, the heuristic rules might not be clear. In [58], Babaguchi et al. investigate multi-modal approaches for semantic content analysis in sports video domain. These approaches are categorized into three classes: collaboration between text and visual streams, collaboration among text, auditory and visual streams and collaboration between graphics stream and external metadata. In [18,19,21], Chaisorn et al. propose a multi-modal two-level framework. Eight categories are created, and based on which, the authors solve story segmentation problem. Their approach achieves very satisfactory result. However, so far, their approach is applied in news video domain only.

2.3 Discussion

According to our reviews, most of the rule-based approaches have one or two of the following drawbacks:

1. The approaches, either two-level or one-level, need to have the heuristic rules pre-created manually in advance. The heuristic rules have to be changed when a new event is to be detected.
2. Some approaches use much domain specific information and features. Generally, these approaches are very effective and achieve very high accuracy. But due to the domain specific features they use, these approaches are not reusable. Some approaches are difficult to apply to different types of videos in the same domain such as another kind of sports video.
3. Some approaches do not use much domain specific information, but the accuracy is lower.

For the statistical approaches, they use less domain specific features than some rule-based approaches. But in general, their performance on average is lower than those of the rule-based approaches. One observation is that quite a few approaches are presented to detect events such as goals in soccer video using statistical model due to the complex structure of soccer video. By analyzing these statistical approaches, we think that most of them can be improved in one or two of the following aspects:

1. Some approaches feed low-level features directly to the statistical models leaving a large semantic gap between computable features and semantics as understood by humans. These approaches can be improved by adding a mid-level representation.
2. Some approaches use only one of the accessible low-level features so that their statistical models cannot achieve good result due to lack of information. These approaches can be improved by combining different low-level features together such as visual, audio and text, etc.

For the multi-modal based approaches, they use more low-level information than other kinds of approaches and achieve higher overall performances. Recently, multi-modal based model becomes an interesting direction. However, in sports video domain, most of the multi-modal based approaches known to us so far use some heuristic rules which makes these approaches not

flexible. Nevertheless, the statistical based method proposed in [18,19,21] for news story segmentation does not rely on any heuristic rules and attracts our attention. We believe that a statistical based multi-modal integration method should also work fine in sports video domain. Based on our observations, we introduce a mid-level representation called Audio Visual Keyword (AVK) that can be learned and detected from video segments. Based on the AVKs, we propose a multi-modal two-level framework fusing both visual and audio features for event detection in sports video and applied our framework to goal detection in soccer videos. In the next chapter, we will explain the details of our AVK.

Chapter 3

AVK: A Mid-Level Abstraction for Event Detection

In Chapter 1, we introduce a two-level event detection framework. As we can see, the Audio and Visual Keyword serves as a key component in our system. In this chapter, we give the definition and introduce the different semantic meaning of the audio and visual keywords used in our system. We also make comparisons and contrasts between our definition and definitions given by other researchers and explain the motivation of our definition. In the last section of this chapter, we introduce how we segment video stream into video segments.

The notion of visual keywords was initially introduced for content-based image retrieval [59,60]. In the case of images, visual keywords are salient image regions that exhibit semantic meanings and that can be learned from sample images to span a new indexing space of semantic axes such as face, crowd, building, sky, foliage, water etc. In the context of video, visual keywords are extended to cover recurrent and meaningful spatio-temporal patterns of video segments. They are characterized using low-level features such as motion, color, texture etc and detected using

classifiers trained a prior. Similarly, we also use audio keywords to characterize the meaning of the audio signal.

In our system, we use Audio and Visual Keyword (AVK) as a mid-level representation to bridge the semantic gap between low-level features and content meaning as understood by humans. Each of the AVKs defined in our vocabulary has its semantic meaning. Hence, in the second level of our system, we can detect the events we are interested in by modeling the temporal transitions embedded in AVK sequence.

3.1 Visual Keywords for Soccer Video

We define a set of simple and atomic semantic labels called visual keywords for soccer videos. These visual keywords form the basis for event detection in soccer video.

To properly define the visual keywords, we first investigate other researchers' work. In [36], the authors define three basic kinds of views in soccer video: global, zoom-in and close-up, based on which plays and breaks in soccer games are detected. Although good experimental results are achieved, three view types are too few to be used for more complex event detection such as goal, corner-kick, etc. In [24], Ekin et al. introduce the similar definition: long shot, in-field medium shot and close-up or out-of-field shot. In order to detect the goals, the authors use one more visual descriptor i.e. slow-motion shot which only can be detected based on a very important assumption: all the slow motion replay segment starts and ends with a special editing effect which can be detected. Since this assumption is not always satisfied, their approach does not work on some soccer videos. In [25], Duan et al. define eight semantic shot categories for soccer game. Along with the heuristic rules pre-defined, their system achieves very good result. But their definition is not very suitable for statistical based approach. For example: although the two categories "player

following” and “player medium view” share the same semantic meaning except that “player following” has higher motion intensity, they are regarded as two absolutely different categories.

Based on our investigations, we present our definition in this section. From the focus of the camera and the moving status of the camera point of views, we classify the visual keywords into two categories: static visual keywords and dynamic visual keywords. Static visual keywords are used to describe the intended focus of the camera by the camera-man while dynamic visual keywords are used to describe the direction of the camera movement.

(1) Static visual keywords

Visual keywords under this category are listed in Table 3-1.

Table 3-1 Static visual keywords defined for soccer videos

Keywords	Abbreviation
Far view group	
• Far view of whole field	FW
• Far view of half field	FH
Mid Range view group	
• Mid range view (whole body visible)	MW
Close up view group	
• Close-up view (inside field)	IF
• Close-up view(edge field)	EF
• Close-up view(outside field)	OF

In the sports video, the camera might take the playing field or the people outside the playing field from “far view”, “mid range view” or “close-up view” (Fig. 3-1).



Fig. 3-1 Far view (left) mid range view (middle) close-up view (right)

Generally, “far view” indicates that the game is playing and no special event happens so the camera captures the field from far to show the whole status of the game. “Mid range view” always indicates the potential defend and attack so that the camera captures players and ball to follow the actions closely. “Close-up view” indicates that the game might be paused due to the foul or the events like goal, corner-kick etc so that camera captures the players closely to follow their emotions and actions. In the slow motion replay segment and segments before the corner-kick and free-kick etc, camera is always in “mid range view” or “close-up view”. For other segments, camera is always in “far view”.

Hence, we define three groups under this category: “far view group”, “mid range view group” and “close-up view group”.

As we discussed before, three static visual keywords “FW” “MW” and “CL” cannot get good result in the second level of our system. Because of this, within each group, we further define one to three static visual keywords.

For “far view” group, we define “FW” and “FH” (Fig. 3-2). If camera captures only the half field so that the whole goal post area or part of the goal post area could be seen, we define it as “FH”. We include “FH” in our vocabulary because video segment that is labeled as “FH” gives us more detailed information than “FW”. It tells us that, at the moment, the ball is near the goal post, suggesting an attack or some potential goals. Generally, most of the interesting events like goal, free-kick (near penalty area) and corner-kick all start from a video segment labeled as “FH”. Indeed, from our experiments, we verify that the use of “FH” improves the accuracy in event detection greatly.



Fig. 3-2 Far view of whole field (left) and far view of half field (right)

For “mid range view” group, we only define one visual keyword: “MW” which stands for “Mid range view (whole body is visible)” (Fig. 3-3). Generally, short-length “MW” video segment indicates the potential attack and defend. Long-length “MW” video segment indicates that the game is paused. For example, when the referee shows the red card, some players run to argue with the referee. The whole process which lasts for more than ten seconds might all be “mid range view”.



Fig. 3-3 Two examples for mid range view (whole body is visible)

For “CL” group, we define “OF”, “IF” and “EF”. We will explain the definition and reason for each visual keyword one by one.

When camera captures the playing field as background and zooms in on a player, it is labeled as “IF” which stands for “In the field”. When camera captures part of the playing field as background and one player stands at the edge or inside the playing field, it is

labeled with “EF” which stands for “edge of the field”. When camera does not capture playing field at all, it is labeled as “OF” which stands for “Out of field”.

When the ball goes out of the field, the game will pause for a while. Later, one player runs to get the ball back and then makes a serve. It is at this moment that the “EF” shot appears. Generally, the appearance of “EF” shot always accompanies the event like throw in, corner-kick etc (Fig. 3-4).



Fig. 3-4 Edge of the field

If for some reasons (such as foul, after the goal and so on), the game pauses for a relatively long time (such as several seconds or longer), there is no interesting action happening in the playing field, then, the camera will focus on the audience and coaches. Especially for the video segment after goal event, the audience and some coaches are cheering while some coaches look very sad. The camera will continue to focus on the audience and coaches for several seconds. In that case, there might be several consecutive “OF” shots (Fig. 3-5).



Fig. 3-5 Out of the field

There are many places that the “IF” segment might appear: after the foul, when the ball goes out of the field, after the goal event and so on. The appearance of the “IF” segment does not give us much useful information in event detection. Generally, we only know that the game might be suspended when we see this keyword (Fig. 3-6).



Fig. 3-6 Inside the field

Initially, we also include some visual keywords like the visual appearance of the referee, coach and goalkeeper in our vocabulary. Later, we found that using these visual keywords does not improve the accuracy much while we had to extract many domain features such as the color of referees and coaches in order to distinguish players from referees or coaches. Consequently, we removed those visual keywords from our visual keyword set and both referee and coach are treated in the same way as players. Meanwhile, we also tried to include a visual keyword --- “Slow Motion” in our vocabulary. But unfortunately, different broadcast companies use different special editing effect before and after a slow motion replay segment. Moreover, for some soccer videos, there is not any special editing effect used before and after slow motion replay segments at all. Because of this, we removed that visual keyword from our vocabulary.

(2) Dynamic visual keywords

Visual keywords under this category are listed in Table 3-2.

Table 3-2 Dynamic visual keywords defined for soccer videos

Keywords	Abbreviation
Still	ST
Moving	MV
Fast moving	FM

In essence, dynamic visual keywords based on motion features intend to describe the camera's motion. Below are some examples for the dynamic visual keywords in which the superimposed black edges are the motion vectors. (Fig. 3-7)

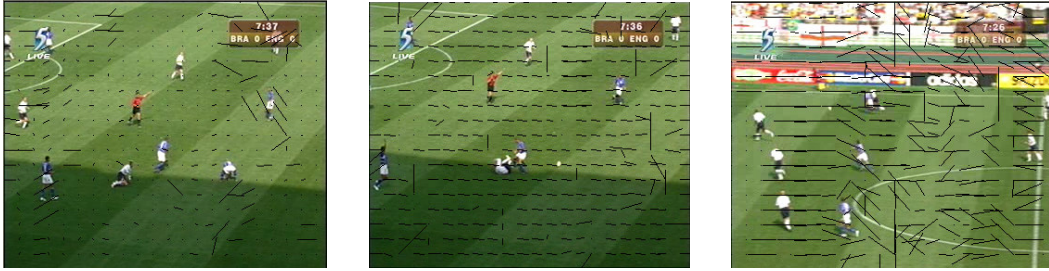


Fig. 3-7 Examples for dynamic visual keywords
still (left) moving(middle) fast moving(right)

Generally, if the game is in play, the camera always follows the ball. If the game is in break, the camera tends to capture the people in the game. Hence, if the camera moves very fast, it indicates that either the ball is moving very fast or the players are running. For example: given a “far view” video segment, if the camera is moving, it indicates that the game is playing and the camera is following the ball; if the camera is not moving, it indicates that the ball is static or moving slowly which might indicate the preparation stage before the free-kick or corner-kick in which the camera tries to capture the distribution of the players from far.

In practice, we label each video segment with two visual keywords: one static visual keyword and one dynamic visual keyword.

3.2 Audio Keywords for Soccer Video

In soccer videos, the audio signal consists of the speech of the commentators, cheers of the audience, shout of the players, whistling of the referee and environment noise. The whistling, excited speech of commentators and sound of audience are directly related to the actions of the people in the game which are very useful for structure analysis and event detection.

Recent years, many approaches have been presented to detect the excited audio portions [33-36]. For our system, we define three audio keywords: “Non-Excited”, “Excited” and “Very Excited” for soccer videos. In practice, we sort the video segments according to their average excitement intensity. The top 10% video segments are labeled with “Very Excited”, video segments whose average excitement intensity are below top 10% higher than top 15% are labeled with “Excited”. Other video segments are labeled with “Non-Excited”. Initially, we also include another audio keyword “Whistle” in our vocabulary. According to soccer games rules, most of the highlights happen along with different kinds of whistling. For example: Long whistling always indicates the start of corner-kick, free-kick or penalty kick. Three consecutive whistling indicate the start or end of the game. Ideally, detection of whistling should facilitate the event detection in soccer videos greatly. Unfortunately, the sound of the whistling is sometimes overwhelmed by the noise of the audience and environment. Hence, we remove the “whistle” from our audio keywords vocabulary.

3.3 Video Segmentation

Generally, the first step in video processing is to detect the shot boundaries and segment video stream into shots which are usually defined as the smallest continuous unit of a video document. But the traditional shot might not correspond to the semantic meaning in soccer video quite well. For some video shots, different parts of them have different semantic meaning and ought to be further divided into several sub-shots.

For example: when the camera pans from mid field to goal area, according to custom shot definition, there is only one shot. But since the semantic meaning of mid field and goal area are different, we need to further segment that shot into two sub shots (Fig. 3-8).

