Semantic Soccer Video Analysis

SUN HAIPING

(B.Eng., Beijing University of Technology, China)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE
INSTITUTE FOR INFOCOMM RESEARCH
SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE

2004

Acknowledgements

First of all, I would like to express my sincerest gratitude to my supervisors, Dr. Qi Tian, Dr. Joo-Hwee Lim, and Dr. Mohan S. Kankanhalli, for their invaluable guidance and suggestions during my two years of study. It is wonderful to study under their supervision.

Although being very busy everyday, Dr. Qi Tian opened a door for research work in front of me and continually advised and encouraged me. Without his help, I would never finish my research work and this thesis. Dr. Joo-Hwee Lim spent a lot of precious time discussing with me about my research work and providing much precious advice. His rigorous work attitude, wide knowledge and sagacious vision have influenced me deeply and will benefit me continually throughout my career. As my co-supervisor in SOC, Dr. Mohan provided many practical suggestions to help me overcome many difficulties in my research work.

I am very grateful to Institute for Infocomm Research for providing me such an excellent environment. Also, I would like to thank National University of Singapore for the financial support. Without this generous assistance, it is hard for me to finish my study here.

Finally, I would like to thank my parents, my sister and brother-in-low, and, my wife for their constant support and encouragement. I will love you forever.

Table of Contents

List of Figures

List of Tables

Summary

Soccer video analysis is concerned with the extraction of valuable semantics by efficient and effective processing of combination of visual, audio and text information. However, one of the major limitations of current soccer analysis is the semantic gap between the low-level features and mid-level representation.

This thesis proposes such a solution that targets at bridging the semantic gap and building an innovative intermediate representation of high- and low-level video information to aid in indexing, retrieval, and browsing. This solution is based on an understanding of broadcast soccer video.

Upon the study of soccer video structure, we found that for the purpose of semantic description, shot is not suitable as a mid-level representation (e.g. too long to be delineated by a semantic word). This means video analysis on a shot basis could not fully use all the essential information contained in soccer videos, which will result in the limitation in further analysis such as event detection and summarization. Instead, we introduce a structural-semantic video representation for efficient description of low-level video features. Firstly, we define 7 categories for soccer video classification, and seven ***Semantic Descriptors*** (close-up view, audience, far view of whole field, far view of penalty box, goal post in close-up view, player/players and mid-range view) are associated with them to delineate their semantic meanings. Therefore, a soccer video stream can be divided into segments, each of which belongs to one of these 7 categories. Or, this video stream can be delineated by a semantic descriptor sequence. This is our proposed mid-level representation of the soccer game.

In order to achieve this mid-level representation, a computational framework is proposed and two approaches are adopted to realize this framework. One approach adopting less domain knowledge is designed to explore a generic method which can be used to analyze other types of sports video; another one uses much more domain knowledge to provide an effective analysis for only soccer video. They shared the same pre-processing and post-processing stages but they are different in the processing stage. In the pre-processing stage, which is designed to reduce the computational complexity, motion magnitude is used to preliminarily segment a video stream into relatively static parts and active parts. Motion features in the static parts are ignored, and these static parts will be processed again in the post-processing stage. Segmentation and classification are done to these active parts in the processing stage. In practice, each P frame is divided into a 4 by 6 gird, each of which is called a ***block***; proposed analyses for the two approaches are based on each P frame.

In the first approach, a video stream is divided into segments instead of shots according to our predefined 4 view types. Each of the segments is defined as a ***unit*** and this approach is