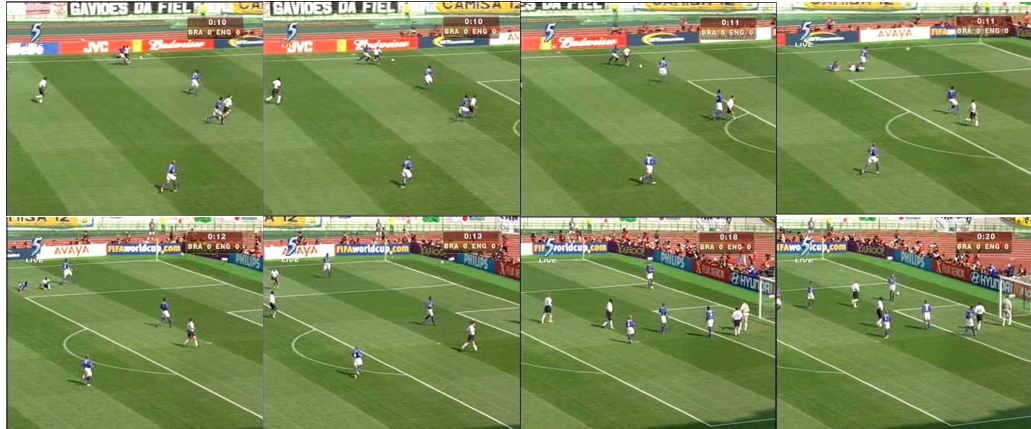


Fig. 3-8 Different semantic meaning within one same video shot

Here is another example: Fig. 3-9 shows several image frames that are extracted from a video shot. The first half part of this video shot shows several players, some of them are defending and one of them is attacking. The game is still in play. And the camera captures the whole body of the players along with the ball in order to follow the players' actions. In the second half of this video shot, the game is paused due to the goal. The camera zooms in a little and focuses at the upper-half body of the attacking player to capture his emotions. Although the two halves of the video shot have different semantic meaning, they are segmented into one video shot using traditional shot segmentation approach.



Fig. 3-9 Different semantic meaning within one same video shot

Another problem we met is that the accuracy of the shot segmentation approaches based on color histogram in sports domain is not as high as in other domains. Generally, these shot segmentation algorithms locate the shot boundary by detecting a large change in color histogram differences. However, the similar color within the playing field and the high ratio appearance of the playing field makes the color histogram difference between two consecutive shots lower in sports domain. Moreover, the frequent used gradual transition effect between two consecutive shots in soccer videos makes shot boundary detection more difficult (Fig. 3-10).



Fig. 3-10 Gradual transition effect between two consecutive shots

Using motion, edge and other information in shot segmentation stage could improve the shot segmentation accuracy [61]. But meanwhile, it also increases the computational complexity. Since our objective in this thesis is event detection, we are not going to spend much effort in shot segmentation stage. Hence, we have decided to further segment the video shots into sub-shots instead. In practice, we perform conventional shot classification using color histogram approach, and insert shot boundaries within a shot whose length is longer than 100 frames to further segment the shot into sub shots evenly. For instance, a 130-frame shot will be further segmented into two sub-shots evenly, namely 65-frame each.

Chapter 4

Visual Keyword Labeling

In Chapter 3, we define six static visual keywords, three dynamic visual keywords and three audio keywords. In this chapter, we will describe how to extract low-level features and label each video segment with one static visual keyword [62] and one dynamic visual keyword.

The key objective of visual keywords labeling is to use the labeled segments for event detection and structure analysis later. In our system, visual keywords are labeled on frame level. I-Frame (also called a key-frame) has the highest quality since it is the frame that compressor examines independent of the frames that proceed and follow it. Hence, we label two visual keywords for every I-Frame in a video segment, and then, we label the video segment with the visual keywords of the majority of frames. Our approach comprises five steps of processing (Fig. 4-1):

1. Pre-processing: In this step, we use Sobel edge detector [63] to extract all the edge points within each I-Frame and convert each I-Frame of the video stream into edge-based binary map. At the same time, we also convert each I-Frame into color-based binary map by detecting dominant color points.

2. Motion information extraction: In this step, some basic motion information is extracted such as the motion vector magnitude, etc.
3. Playing field detection and Regions of Interest (ROIs) segmentation: In this step, we detect the playing field region from the color-based binary map and then we segment the ROIs within the playing field region.
4. ROI feature extraction: In this step, ROI properties such as size, position, shape, and texture ratio are extracted from the color-based binary map and edge-based binary map we computed in Step 1.
5. Keyword labeling: Two SVM classifiers and some decision rules are applied to the ROI properties we extracted in Step 4 and playing field region we obtained in Step 3 to label each I-Frame with one static visual keyword. Motion information extracted in Step 2 is also used to label each I-Frame with one dynamic visual keyword.

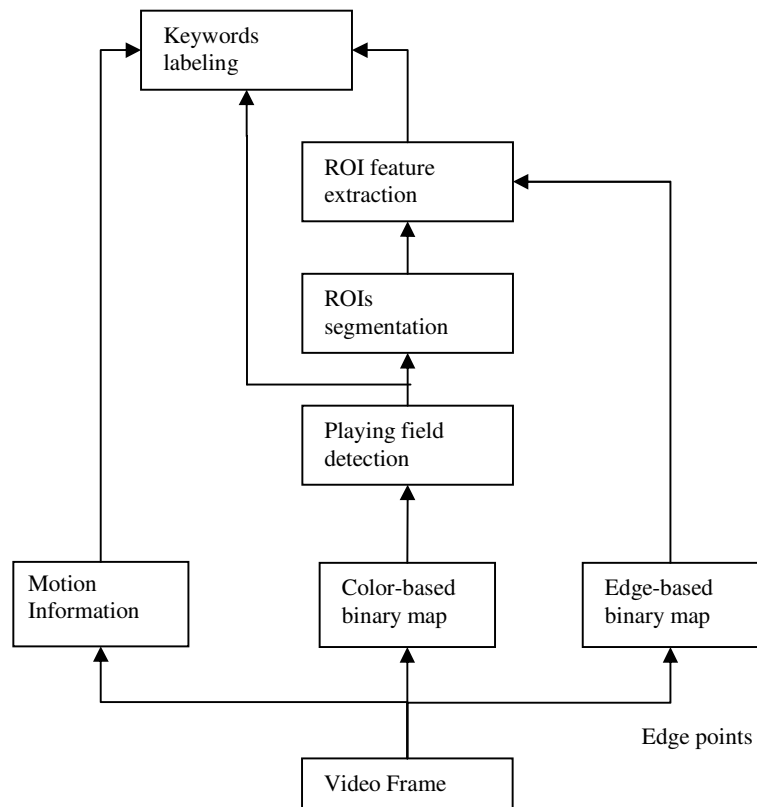


Fig. 4-1 Five steps of processing

This chapter is organized as follow: Section 4.1 describes pre-processing stage; it includes how to extract the edge points and dominant color points. Feature extraction and keywords labeling are explained in Section 4.2 and Section 4.3 respectively. Last but not least, in Section 4.4, we report the promising experimental result.

4.1 Pre-Processing

4.1.1 Edge Points Extraction

It has been shown that the edge map of the image contains a lot of essential information. Before we begin our consideration of video segment labeling, we need to consider the problem of edge detection.

There are some popular gradient edge detectors like Roberts, Sobel, and so on. Since we need to detect both horizontal and vertical edge components, we have selected the Sobel operator as our edge detector.

Given the I-Frame bitmap $Map_{original}$, we use three steps to get edge-based binary map.

- (1) We convolve the Sobel kernels to $Map_{original}$.

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad \text{Equ. 4-1}$$

$$K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad \text{Equ. 4-2}$$

$$Map_{gradient(x)} = conv(Map_{original}, K_x) \quad \text{Equ. 4-3}$$

$$Map_{gradient(y)} = conv(Map_{original}, K_y) \quad \text{Equ. 4-4}$$

$$Map_{gradient}[x, y] = |Map_{gradient(x)}[x, y]| + |Map_{gradient(y)}[x, y]| \quad \text{Equ. 4-5}$$

$C = conv(A, B)$ where A is a $w_a \times h_a$ matrix, B is a $w_b \times h_b$ matrix is defined as:

$$c(x, y) = \sum_{i=1}^{w_b} \sum_{j=1}^{h_b} a(x+i-1, y+j-1) \times b(i, j) \quad \text{Equ. 4-6}$$

- (2) Use a liner filter to map all the elements in $Map_{gradient}$ to the numbers range from 0 to 255.

$$E_{\min} = \min_{i=1, \dots, width, j=1, \dots, height} Map_{gradient}[i, j] \quad \text{Equ. 4-7}$$

$$E_{\max} = \max_{i=1, \dots, width, j=1, \dots, height} Map_{gradient}[i, j] \quad \text{Equ. 4-8}$$

$$E_{dis} = E_{\max} - E_{\min} \quad \text{Equ. 4-9}$$

$$Map'_{gradient}[x, y] = \frac{(Map_{gradient}[x, y] - E_{\min}) \times 255}{E_{dis}} \quad \text{Equ. 4-10}$$

- (3) The result is consolidated into binary map by a threshold and setting all the points that is greater than the threshold to 1 and the others to 0. Finally, we get the edge-based binary map Map_{edge} (Fig. 4-2).

$$Map_{edge}[x, y] = \begin{cases} 0 & Map'_{gradient}[x, y] < t \\ 1 & Map'_{gradient}[x, y] > t \end{cases} \quad \text{Equ. 4-11}$$

where t is the threshold, and it is set to 125 in practice.

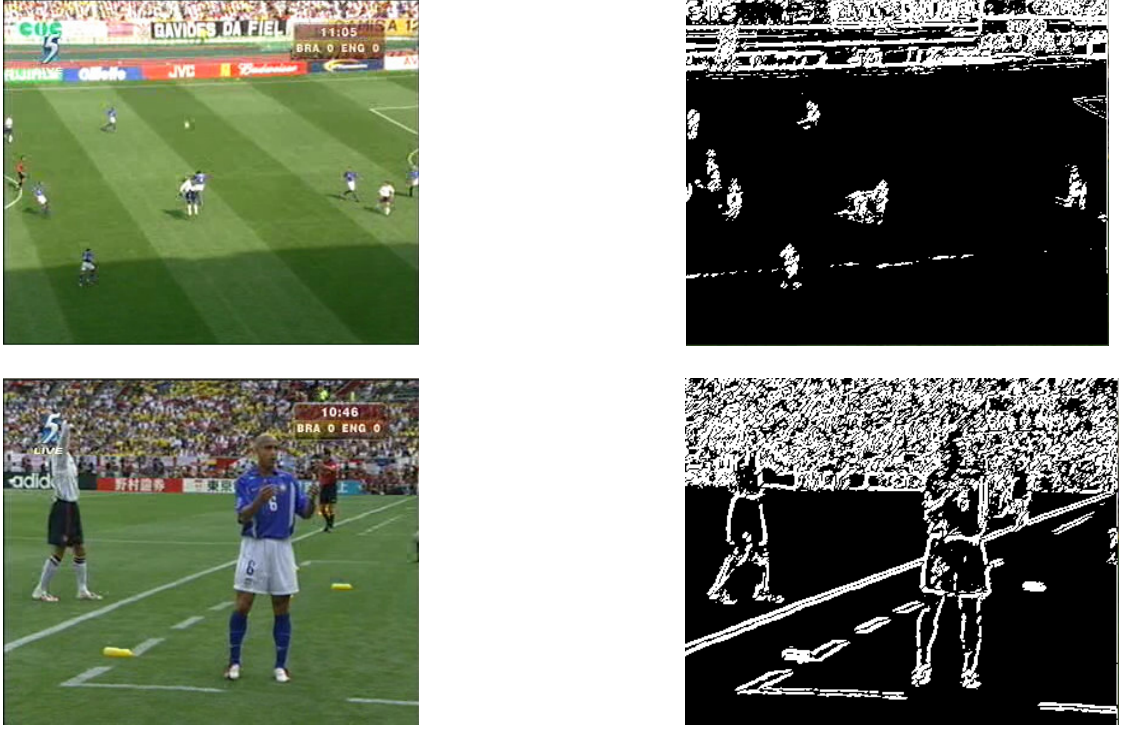


Fig. 4-2 I-Frame (left) and its edge-based map (right)

4.1.2 Dominant Color Points Extraction

For most of the sports videos, there is a playing field with players. Since most of the visual keywords we define are related to the playing field, and the distribution of the field pixels within each frame can help us in determining which visual keyword the frame should be labeled with, we detect the field region as our first step. To do that, we convert each I-Frame bitmap into a color-based binary map by setting all the pixels that are within the field region into black pixels and other pixels into white pixels. Since, for soccer videos, the field is always characterized by one dominant color, we simply get the color-based binary map by mapping all the dominant color pixels into black pixels and non-dominant color pixels into white pixels.

In order to deduce the information we need to process, we sub sample the color based binary image with a 8x8 window into a 35x 43 matrix denoted as Map_{color} (Fig. 4-3).

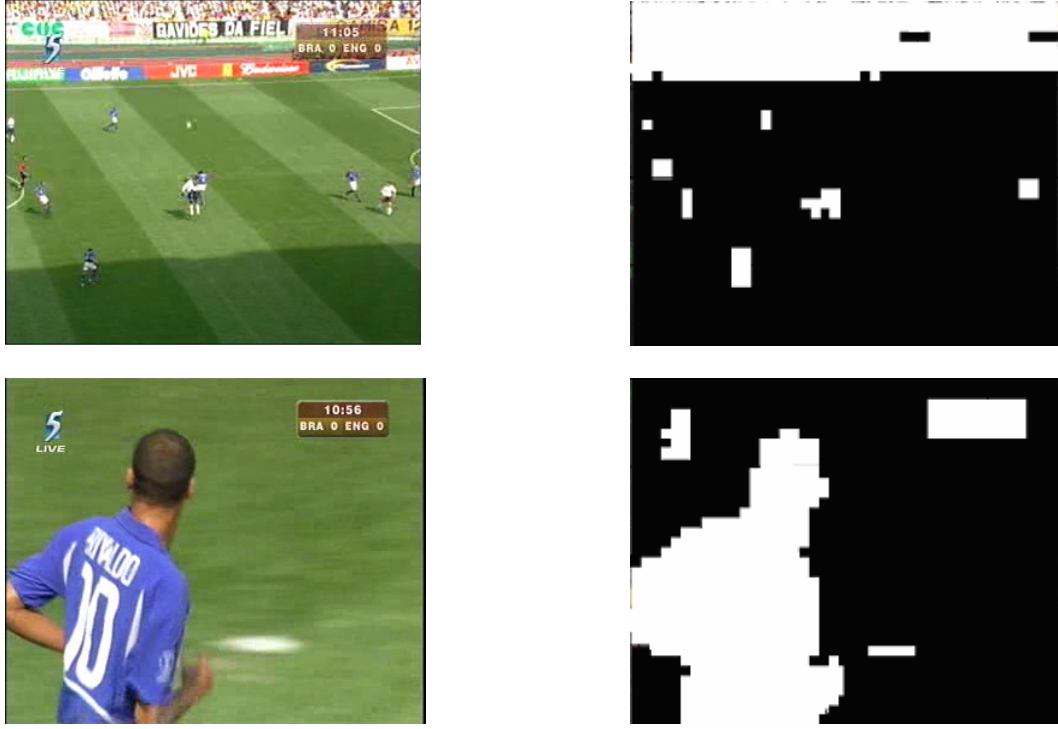


Fig. 4-3 I-Frame (left) and its color-based map (right)

4.2 Feature Extraction

4.2.1 Color Feature Extraction

Color is a very important feature for visual keyword labeling and color information is obtained by decoding each I-Frame in the video stream.

Our basic idea is to detect the playing field region and the ROIs inside playing field region first, and then, we use the information we get from the ROIs inside playing field and the position of the playing field to label each I-Frame with one static visual keyword.

After decoding a I-Frame, we convert it into a color-based binary map by setting all the dominant color pixels into black pixels and non-dominant color pixels into white pixels. The color-based binary map is denoted as Map_{color} . Meanwhile, we also extract all the edge points in the I-Frame to get an edge-based binary map denoted as Map_{edge} .

Y-axis Projection

Given the color-based binary map Map_{color} , we project it to the Y-axis by the following formula.

$$P_y(j) = \sum_{i=1}^{43} Map_{color}[i, j] \quad j = (1, 2, 3, \dots, 33, 34, 35) \quad \text{Equ. 4-12}$$

(Map_{color} is a 35×43 matrix)

P_y is very useful in deciding whether a frame is in “far view” or not. For “far view” frame, there are many elements of P_y that are very small and some of them are equal to zero. Otherwise, most of the elements of P_y are very large and some of them are even bigger than 30. For some non-“far view” frame, there will be several elements of P_y which are equal to zero, but the number of the zero elements of P_y is much less.

Field Edge Position

For sports video, the color information outside playing field is less important than the color information inside field. Because of this, we extract more color features from inside field than from outside field. To do that, we need to detect the field region first. By studying the soccer videos, we observe that, for most of the frames, there is a very clear edge between the soccer field and other regions. Generally, that edge is consisted of two horizontal lines. We use two variables --- H_1 & H_2 --- to describe the edge. H_1 is the distance between the top edge line to the top border. Similarly, H_2 is the distance between the bottom edge line to the bottom border.

In practice, we use the following formulas to get H_1 & H_2 .

$$H_1 = \begin{cases} \min(j \mid P_{j-1} > t, P_j < t) & P_1 > t \\ 1 & P_1 < t \end{cases} \quad \text{Equ. 4-13}$$

$$H_2 = \begin{cases} \max(j \mid P_{j-1} > t, P_j < t) & P_{35} > t \\ 1 & P_{35} < t \end{cases} \quad \text{Equ. 4-14}$$

where t is the threshold and is set to be $43 \times \frac{6}{7} \approx 36$ in practice

The positions of those lines are very helpful in video segments labeling. Generally, the H_1 in “Far View” shot / frame ranges from 0 to 18. For “Mid range view” segment, H_1 might be 0 or varies from 15-34. For “Close up view”, H_1 is equal to 0 or 35, for some cases, H_1 might be a number between 10 and 20.

ROI segmentation

Given the color-based binary map of the I-Frame, it is quite easy for us to segment the whole bitmap into ROIs simply by segmenting each consecutive region as one separate ROI. As we mentioned before, the color information outside field is less important than the color information inside field. We only segment the ROIs within the field region.

The ROIs we segment from color-based binary map are denoted as

$$R = \{R_1, R_2, R_3, \dots, R_{n-1}, R_n\}$$

where n is the number of the ROIs

A ROI R_j is denoted as

$$R_j = \{D_{j1}, D_{j2}, D_{j3}, \dots, D_{jm(j)-1}, D_{jm(j)}\}$$

where $m(j)$ is the number of pixels within ROI R_j

A pixel D_{ji} is denoted as

$$D_{ji} = (x_{ji}, y_{ji})$$

where x_{ji}, y_{ji} is the coordinate for point D_{ji} .

After we segmented the ROIs, we need to compute some properties about the ROIs.

Basic ROI Information

For a ROI R_j , it is very easy for us to compute its size as

$$Size_j = m(j) \quad \text{Equ. 4-15}$$

The size of the ROI varies for different visual keywords. For “Far view”, the ROIs are always smaller than 20 pixels. For “Mid range view” and “Close up view”, the ROI size is larger.

We also find the position of each ROI by finding the left-top corner and right-down corner of the minimum rectangle which can accommodate the ROI.

$$D_{top-left} = (x_{top-left}, y_{top-left}) \quad \text{Equ. 4-16}$$

$$x_{top-left} = \min_{i=1,2,\dots,m(j)} (x_{ji}), y_{top-left} = \min_{i=1,2,\dots,m(j)} (y_{ji})$$
$$D_{bottom-right} = (x_{bottom-right}, y_{bottom-right})$$
$$x_{bottom-right} = \max_{i=1,2,\dots,m(j)} (x_{ji}), y_{bottom-right} = \max_{i=1,2,\dots,m(j)} (y_{ji}) \quad \text{Equ. 4-17}$$

ROI shape

Generally, the possible ROIs inside field area include: player, ball, line, goal net, score board and so on. Some ROIs are regarded as noises because their existence affects our accuracy in visual keyword labeling. Since different kinds of ROI tend to have different shape, our basic idea is to use ROI shape to discard those irrelevant ROIs such as the score board and line. Basically, we classify ROI shape into 3 classes:

(1) Rectangles

Generally, the score board and some of the lines inside playing field appear as a rectangle inside playing field. For this kind of ROI, their information does not help us in visual keyword labeling much. We discard the ROI with this shape.

(2) Triangles

Generally, ROIs that appear as a triangle are the “goal post” area that stands at the edge of the field. This kind of ROIs always appears in “FH” shot. Hence, we use the positions and the shapes of this kind of ROIs to detect “FH” shot.

(3) Others

Generally, ROIs in this class are most likely to be players. ROIs in this class are useful for us in visual keyword labeling. Only the information of the ROIs in this class will be input into our SVM classifier.

In order to classify each ROI into one of the three classes in terms of their shape, we first find the minimum rectangle that contains the ROI and then divide the rectangle into 4 areas which is shown in Fig. 4-4:

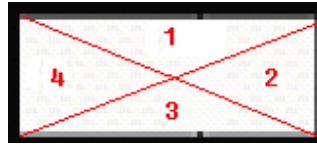


Fig. 4-4 Template for ROI shape classification

We calculate the number of the whites pixels within each area denoted as T_1, T_2, T_3, T_4 . Let $T = T_1 + T_2 + T_3 + T_4$ and S represents the size of the rectangle. We use the following rules to classify the ROI shape into one of three classes mentioned before (Table 4-1):

Table 4-1 Rules to classify the ROI shape

Condition	ROI Shape
$T \sim S$	<i>rectangle</i>
$T_i + T_{i+1} \sim \frac{S}{2}, T - T_i - T_{i+1} \sim 0$	<i>triangle</i>
$T_1 + T_4 \sim \frac{S}{2}, T - T_1 - T_4 \sim 0$	<i>triangle</i>
<i>others</i>	<i>others</i>

The shape of the ROI can help us in discarding those noisy ROIs such as the score board and lines. In the classification stage, we process only on the ROIs that are believed to be players.

Texture ratio

We define the texture ratio as

$$R_{texture} = \frac{Num_{edge}}{Size} \times 100\% \quad \text{Equ. 4-18}$$

For “far view” segments, since the ROIs inside the playing field are relatively small, the texture ratio should be relatively higher. For “close-up view” and “mid range view” segments, since the ROIs sizes are relatively larger and there are not many edges inside the ROIs, the texture ratio should be relatively lower.

4.2.2 Motion Feature Extraction

Since the motion vector information is coded into compressed MPEG video streams, we can extract the motion features from MPEG video streams directly. We use the distribution of the directions and magnitudes of the motion vectors to label segment with dynamic visual keywords.

In practice, we first classify each motion vector into one of nine regions according to their directions and then we calculate the number of motion vectors within each region (Fig. 4-5) denoted as $Region_{motion}(i)$ $i = 1, 2, \dots, 7, 8, 9$.

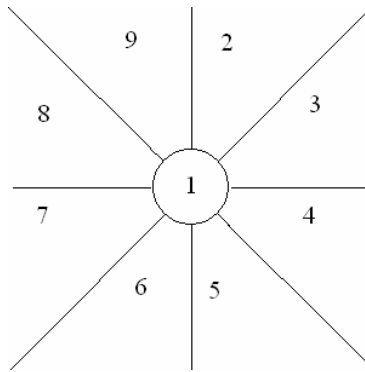


Fig. 4-5 Nine regions for motion vectors

Later, we calculate the mean and standard deviation of the $Region_{motion}(i)$ by:

$$Region_{mean} = \frac{\sum_{i=2}^9 Region_{motion}(i)}{8} \quad \text{Equ. 4-19}$$

$$Region_{std} = \sqrt{\frac{\sum_{i=2}^9 (Region_{motion}(i) - Region_{mean})^2}{8}} \quad \text{Equ. 4-20}$$

We also need to know the scale of the motion vectors. We calculate the average magnitude of all the motion vectors denoted as Mag_{motion} .

If $Region_{std}$ is relatively big, it means that there is one dominant direction among all the motion vectors. If $Region_{std}$ is relatively small, it means that motion vectors tend to have different directions.

4.3 Visual Keyword Classification

We label two visual keywords for every I-Frame in a video segment, and then, we label the video segment with the visual keywords of the majority of frames.

4.3.1 Static Visual Keyword Labeling

After feature extraction, we use those features to label the video segments with visual keywords. Since different features have different discrimination power in different visual keyword labeling, we do not use one single SVM classifier with all the features. Instead, we adopt a progressive classification approach with a hierarchical classifier structure (Fig. 4-6). Two SVM classifiers are applied to the features we extracted to classify the I-Frame bitmap into “far view”, “mid range view” or “close-up view”. Then, different decision rules and threshold are used to label each I-Frame with a certain static visual keyword. For each SVM classifier, we choose the most suitable features for it.

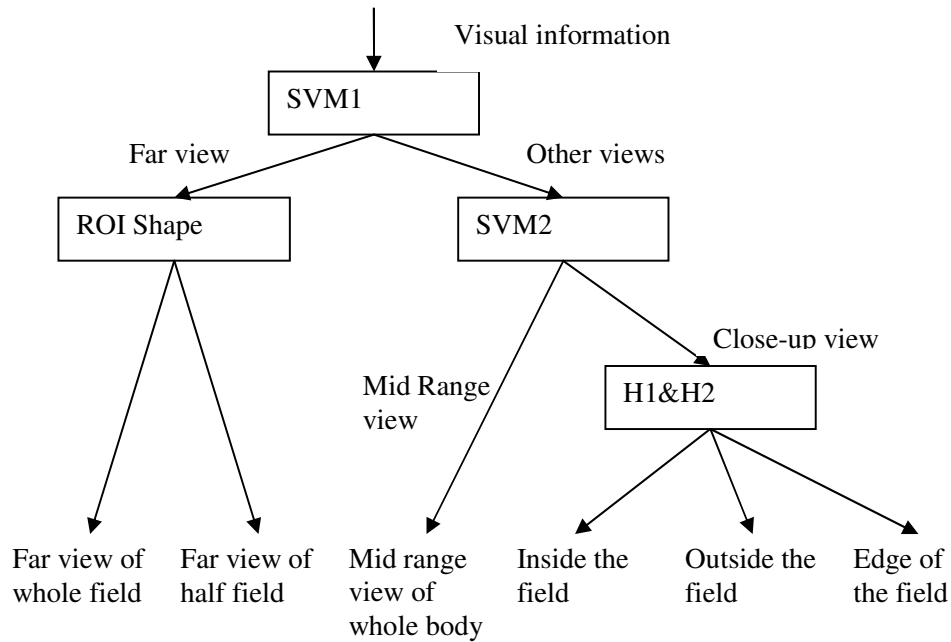


Fig.4-6 Classifiers for color based keywords classification

Features that are useful to classify a video segment into “far view” or “other views” includes: playing field position and some basic information of the ROIs. Since there might be more than one player in the playing field, we sort the ROIs in size and pay more attention to big ROIs which are usually the focus. In practice, we send y-axis projection, field edge position, texture ratio and the sizes of the largest two ROIs whose shape is in “others” class to the first SVM classifier to classify the I-Frame into “far view” or “other views”.

For “other views” I-Frame, a second SVM classifier is applied to further classify it into “mid range view” or “close-up view”. The input of SVM2 includes field edge position, texture ratio, ROI position and the sizes of the largest two ROIs whose shape is in “others” class.

We could have used two more SVM classifiers to further classify the video segments. But, since further classifications can be easily achieved by using two sets of decision rules, we tune the thresholds of the decision rules empirically instead.

If a “far view” video segment has at least one triangle shaped ROIs spotted, “FH” will be labeled to the I-Frame; otherwise, we will label “FW” to the I-Frame.

For “close-up view” video segment, we use the following rule to decide which visual keyword should be labeled to the I-Frame:

$H_1 = 0,1$	IF	Equ. 4-21
$1 < H_1 < 34$	EF	
$H_1 = 34,35$	OF	

4.3.2 Dynamic Visual Keyword Labeling

We use the following rules to label the dynamic visual keywords (Fig. 4-7):

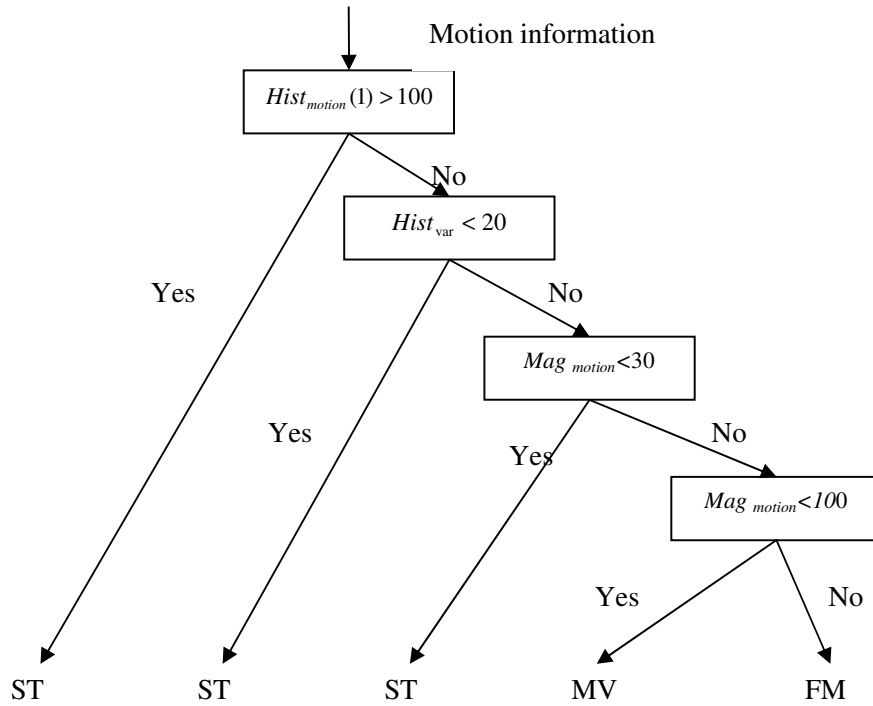


Fig. 4-7 Rules for dynamic visual keyword labeling

4.4 Experimental Results

The proposed framework has been implemented on a windows XP platform. We implemented two tools: one for low level feature extraction (Fig. 4-8), the other for ground truth labeling (Fig. 4-9).

In Fig. 4-8, the left image is the I-Frame of the video stream; those black edges on the left image are the motion vectors. The right image is the color-based binary bitmap. In Fig. 4-9, there is one grid control which lists the detailed information of the video segments such as the start frame number and end frame number of the video segment, etc. The right-top image is the video stream that is showing. The bitmaps at the bottom are the key-frames of six consecutive video segments start from the video segment currently selected by the user.

We use SVM Light toolkit [64] for classification. The kernel function is set to be linear function and the parameter is set to 3. 40-dimension feature vectors are applied to SVM1 and 12-dimension feature vectors are applied to SVM2.

Ten halves soccer matches (8214 video segments, 459 minutes) from FIFA 2002 and UEFA 2002 were used in the experiments to evaluate the system performance on static visual keywords labeling. Three of ten halves soccer matches are used for training and the rest of the soccer videos (5750 video segments, 311 minutes) are used for testing.

Table 4-2 lists the numbers of ground truth, the numbers of false negatives and the numbers of false positives for each match in detail. We also calculate the precision and recall for each static visual keywords and listed in Table 4-3. For the dynamic visual keywords, since they are classified based on some decision rules, we do not evaluate its performance here.

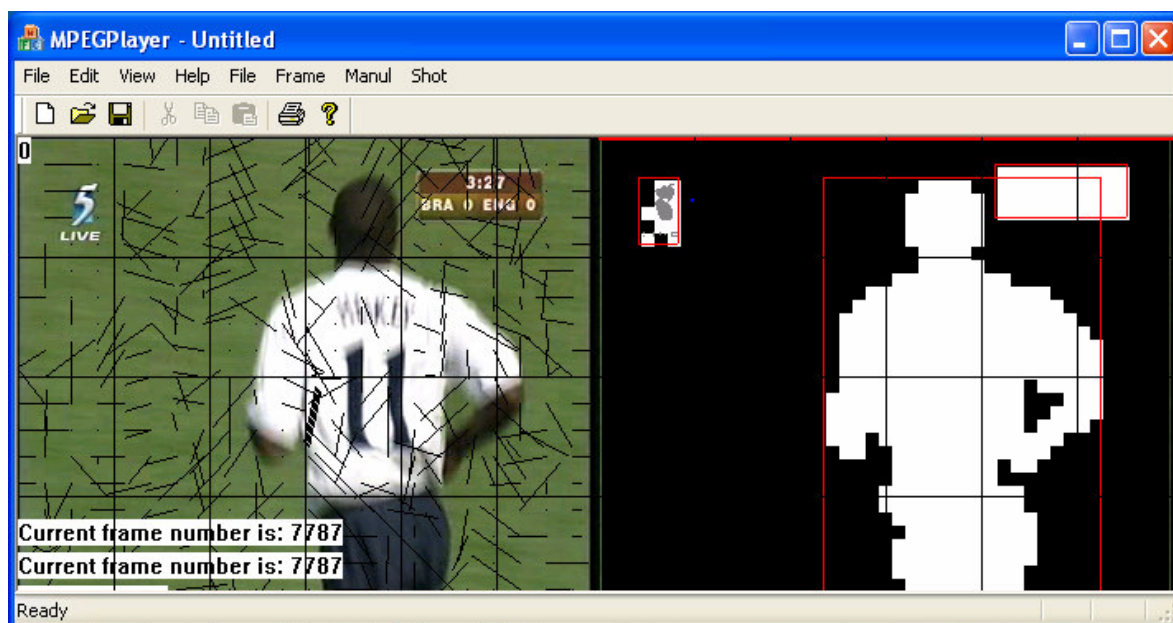


Fig. 4-8 Tool implemented for feature extraction

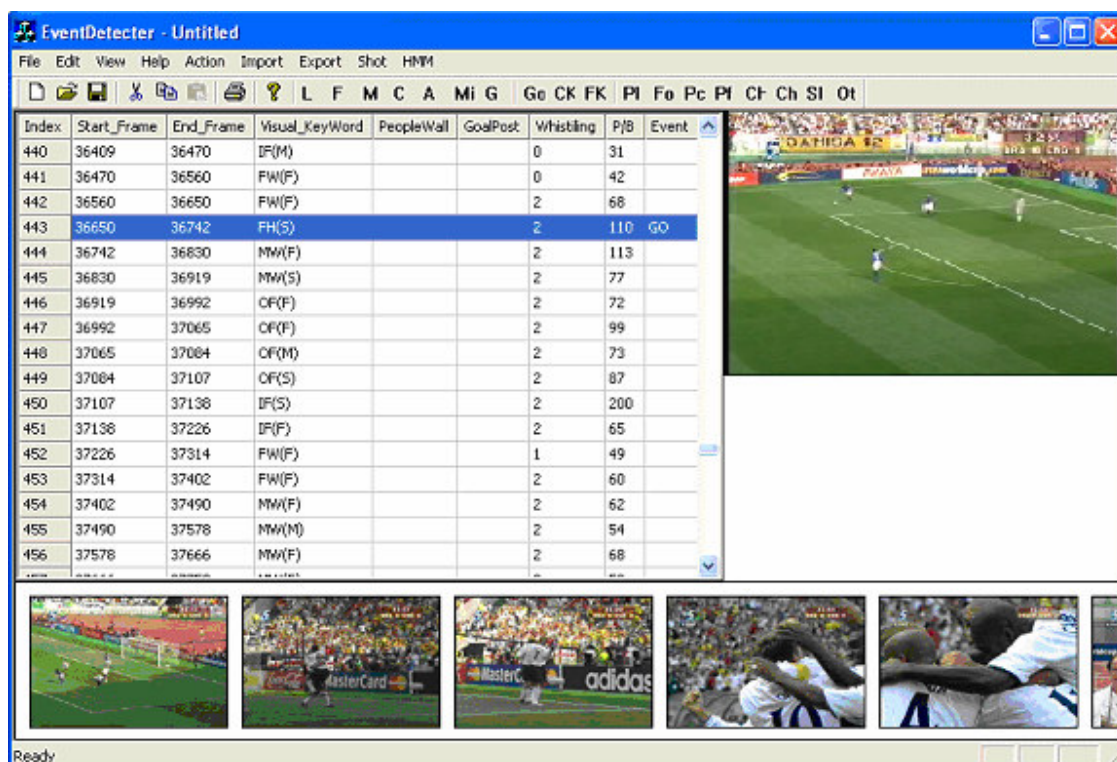


Fig. 4-9 Tool implemented for ground truth labeling

Table 4-2 Experimental Results

Match	KW	FW	FH	MW	IF	EF	OF
SEN	Truth	411	62	145	70	32	93
Vs	Miss	18	8	19	0	0	0
TUR	False	17	9	10	6	3	0
GER	Truth	429	102	51	124	19	116
Vs	Miss	21	17	11	0	0	0
ENG	False	11	10	6	15	7	0
LEV	Truth	272	128	103	79	27	152
Vs	Miss	20	26	11	0	0	0
LIV	False	23	10	9	12	3	0
LIV	Truth	368	129	92	83	13	143
Vs	Miss	29	17	19	0	0	0
LEV	False	27	9	18	8	3	0
BRA	Truth	384	85	93	108	28	98
Vs	Miss	21	15	18	0	0	0
GER	False	18	9	12	9	6	0
USA	Truth	301	125	176	82	36	120
Vs	Miss	22	15	16	0	0	0
GER	False	16	9	13	7	8	0
KOR	Truth	387	98	136	93	30	127
Vs	Miss	24	20	20	0	0	0
TUR	False	20	10	16	10	8	0
Total	Truth	2552	729	796	639	185	849
	Miss	155	118	114	0	0	0
	False	132	66	84	67	38	0

On average, we have achieved 90.7% precision for all the six static visual keywords and 89.6% recall for the first three static visual keywords, namely, “FW”, “FH” and “MW”. According to our experimental result, our system achieves good precision and recall in “FW” and “FH” labeling which are classified by SVM1. By comparison, both the precision and recall achieved by SVM2 is not as high as the precision and recall achieved by SVM1. Two factors are affecting

the accuracy for SVM2. First, SVM2 takes SVM1's output as its input; errors generated by SVM1 will affect SVM1's performance. Second, our ROI segmentation method achieves better performance for images under "far view" group than images under "mid range view" and "close-up view" groups. Since our SVM classifiers take the properties of ROI as input, failure in the ROI segmentation stage will cause the misclassification of the whole image.

Table 4-3 Precision and Recall

Keyword \	FW	FH	MW	IF	EF	OF
Truth	2552	729	796	639	185	849
Miss	155	118	114	0	0	0
False	132	66	84	67	38	0
Precision (%)	94.8	90.9	89.4	89.5	79.5	100
Recall (%)	94.2	86.1	87.5	100	100	100

Precision for "EF" is not high (**79.5%**) due to the errors incurred in ROI segmentation stage when many players stand together (Fig. 4-10). However, this drawback can be improved by using players' motion trajectory in ROI segmentation stage which is one of our future works.



Fig. 4-10 "MW" segment which is labeled as "EF" wrongly

Chapter 5

Audio Keyword Labeling

In Chapter 3, we define three audio keywords: “Non-excited”, “Excited” and “Very excited”. In this chapter, we will explain how we label the video segment with one of these three pre-defined audio keywords.

Fig. 5-1 shows our framework for audio keyword labeling. First, audio signal is partitioned into equal-length audio segments with 25% overlap between each two consecutive audio segments. Then, we use a twice-iterated Fourier Transform [41] to compute the excitement intensity of the audio signal within each audio segment. Next, for each video segment, we compute the average excitement intensity of the audio segment within that video segment. Then, that video segment is labeled with one of the three audio keywords “Non-excited”, “Excited” and “Very excited” according to the average excitement intensity.

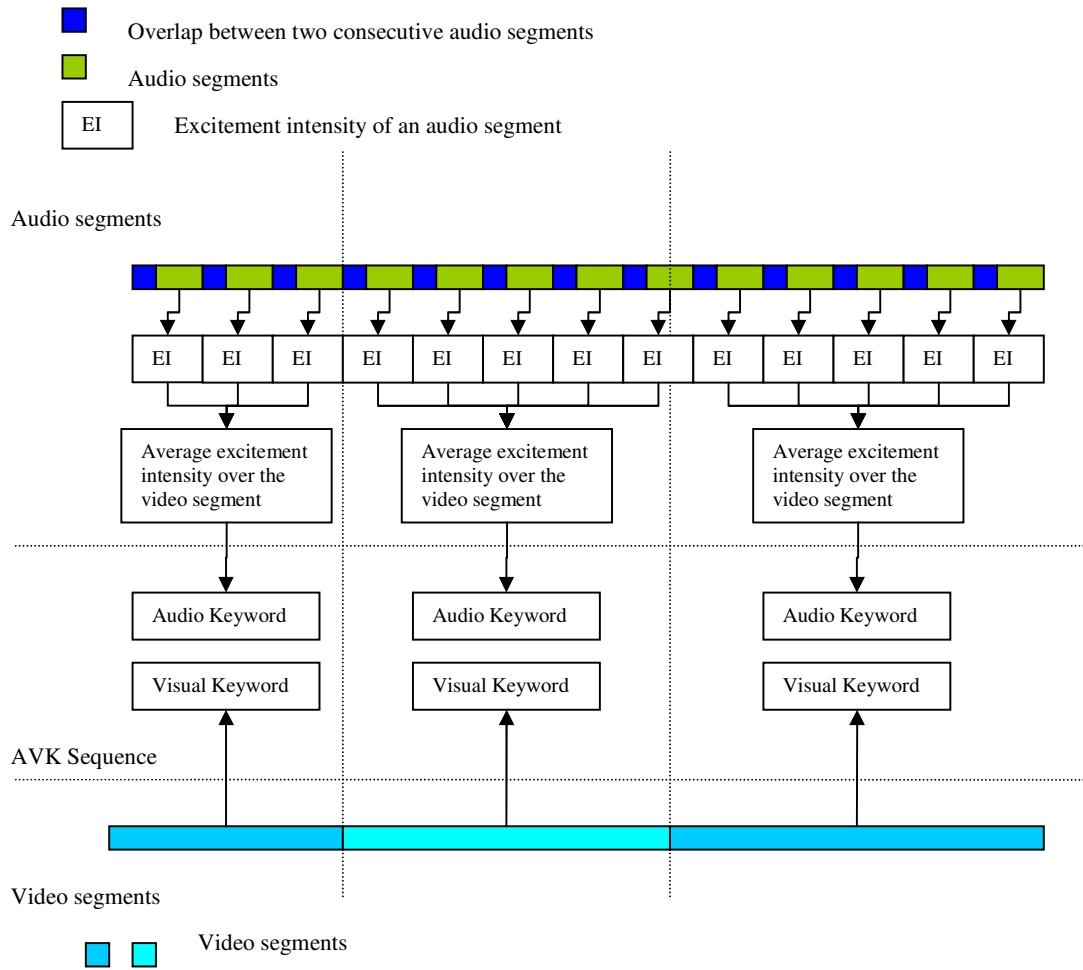


Fig. 5-1 Framework for audio keyword labeling

5.1 Feature Extraction

Among all the approaches known to us, the Double FFT method presented in [41] gives us a deep impression for its good performance in sports video domain. The authors apply a twice-iterated Fourier transform to the audio signal and computer the excitement intensity.

The principle of the Double FFT method is explained in [41] very clearly. Therefore we are not going to repeat the details of their work in this thesis. We just briefly describe the following 3 steps we used to compute the excitement intensity e_i :

1. The audio signal (44.1 KHz, 16bps, mono) is first extracted from video stream and is divided into small audio segments, 100ms each. To smooth the continuity, we let every two consecutive segments to have 25ms overlap. The audio segments are processed sequentially from the beginning to the end.
2. For each audio segment a_i , $f(n)$ $0 \leq n < 4,410$ is the audio signal within a_i . We use the following 3 sub-steps to computer e_i which is an integer we use to characterize the dominant speech within a_i .

- a. A first Fourier Transform is applied to $f(n)$

$$g(k) = \left| \frac{1}{N} \sum_{n=0}^{N-1} f(n) e^{-jk\omega_0 n} \right| \quad \text{Equ. 5-1}$$

$$\text{where } N = 4,410, \omega_0 = \frac{\pi}{N}, g(k) \in R$$

- b. We get a portion of $g(k)$ ranges from 100-400Hz denoted as $g'(n)$ ($0 \leq n < 300$).

Then, a second Fourier Transform is applied to $g'(n)$.

$$h(k) = \left| \frac{1}{N} \sum_{n=0}^{N-1} g'(n) e^{-jk\omega_0 n} \right| \quad \text{Equ. 5-2}$$

$$\text{where } N = 300, \omega_0 = \frac{\pi}{N}, h(k) \in R$$

- c. We compute the density of the peak in the $h(k)$ profile by counting the number of peaks in the $h(k)$ profile, noted as e_i .

3. The e_i sequence is normalized by linearly mapping the minimum and maximum elements to zero and one hundred respectively.

By comparing the 1D graph profile of e_i sequence and the audio signal, we find that the Double FFT method could characterize the dominant speech in soccer videos well. For the audio signal segments near the goal events, the relative e_i are very big. From our experiment (based on 3 hours of audio signal extracted from soccer videos), we find that most of the goal events are within the audio segments whose e_i values are among the top 10% in the whole e_i sequence. By changing the threshold from 10% to 15%, we can enclose all the goal events. Therefore, we regard the e_i as the excitement intensity within the relative audio segment a_i .

5.2 Audio Keyword Classification

So far, we have computed the excitement intensities for all the audio segments. Now, we need to calculate the average excitement intensity for a certain video segment.

Since there are 25 visual frames within one second, each visual frame lasts for $\frac{1000}{25} = 40ms$.

The audio segments all last for 100ms with 25ms overlap and the average length of the audio segment is $100-25=75ms$. Hence, given the number of the start and end frame of a video segment:

$Start_{Video}$ and End_{Video} , we can find the relative audio segments by:

$$Start_{audio} = \frac{Start_{video} \times 40}{75} \quad \text{Equ. 5-3}$$

$$End_{audio} = \frac{End_{video} \times 40}{75} \quad \text{Equ. 5-4}$$

Then, we can calculate the average excitement intensity within the video segment denoted as \tilde{e}_i with Equ. 5-5.

$$\tilde{e}_i = \frac{1}{End_{audio} - Start_{audio} + 1} \sum_{j=Start_{audio}}^{End_{audio}} e_j \quad \text{Equ. 5-5}$$

In the end, we quantize the \tilde{e}_i value to one of the three possible levels and classify the relative video segment into “NE”, “EX” or “VE” accordingly.

- “VE”: The video segments whose \tilde{e}_i values are among the top 10% in the whole \tilde{e}_i sequence.
- “EX”: The video segments whose \tilde{e}_i values are below the top 10%, above the top 15% in the whole \tilde{e}_i sequence.
- “NE”: Other video segments.

Chapter 6

Event Detection

In the first level of our system (Fig. 1.1), video streams are analyzed and visual / audio keyword sequences are computed. In the second level, we deal with event detection. In general, the probabilistic mapping between the AVK sequence and the events can be modeled either statistically (e.g. HMM) or syntactically (e.g. grammar). In this chapter, we detect the goal event with both approaches, namely grammar-based approach (Section 6.1) and HMM-based approach (Section 6.2). In section 6.3, we compare these two approaches by analyzing the advantages and disadvantages of them.








6.1 Grammar-Based Event Detector

In this section, we focus on utilizing a grammar-based approach [65] to detect corner-kick and goal events in soccer video based on visual keyword sequences that capture the temporal visual patterns of soccer games (as a result of broadcast production).

A set of detection rules called *event detection grammar* can be pre-defined or learned from the training data. These detection rules capture the semantic meaning of soccer events based on the visual keywords defined for soccer video. To detect the events of interest, we apply the event detection rules to the visual keyword sequence recursively to generate some *grammar trees*. We say that a visual keyword sequence contains a certain event if the visual keyword relevant to that event is spotted at either one of the roots of the grammar trees. By studying the position of the “head” visual keyword, the exact position of the segment that is the key to the event can be spotted.

6.1.1 Visual Keyword Definition

Table 6-1 Visual keywords used by grammar-based approach

Keywords	Abbreviation	Relative Segment Example
Far view of whole field	FW	
Far view of half field	FH	
Mid range view (whole body of player visible)	MW	
Close up view of multiple players	CS	
Close up view of single player	CP	
Goal Post	GP	
Audience	AD	

Our grammar-based approach is based on the visual keywords that are introduced in [66] and the definition of the visual keywords is similar but not exactly the same with the visual keywords we defined in Chapter 3. In Table 6-1, we list the visual keywords used by our grammar-based approach.

6.1.2 Event Detection Rules

Typically an event detection rule i is of the form:

$$\langle V_1 \rangle, \langle V_2 \rangle, \dots, \langle V_n \rangle \rightarrow \langle E_i \rangle$$

We say a detection rule is applicable to a certain visual keyword sequence if and only if the visual keyword sequence contains exactly the same consecutive visual keywords (or terms) $\langle V_1 \rangle, \langle V_2 \rangle, \dots, \langle V_n \rangle$. If a detection rule is applicable to a visual keyword sequence, we could apply the detection rule to the visual keyword sequence by removing the visual keywords (or teams) $\langle V_1 \rangle, \langle V_2 \rangle, \dots, \langle V_n \rangle$ and inserting a new term $\langle E_i \rangle$ which lies right to the arrow. Note that the left side can contain terms generated by other rules. Some rules have a “head” term (e.g. boldfaced V_1). The “head” term is very important as the segment associated with is the segment upon which the event anchors.

For example: given a detection rule $\langle FW \rangle \langle CL \rangle \rightarrow \langle OF \rangle$ and two visual keyword sequences $\langle MW \rangle \langle FW \rangle \langle CL \rangle \langle MW \rangle$ and $\langle MW \rangle \langle FW \rangle \langle MW \rangle \langle CL \rangle \langle MW \rangle$, the detection rule is applicable to the first visual keyword sequence. By applying the detection rule, a new visual keyword sequence will be generated as follow: $\langle MW \rangle \langle OF \rangle \langle MW \rangle$. The detection rule is not applicable to the second visual keyword sequence because the second visual keyword sequence does not have consecutive visual keywords $\langle FW \rangle \langle CL \rangle$.

By using or more detection rules, it is possible to provide AND-OR choices. For example, given detection rules:

1. $\langle FH \rangle \langle CL \rangle \rightarrow \langle OF \rangle$

2. $\langle \text{FW} \rangle \langle \text{CL} \rangle \rightarrow \langle \text{OF} \rangle$

3. $\langle \text{OF} \rangle \langle \text{PR} \rangle \rightarrow \langle \text{SF} \rangle$

The third rule means that: if and only if both $\langle \text{OF} \rangle$ and $\langle \text{PR} \rangle$ are generated, then, $\langle \text{SF} \rangle$ could be generated. The first two detection rules means: either $\langle \text{FH} \rangle \langle \text{CL} \rangle$ is spotted or $\langle \text{FW} \rangle \langle \text{CL} \rangle$ is spotted, $\langle \text{OF} \rangle$ could be generated.

In essence, an event detection rule captures the production intent of the cameraman such that the semantic meaning of the term on the right of the arrow is expressed as a sequence of visual keywords (or terms) on the left of the arrow. For example, one of the detection rules we use in corner-kick detection grammar is: $\langle \text{FH} \rangle \langle \text{CP} \rangle \rightarrow \langle \text{OF} \rangle$. In this rule, $\langle \text{FH} \rangle$ followed by $\langle \text{CP} \rangle$ means the focus of the camera moves from the whole field (both player and ball can be seen) to a certain player (only player can be seen) which indicates that the ball might go out of the field which is the semantic meaning we set to the term $\langle \text{OF} \rangle$ (i.e. out-of-field).

6.1.3 Event Parser

When an event detection grammar is applied to a visual keyword sequence, the event parser will apply all the applicable rules to the visual keyword sequence recursively and construct grammar trees.

The event parser is implemented with a recursion function. It starts from recursion layer 1 and at recursion layer N, it carries out the following steps:

1. If the pre-defined event is detected, then quit the function.
2. Backup the visual keyword sequence.
3. Find all the applicable rules.
4. If there are no applicable rules, then quit this layer to layer N-1.
5. Apply the 1st applicable rule to the visual keyword sequence.
6. Go to the start of layer N+1.

7. Wait until function quits from layer N+1.
8. If all the applicable rules have been applied, go to step 11.
9. Restore the visual keyword sequence to the state backuped at step 2 in this layer and then apply the next applicable rule to the sequence.
10. Go to step 6.
11. If this is layer 1, then quit the function.
Otherwise, quit this layer to layer N-1.

6.1.4 Event Detection Grammar

Although events like goal and corner-kick always anchor upon a single segment, we can hardly detect them only by studying the visual keyword labeled of that segment. The semantic meaning of the visual keywords labeled on the segments around that segment can also help us in locating the event. Hence we formulate sub-events that occur along with the event we are trying to detect and instead of detecting the event, we detect the sub-events first. Later, we use the “head” term to locate the exact position of the segment that the event occurs.

In practice, we define 2 detection grammars, one for corner-kick detection and the other one for goal detection.

6.1.4.1. Corner-kick Detection

Generally, the occurrence of a corner-kick <CK> event is accompanied with the following three sub-events:

1. <OF>: ball goes out of the field from the bottom line.
2. <PR>: one player runs to the corner while other players gather around the goal post.
3. <SB>: the player who runs to the corner serves the ball. It is within this sub-event that the corner-kick event anchors

To detect the corner-kick event, we define rules to detect these three sub-events respectively.

Then, we use the following rule to detect the corner-kick event:

$$\langle \text{OF} \rangle, \langle \text{PR} \rangle, \langle \text{SB} \rangle \rightarrow \langle \text{CK} \rangle.$$

To define the rules for $\langle \text{OF} \rangle$ sub-event, we studied several soccer games videos and found that $\langle \text{FH} \rangle$ followed by $\langle \text{CP} \rangle$ or $\langle \text{CS} \rangle$ always indicates that the ball goes out of the field. Sometimes, there is also short visual keyword sequence consists of $\langle \text{AU} \rangle$, $\langle \text{CP} \rangle$ followed by $\langle \text{FH} \rangle$, $\langle \text{CP} \rangle$ and it also indicates $\langle \text{OF} \rangle$. Hence, we have four rules for $\langle \text{OF} \rangle$ in our grammar which are presented at the first four lines in Table 6-2.

Similarly, we can define other rules for $\langle \text{PR} \rangle$ and $\langle \text{SB} \rangle$. In practice, we define 16 rules for corner kick detection as shown in Table 6-2. The last column in Table 6-2 (also in Table 6-3) is the number of times that the rule is used when the grammar is applied to the test data. Fig. 6-1 shows the grammar tree of a typical visual keyword sequence containing the corner-kick event. The boxes in blue rectangles are “head” terms.

Table 6-2 Grammar for corner-kick detection

ID	Rules	Rules	Frequency
1	$\langle \text{FH} \rangle \langle \text{CP} \rangle$	$\langle \text{OF} \rangle$	14
2	$\langle \text{FH} \rangle \langle \text{CS} \rangle$	$\langle \text{OF} \rangle$	2
3	$\langle \text{OF} \rangle \langle \text{AU} \rangle$	$\langle \text{OF} \rangle$	1
4	$\langle \text{OF} \rangle \langle \text{CP} \rangle$	$\langle \text{OF} \rangle$	3
5	$\langle \text{MW} \rangle \langle \text{FH} \rangle$	$\langle \text{PR} \rangle$	2
6	$\langle \text{FH} \rangle \langle \text{MW} \rangle$	$\langle \text{PR} \rangle$	1
7	$\langle \text{GP} \rangle \langle \text{CP} \rangle$	$\langle \text{PR} \rangle$	2
8	$\langle \text{FH} \rangle \langle \text{CS} \rangle$	$\langle \text{PR} \rangle$	1
9	$\langle \text{PR} \rangle \langle \text{CP} \rangle$	$\langle \text{PR} \rangle$	4
10	$\langle \text{PR} \rangle \langle \text{MW} \rangle$	$\langle \text{PR} \rangle$	24
11	$\langle \text{CP} \rangle \langle \text{PR} \rangle$	$\langle \text{PR} \rangle$	0
12	$\langle \text{CS} \rangle$	$\langle \text{PR} \rangle$	5
13	$\langle \text{MW} \rangle$	$\langle \text{PR} \rangle$	5
14	$\langle \text{FH} \rangle$	$\langle \text{SB} \rangle$	14
15	$\langle \text{OF} \rangle \langle \text{PR} \rangle \langle \text{SB} \rangle$	$\langle \text{CK} \rangle$	14
16	$\langle \text{OF} \rangle \langle \text{FW} \rangle \langle \text{PR} \rangle \langle \text{SB} \rangle$	$\langle \text{CK} \rangle$	2

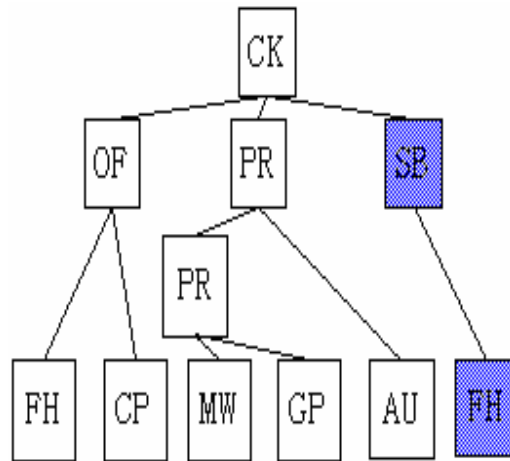


Fig. 6-1 Grammar tree for corner-kick

6.1.4.2. Goal Detection

We define 4 sub-events for the goal <GL> event:

1. <ST>: Player shoots and goals. Within this sub-event, goal event anchors.
2. <CR>: Players cheer
3. <SM>: Slow-motion replay
4. <SM>: Slow-motion again.

Table 6-3 Grammar for goal detection

ID	Rules	Rules	Frequency
1	<CP>	<CR>	4
2	<CS>	<CR>	1
3	<CR> <AU>	<CR>	2
4	<CR> <CP>	<CR>	14
5	<CR> <CS>	<CR>	4
6	<FH> <GP>	<SM>	2
7	<MW> <GP>	<SM>	3
8	<CS> <GP>	<SM>	0
9	<FW><MW><GP>	<SM>	0
10	<SM> <GP>	<SM>	2
11	<SM> <SM>	<SM2>	0
12	<FW> <SM>	<SM2>	0
13	<MW> <SM>	<SM2>	5
14	<FH>	<ST>	5
15	<ST><CR> <SM2>	<GL>	5

Besides these four sub-events, we also define <SM2> as two consecutive slow-motion replays. Fifteen rules have been defined and shown in Table 6-3. Figure 6-2 shows a grammar tree of keyword sequence that contains goal event.

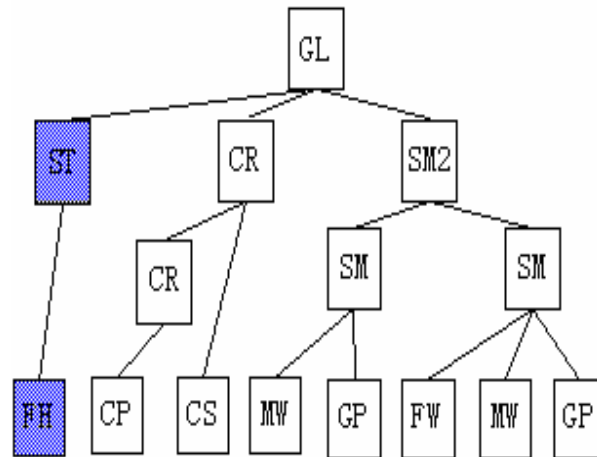


Fig.6-2 Grammar tree for goal

6.1.5 Experimental Results

The proposed framework has been implemented on a windows XP platform. The portions of 4 FIFA World Cup 2002 soccer matches (1666 segments) were used in the experiments to evaluate the system performance. There are 17 corner kicks and 7 goals in total. The boundaries of video segments and visual keywords are labeled manually. A grammar for goal detection and another grammar for corner-kick detection are applied to the visual keyword sequences. Table 6-4 and Table 6-5 list the number of the events, the number of events that are detected correctly and the number of false-alarms.

Table 6-4 Result for corner-kick detection

Soccer Video	Corner Kick	Correctly Detected	False Alarmed
BRA vs ENG	3	2	1
GER vs BRA	5	4	1
KOR vs TUR	5	4	0
GER vs USA	4	4	1
Total	17	14	3

Table 6-5 Result for goal detection

Soccer Video	Goal	Correctly Detected	False Alarmed
BRA vs ENG	2	2	0
GER vs BRA	0	0	0
KOR vs TUR	4	3	0
GER vs USA	1	1	0
Total	7	6	0

Out of the 17 corner-kick events, 14 of them are detected correctly while 3 of them are missed. Two of the three misses are due to the insertion of slow motion in the preparation sub-event. Another one is missed because only one visual keyword <CS> occurs in its out-of-field sub-event. This is exceptional.

For goal detection, there are 7 goal events while 6 of them are detected successfully. One event is missed because our grammar failed to detect its second slow-motion replay sub-event which is also exceptional.

6.2 HMM-based Event Detector

In this section, we focus on the use of statistical model for event detection [64]. More precisely, Hidden Markov Models (HMM) [52] are applied to AVK sequences in order to detect the goal event automatically.

We use HMM as our classifier here because HMM is widely used in news video and sports video domain and has shown to be a very effective approach for content-based video analysis [36,54,55]. One limitation of HMM is that HMM support Gaussian Mixture Model only while modeling observation probabilities. However, this limitation is alleviated to some extent by applying SVM to the features extracted from video streams in the first level of our system.

The AVK sequences that follow the goal events share similar AVK pattern which makes it possible for us to use HMM to model the temporal shot transition patterns automatically so that detection rules are no longer required to be defined in advance. Generally, after the goal, the game will pause for a while (around 30-60 seconds). During that break period, the camera first zooms into the players to capture their emotions and people cheer for the goal. Next, two to three slow motion replays are presented to show the actions of the goalkeeper and shooter to the audience again. Then, the focus of the camera might go back to the field to show the exciting emotion of the players again for several seconds. In the end, the game resumes. Fig. 6-3 shows an example of the typical pattern followed the goal event.



Fig. 6-3 Special pattern that follows the goal event

Instead of sending all the video portions to HMM, we first extract the some candidate portions (exciting break portions) from the AVK sequence. Then, two HMMs are applied to the candidates to model the exciting break portions with goal pattern and without goal pattern respectively.

In this section, we will first describe how we extract exciting break portions from the AVK sequences. Next, we introduce how we combine the AVK together and map them into a 13-dimension feature vector. Then, we will describe how we use two HMMs to model the goal pattern. In the end, we will report our experimental results.

6.2.1 Exciting Break Portion Extraction

In Chapter 3, we have briefly described the semantic meaning of the static visual keywords. Generally, long “far view” segment always indicates that the game is in play and short “far view” segment is sometimes used during a break. Hence, we extract play portions by detecting four or more consecutive “far view” video segments. For break portions, we scan the static visual keyword sequence from the beginning to the end sequentially. When we spot a “far view” segment, a portion that starts from the first non-“far view” segment thereafter ends at the start of the next play portion is extracted and regarded as a break portion. (Fig. 6.4)

After break portions extraction, we use audio keyword to further extract exciting break portions. For each break portion, we compute the number of “EX” and “VE” keywords that are labeled to it, denoted as EX_{num} and VE_{num} . The excitement intensity and excitement intensity ratio of this break portion is computed as:

$$Excitement = 2 \times VE_{num} + EX_{num} \quad \text{Equ. 6-1}$$

$$Ratio = \frac{Excitement}{Length} \quad \text{Equ. 6-2}$$

where *Length* is the number of the video segments within the break portion

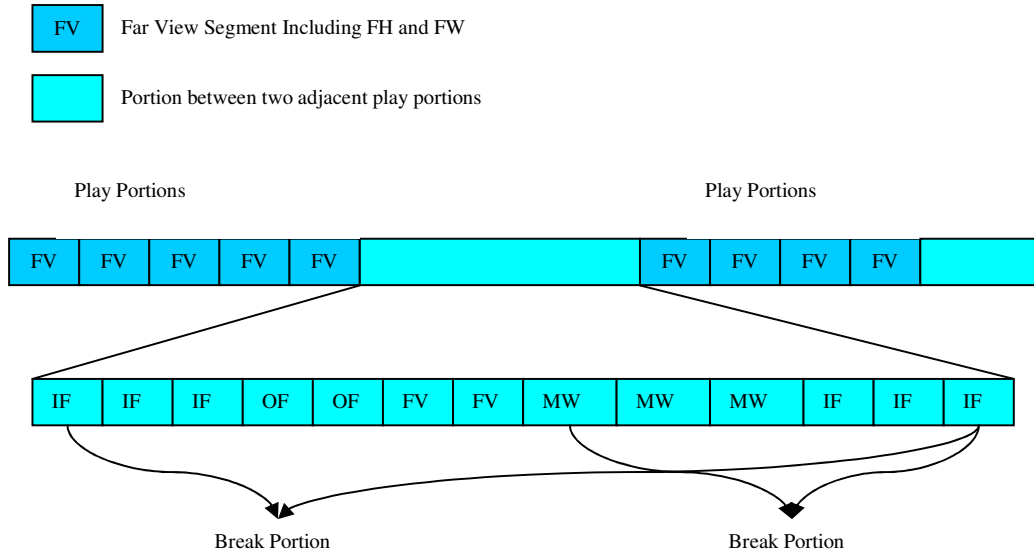


Fig. 6-4 Break portions extraction

By setting thresholds for excitement intensity ratio (T_{Ratio}) and excitement intensity ($T_{Excitement}$) respectively, we extract the exciting break portions. Our experiments are based on two sets of thresholds, they are set to be (0.4 and 9) and (0.3 and 7) respectively.

6.2.2 Feature Vector

Generally, two kinds of methods are used to combine the visual and audio features:

1. Before the classification, visual and audio features are fused together into one single audiovisual feature vector.
2. Visual and audio features are treated as two different features and classified separately first. Then, the classification results are combined into final classification decision.

For our case, we decide to use the first method for our system. If we need to add more keyword sequence in the future such as text caption keywords, etc, new information can be easily

incorporated into our system simply by expanding the dimensions of the feature space with the new keyword sequence.

For each video segment, we label one static visual keyword, one dynamic visual keyword and one audio keyword. Including the length of the video segment, we use a 13-dimensions feature vector to represent one video segment.

We have defined 12 AVKs in total and the first 12-dimensions correspond to the 12 AVKs. Given a video segment, only the dimensions that correspond to the AVKs labeled to the video segment are set to one and, other dimensions are all set to zero. The last dimension is used to describe the length of the video segment by a number between zero and one which is the normalized version of the number of the frames of the video segment.

6.2.3 Goal and Non-Goal HMM

Hidden Markov Model is a powerful tool for analyzing the sequential data. It has been applied to many sports video research work with significant success.

Several preliminary experiments are performed to decide the best topology for the HMM. By varying the number of hidden states in the HMM from four to seven, we find out that five-state HMM gives the best result. Since the sports video is always processed according to the temporal order, we limit the model of HMM to left-to-right. Hence, we use two five-state left-right HMMs to model the exciting break portions with goal event (*goal model*) and without goal event (*non-goal model*) respectively (Fig. 6-5). We denote *goal model likelihood* with G and *non-goal model likelihood* with N hereafter. Observations send to HMMs are modeled as single Gaussians.

In practice, HTK [68] is used for HMM modeling. The initial values of the parameters of the HMMs are estimated by repeatedly using Viterbi alignment [69] to segment the training observations and then recomputing the parameters by pooling the vectors in each segment. Then, Baum-Welch algorithm [70] is used to re-estimate the parameters of the HMMs. For each

exciting break portion, we evaluate its feature vector likelihood under both two HMMs and we say the goal event is spotted within this exciting break portion if its G is bigger than its N .

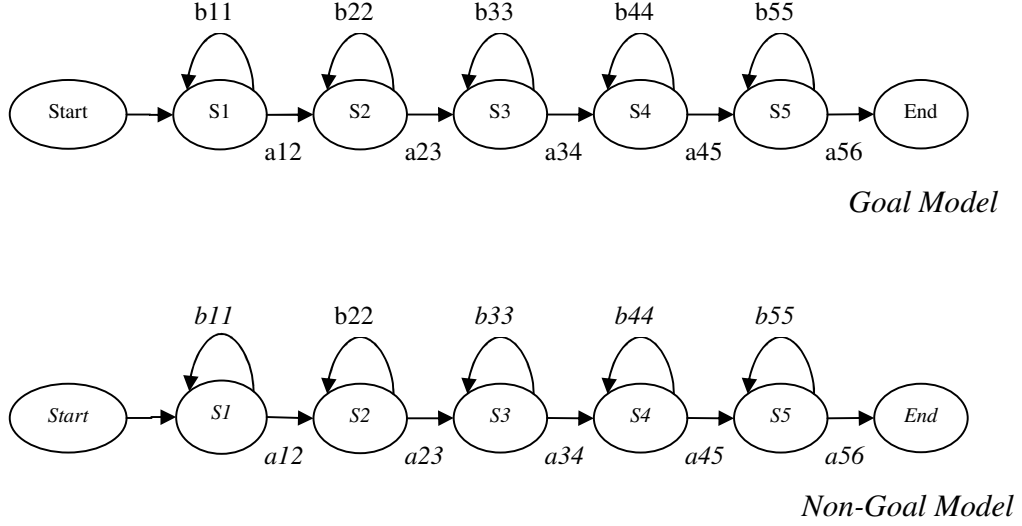


Fig. 6-5 Goal and non-goal HMMs

6.2.4 Experimental Results

In order to train the HMM models for goal and non-goal events, we need large amount of training sets. Unfortunately, in soccer videos, goal happens rarely. We have studied the soccer videos of ten half-matches. On average, there are only 2.1 goals for each half-match. The biased amount of training data will affect the accuracy. Hence, in this thesis, we use cross validation to remedy this problem in some extent.

Ten half-matches of the soccer video (459 minutes, 21 goals) from FIFA 2002 and UEFA 2002 are used in our experiment. The soccer videos are all in MPEG-1 format, 352×288 pixels, 25 frames/second. AVK sequences of seven half-matches are labeled automatically while the other three AVK sequences are labeled manually.(please refer to section 4.4 for the precision and recall). For the purpose of cross validation, for each one of the ten AVK sequences, we use the other nine AVK sequences as training data to detect goal from this AVK sequence. Exciting

break portions are extracted from all the ten AVK sequences automatically using two different sets of threshold settings. The lowest excitement intensity ratio and excitement intensity of the goal portions in our experimental data is 0.6 and 12 respectively. In practice, the thresholds of T_{Ratio} and $T_{Excitement}$ are set to be (0.4 and 9) and (0.3 and 7) respectively and the experimental results are listed in Table 6.6 and Table 6.7.

Table 6-6 Result for goal detection ($T_{Ratio}=0.4$, $T_{Excitement}=9$)

Soccer Video	Goal	Correctly Detected	Miss	False Alarm	Precision	Recall
GER vs ENG	3	3	0	0	100%	100%
LEV vs LIV	4	4	0	0	100%	100%
LIV vs LEV	1	1	0	1	50%	100%
SEN vs TUR	0	0	0	1	-----	-----
BRA vs GER	2	1	1	0	100%	50%
USA vs GER	1	1	0	1	50%	100%
KOR vs TUR	4	3	1	0	100%	75%
ARS vs LEV	2	2	0	0	100%	100%
ENG vs BRA	2	2	0	0	100%	100%
REA vs BAY	2	2	0	1	66.7%	100%
Total	21	19	2	4	82.6%	90.5%

Table 6-7 Result for goal detection ($T_{Ratio}=0.3$, $T_{Excitement}=7$)

Soccer Video	Goal	Correctly Detected	Miss	False Alarm	Precision	Recall
GER vs ENG	3	3	0	1	75%	100%
LEV vs LIV	4	4	0	2	66.7%	100%
LIV vs LEV	1	1	0	2	33.3%	100%
SEN vs TUR	0	0	0	2	-----	-----
BRA vs GER	2	1	1	1	50%	50%
USA vs GER	1	1	0	1	50%	100%
KOR vs TUR	4	3	1	0	100%	75%
ARS vs LEV	2	2	0	1	66.7%	100%
ENG vs BRA	2	2	0	1	66.7%	100%
REA vs BAY	2	1	1	2	33.3%	50%
Total	21	18	3	13	58.0%	85.7%

When we set T_{Ratio} and $T_{Excitement}$ to appropriate values, most of the exciting break portions extracted from the AVK sequence are the corner-kick, free-kick, etc. These portions share similar structure patterns. If T_{Ratio} and $T_{Excitement}$ are set to too low, some portions with different structure patterns might also be extracted which will bring noises to the HMM and lower the system performance.

From our experiments, we can see that our approach achieves 82.6% precision and 90.5% recall when T_{Ratio} and $T_{Excitement}$ is set to be 0.4 and 8 respectively. When T_{Ratio} and $T_{Excitement}$ is relaxed to 0.3 and 7, the precision and recall degenerate. Since, generally, the excitement intensity ratio and excitement intensity is higher than 0.6 and 12, we think 0.4 and 9 is a reasonable setting for goal event detection.

By investigating the misclassified exciting break portions, we find that there are mainly two factors that are affecting the accuracy. First, the “shot and miss” shares similar temporal patterns

of video segments with “goal” which lower the precision of system performance. Second, the accuracy of AVK sequence is also affecting the system accuracy.

6.3 Discussion

The two approaches we proposed in this chapter have different strengths and weaknesses. Although the definitions of AVK of these two approaches are not exactly the same, their basic ideas are the same which are to detect event based on AVK sequence. In this section, we compare these two approaches in the following three aspects:

6.3.1 Effectiveness

The grammar-based approach is very effective. It can detect the goal and corner-kick even without using any audio keywords. But for HMM-based approach, based on the definition of current AVK, we can only detect goal event with good precision and recall but cannot achieve good result in corner-kick event detection. The main reason is that corner-kick and free-kick does not always follow the unique shot transition pattern while most of the goals do. Hence, unlike goal detection, one HMM might not be enough for modeling all the “corner-kick” segments. Two approaches might be useful to resolve this problem: first, use more than one HMMs to model the shot transition of “corner-kick” segments; second, define more audio keywords in our audio keywords vocabulary or include more types of keywords such as text caption keywords, etc.

6.3.2 Robustness

Although the grammar-based approach is very effective, unfortunately, it requires very high accuracy in the first level of our system. In our experiment, we use manually labeled data which is based on the assumption that the first level of our system could achieve 100% accuracy on all the visual keywords defined in our vocabulary which is hardly possible. An error in the first level

such as misclassification of one shot might result in the failure of the whole grammar tree construction. For example: if the first visual keyword of the visual keyword sequence “FH CP CS MW GP FW MW GP” are misclassified into “FW” then the whole keyword sequence becomes “FW CP CS MW GP FW MW GP”. Using the goal event detection grammar we proposed in section 6.1, this goal event will be missed.

Some modifications could be made to our grammar-based approach such as using stochastic grammar parsing. In that case, “FW” followed by “CP” could also generate “OF” like “FH” followed by “CP” but with lower probability. This could improve the robustness of the grammar-based approach.

By comparison, the HMM-based approach is much more robust. One misclassified shot just lowers the probability but it does not necessarily cause the failure of the detection of an event.

6.3.3 Automation

Another aspect we wish to make a comparison here is the automation of these two approaches. For the grammar-based approach, up to now, we cannot get the detection rules automatically. We have to study the soccer videos carefully and define the detection rules manually. The detection rules might require modification in order to adapt to different shooting or attack strategies.

On the other hand, there is no such concern for the HMM-based approach. The two HMM models are trained automatically from the training data. It can be adapted to other shooting or attack strategies simply by re-training the HMM models with new AVK sequences.

Chapter 7

Conclusion and Future Work

7.1 Contribution

In this thesis, we have proposed a multi-modal two-level framework that uses Audio and Visual Keywords (AVKs) to extract break portions and detect goals from soccer video. The notion of AVK indeed facilitates the use of a unique syntactic parsing technique and HMM-based statistical approach to detect events in soccer video. We summarize our contributions as follow:

- (1) We propose a multi-modal two-level framework for sports video event detection and use soccer video as our test bed [67]. Our system successfully fuses the semantic meaning of AVKs by applying HMM in the second-level to the AVKs which are well aligned to the video segments.

- (2) We propose a promising approach for static visual keyword labeling by using ROIs and SVM [62]. When applying our approach to ten half-matches, our system achieves 90.7% precision and 89.6% recall on average.
- (3) Both statistical and syntactic approaches are used in the second level to compare the difference between these two kinds of approaches. Strengths and weaknesses of these two approaches are analyzed.

7.2 Future Work

Event detection is a challenging research topic with great application potential. We have embarked on this research direction and achieved very promising results. However, there is still much room for improvement. In particular, we would like to pursue the following directions not completed in this thesis due to time constraints:

- (1) Apply our framework to event detection in other domains. In this thesis, we concentrate on goal detection in soccer videos. We would like to apply our framework to the detection of other interesting events in other sports videos such as touchdown in football videos, etc.
- (2) Combine the advantages of statistical approach and syntactical approach. Fig. 7.1 shows the relation between these two approaches. By discovering the rules from HMM model, we can explore the relation between the grammar-based approach and the rules discovered. We will also find the relation between the stochastic grammar-based approach and the HMM-based approach. Then, we would like to find an approach to detect the events as effective as the grammar-based approach and as robust as the HMM-based approach.

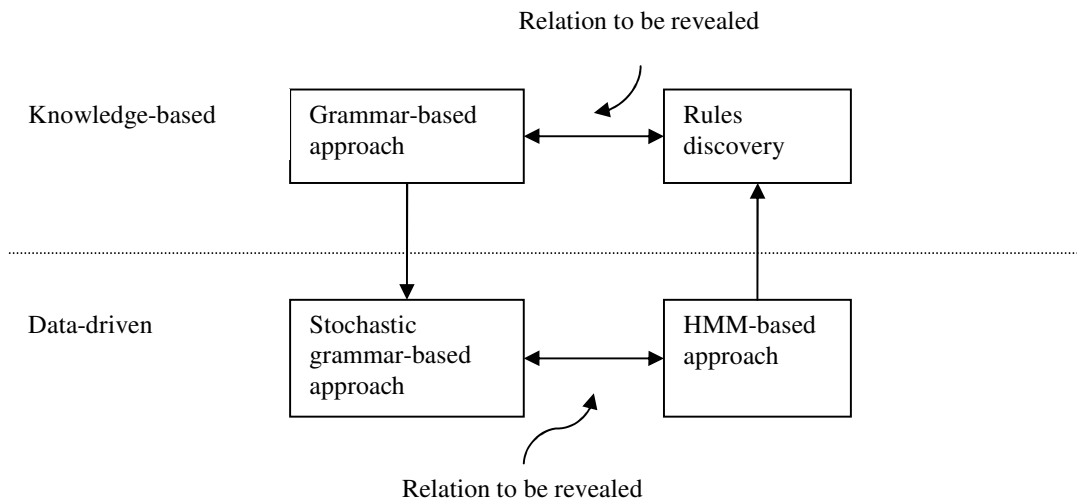


Fig. 7-1 Relation between syntactical approach and statistical approach

- (3) Tolerate uncertainty in the AVKs. So far, we label the video segments with one certain static visual keyword, one certain dynamic visual keyword and one certain audio keyword. We would like to allow uncertainty in the AVK labeling stage. For example: a video segment might be 70% similar to “far view” and 30% similar to “mid range view”.
- (4) Integrate the video segmentation and recognition of AVK into one stage. So far, our system first segment the video streams into video segments. Then, we label each video segment with AVK. We would like to integrate these two steps into one step by HMM.

References

- [1] S. Eickeler and S. Muller. "Content-based video indexing of TV broadcast news using hidden markov models." *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 6. 2997-3000, 1999.
- [2] S. Tsekeridou and I. Pitas. "Content-based video parsing and indexing based on audio-visual interaction.", *IEEE Transaction on Circuits and Systems for Video Technology*. 11(4):522-535, 2000
- [3] S. W. Smoliar and H. Zhang. "Content-based video indexing and retrieval". *IEEE Multimedia*, 1(2): 62-75, 1994.
- [4] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic indexing and retrieval in video", *Proceedings of IEEE International Conference on Multimedia and Expo*, volume. 1 pp. 475-478, New York, NY, July 2000.
- [5] M.R. Naphade and T. S. Huang, "Semantic video indexing using a probabilistic framework." *Proceedings of the ICPR 2002*, volume 2, pages 1005-1008. August 2002.
- [6] M.A. Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization", *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1997.
- [7] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding techniques." *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 775-781, Feb. 1997.
- [8] S.-F. Chang, "Optimal video adaptation and skimming using a utility-based framework", *Tyrrhenian International Workshop on Digital Communications (IWDC-2002)*, Capri Island, Italy, Sept. 2002.

- [9] N. V. Patel and I. K. Sethi. "Video shot detection and characterization for video databases". *Pattern Recognition*, 30(4):583 – 592, April 1997.
- [10] D. Zhang, W. Qi, and H.J. Zhang. "A new shot boundary detection algorithm". *Proceedings of 2nd IEEE Pacific-Rim Conference on Multimedia (PCM 2001)*, pages 64–70, Beijing, China, 2001.
- [11] A. M. Ferman and A. M. Tekalp, "A fuzzy framework for unsupervised video content characterization and shot classification," *Journal of Electronic Imaging*, volume 10, no. 4, pp. 917-929, Oct. 2001.
- [12] R. Hammound and R. Mohr. "A probabilistic framework of selecting effective key-frames for video browsing and indexing." *International workshop on Real-Time Image Sequence Analysis*, pages 79-88, August 2000.
- [13] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra. "Adaptive key frame extraction using unsupervised clustering." *Proceedings of IEEE Conference on Image Processing*, pp. 866-870, Chicago, IL, October 1998.
- [14] JungHwan Oh, Kien A. Hua and Ning Liang. "A content-based scene change detection and classification technique using background tracking." *Proceedings of SPIE Conference Multimedia Computing and Networking*, pp. 254-265, Jan 2000
- [15] M.J. Gauch, S. Gauch, S. Bouix and X.L. Zhu, "Real time video scene detection and classification", *Information Processing and Management* 35 , pp 401-420, 1999
- [16] Y. Nakamura and T Kanade, "Semantic analysis for video contents extraction –spotting by association in news video". *Proceedings of ACM International Multimedia Conference*, pp. 393-401, Seattle, 1997
- [17] X. Shao, C.S. Xu and M.S. Kankanhalli, "Automatically generating summaries for musical video", *Proceedings of IEEE International Conference of Image Processing (ICIP03)*, Barcelona, Spain, 2003

- [18] L. Chaisorn, T.S. Chua and C.H. Lee. "The segmentation of news video into story units". *Proceedings of IEEE International Conference on Multimedia and Expo*, Lausanne, Switzerland. Aug 2002. 26-29
- [19] L. Chaisorn, T.S. Chua and C.H. Lee. "A multi-modal approach to story segmentation for news video". *Journal of World Wide Web. Kluwer Academic Publishers*. 6(3), 187-208, Jun 2003
- [20] Y. Chen and E.K. Wong "A knowledge-based approach to video content classification", *Proceedings of SPIE (Storage and Retrieval for Media Databases)* Vol. 4315, pp. 292-300, Jan 2000
- [21] L. Chaisorn, T.S. Chua, C.K. Koh, Y.L. Zhao, H.X Xu, X.M. Feng and Q. Tian. "A two-level multi-modal approach for story segmentation of large news video corpus". TRECVID Workshop 2003, Gaithersburg, USA, November 2003
- [22] A. Amir, M. Berg, S.F. Chang, W. Hsu, G. Iyengar, C.Y. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, J.R. Smith, B. Tseng, Y. Wu and D.Q. Zhang "IBM research TRECVID-2003 video retrieval system". TRECVID Workshop 2003, Gaithersburg, USA, November 2003
- [23] J.Assfalg, M.Bertini, C. Colombo, A. Del Bimbo, W. Nunziati. "Automatic extraction and annotation of soccer video highlights", *Proceedings of IEEE International Conference on Image Processing*, volume 2, pp. 527-530, 2003
- [24] A. Ekin and A. M. Tekalp, "Generic event detection in sports video using cinematic features", *2nd IEEE Workshop on Event Mining: Detection and Recognition of Events in Video*, pp.34 June 2003
- [25] L.Y. Duan, M. Xu, TS. Chua, Q. Tian, C.S. Xu, "A Mid-level Representation Framework for Semantic Sports Video Analysis", *ACM Multimedia*, 2003.
- [26] V. Tovinkere, R. J. Qian, "Detecting semantic events in soccer games: toward a complete solution", *Proceedings of IEEE International Conference on Multimedia and Expro*, pp. 1040-1043, Tokyo, Japan, 2001

- [27] H. Hashimoto A. Akutsu, Y. Tonomura and Y. Ohba. "Video indexing using motion vectors". *Proceedings of SPIE Visual Communication and Image Processing*, pp. 1522-1530, 1992.
- [28] J. Park, N. Yagi, K. Enami, and E.Kiyoharu. "Estimation of camera parameters from image sequence for model video based coding". *IEEE Transactions on Circuits Systems and Video Technology*, 3(4):288-296, 1994
- [29] M. Srinivasan, S. Venkatesh, and R. Hosie. "Qualitative estimation of camera motion parameters from video sequences". *Pattern Recognition*, 30:593-606, 1997
- [30] A. Bonzanini, R. Leonardi, P. Migliorati, "Semantic video indexing using MPEG motion vectors", *Proceedings of EUSIPCO'2000*, pp. 147-150, Tampere, Finland, Sept. 2000.
- [31] A. Bonzanini, R. Leonardi, P. Migliorati, "Event recognition in sport programs using low-level motion indices", *Proceedings of IEEE International Conference on Multimedia and Expro*, pp. 920-923, Tokyo, Japan, 2001
- [32] C.L. Huang, C.Y. Chang. "Video summarization using Hidden Markov Model". *Proceedings of International Conference on Information Technology: Coding and Computing*, pp. 473-477, 2001.
- [33] H. Pan, P. Beek and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation", *Proceedings of International Conference on Acoustics, Speech and Signal processing (ICASSP)*, volume 3, pp. 1649-1652, 2001.
- [34] M. Lazarescu, S. Venkatesh and G. West, "On the automatic indexing of cricket using camera motion parameters", *Proceedings of IEEE International Conference on Multimedia and Expro*, volume 1, pp. 809-812, 2002.

- [35] D. Zhong, S. F. Chang, "Structure analysis of sports video using domain models", *Proceedings of IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, pp. 920-923, Aug. 22-25, 2001.
- [36] L. Xie, S.F. Chang, A. Divakaran, H. Sun, "Structure analysis of soccer video with Hidden Markov Models", *Proceedings of International Conference on Acoustics, Speech and Signal processing (ICASSP)*, volume 4, pp. 4096-4099, Orlando, FL, USA, 2002.
- [37] Z. Xiong, R. Radhakrishnan, A. Divakaran and T. Huang, "Audio events extraction based highlights extraction from baseball, golf and soccer games in a unified framework" *Proceedings of International Conference on Acoustics, Speech and Signal processing (ICASSP)*, volume 3, pp. 401-404, Hong Kong, April 2003.
- [38] Y. L. Chang, W. Zeng, I. Kamel, and R. Alonso, "Integrated image and speech analysis for content-based video indexing," *Proceedings of International Conference on Multimedia Computing and Systems*, pp.306–313, 1996.
- [39] Y. Rui, A. Gupta, and A. Acero. "Automatically extracting highlights for TV baseball programs". *Proceedings of ACM International Conference on Multimedia*, pp. 105-115, USA, 2000.
- [40] Min Xu, Namunu C. Maddage, Changsheng Xu, Mohan Kankanhalli, Qi Tian, "Creating audio keywords for event detection in soccer video", *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003
- [41] K. Wan, N. Maddage, C.S. Xu, "Characterization of dominant speech for automatic sports highlight generation" *submitted to Proceedings of International Conference on Acoustics, Speech and Signal processing (ICASSP)*, 2004.

- [42] N. Babaguchi, Y. Kawai and T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration", *IEEE Transactions on Multimedia*, volume 4, pp. 68-75, March, 2002.
- [43] D.Q. Zhang, R. Raj, S.-F Chang, "General and domain-specific techniques for detecting and recognizing superimposed text in video", *Proceedings of IEEE International Conference on Image Processing*, volume 1, pp. 593-596, 2002.
- [44] Y. H. Gong, L. T. Sin, C. H. Chuan, H. J. Zhang and M. Sakauchi, "Automatic parsing of TV soccer programs", *Proceedings of International Conference on Multimedia Computing and Systems*, pp. 167-174, 1995.
- [45] D. Yow, B. Yeo, M. Yeung and B. Liu, "Analysis and presentation of soccer highlights from digital video". *Proceedings of Second Asian Conference on Computer Vision*, pp. 499-503, 1995
- [46] X. Yu, C.S. Xu, Q. Tian and H. Leong, "A ball tracking framework for broadcast soccer video", *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003
- [47] X.G. Yu, Q. Tian and K. Wan, "A novel ball detection framework for real soccer video", *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003
- [48] W. Zhou, A. Vellaikal and C. Jay Kuo, "Rule-based video classification system for basketball video indexing". *Proceedings of ACM multimedia workshops*, pp. 213-216, 2000
- [49] A. Hanjalic, "Generic approach to highlights extraction from a sport video", *Proceedings of IEEE International Conference of Image Processing (ICIP03)*, Barcelona, Spain, 2003
- [50] P. Xu, L. Xie, S. F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and systems for segmentation and structure analysis in soccer video", *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 721-724, Tokyo, Japan, Aug. 22-25, 2001.
- [51] A.A. Alatan, A.N. Akansu, and W. Wolf, "Multi-modal dialogue scene detecting using hidden markov models for content-based multimedia indexing", *Multimedia Tools and Applications* 14(2): 137-151, 2001.

- [52] J.Huang, Z. Liu, Y. Wang, Y. Chen and E. K. Wong, "Integration of multimodal features for video scene classification based on HMM", 2001
- [53] L. Xie, S.F. Chang, A. Divakaran, H. Sun, "Structure analysis of soccer video with Hidden Markov Models", *Proceedings of International Conference on Acoustics, Speech and Signal processing (ICASSP)*, volume 4, pp. 4096-4099, Orlando, FL, USA, 2002.
- [54] X. Gibert, H. Li and D. Doermann. "Sports video classification using HMMs". *Proceedings of IEEE International Conference on Multimedia and Expo*, volume 2, pp. 345-348, 2003.
- [55] G. Xu, Y.F. Ma, H.J. Zhang, S. Yang. "A HMM based semantic analysis framework for sports game event detection." *Proceedings of International Conference of Image Processing*, volume 1, pp. 25-28, 2003.
- [56] Z. Xiong, R. Radhakrishnan and A. Divakaran, "Generation of sports highlights using motion activity in combination with a common audio feature extraction framework". *IEEE International Conference on Image Processing*, 2003
- [57] R. Leonaidi, P. Migliorati, M. Prandini, "Semantic indexing of sports program sequences by audio-visual analysis", *Proceedings of International Conference of Image Processing*, volume 1, pp. 9-12, 2003
- [58] N. Babaguchi and N. Nitta, "Intermodal collaboration: a strategy for semantic content analysis for broadcasted sports video". *Proceedings of IEEE International Conference of Image Processing (ICIP03)*, Barcelona, Spain, 2003
- [59] J.H. Lim (2001). Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications*, 4(2/3): 125-139.
- [60] J.H. Lim, Q. Tian, & Mulhem, P. (2003). Home photo content modeling for personalized event-based retrieval. *IEEE Multimedia*, 10(4): 28-37

- [61] Yi Lin, Mohan S Kankanhalli, and Tat-Seng Chua, "Temporal multi-resolution analysis for video segmentation", *Proceedings of SPIE (Storage and Retrieval for Media Databases)*. vol 3972, pp. 494-505, Jan 2000
- [62] Y.L. Kang, J.H. Lim, Q. Tian, M.S. Kankanhalli and C.S. Xu. "Visual keywords labeling in soccer video". *Proceedings of IEEE International Conference on Pattern Recognition*, 2004.
- [63] R. Gonzalez and R. Woods, "Digital Image Processing", Addison Wesley, 1992, pp 414-428
- [64] T. Joachims "Making large-Scale SVM Learning Practical", "Advances in Kernel Methods - Support Vector Learning", B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- [65] Y.L. Kang, J.H. Lim, Q. Tian and M.S. Kankanhalli. "Soccer video event detection with visual keywords". *IEEE Pacific-Rim Conference on Multimedia*, 2003.
- [66] J.H. Lim, H.P. Sun, Q. Tian and M.S. Kankanhalli. "Semantic labeling of soccer video". *IEEE Pacific-Rim Conference on Multimedia*, 2003.
- [67] Y.L. Kang, J.H. Lim, M.S. Kankanhalli, C.S. Xu and Q. Tian. "Goal detection in soccer video using audio/video keywords". *Proceedings of IEEE International Conference on Image Processing*, 2004
- [68] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "The HTK book" version 3.2, CUED, Speech Group, 2002
- [69] G.D.Forney. "The viterbi alignment", *Proceedings of IEEE*, vol 61, no. 3, pp. 263-278, March, 1973
- [70] L.R. Rabiner, B.H. Juang, "An introduction to Hidden Markov Models", *IEEE ASSP Magazine*, vol. 3, February 1986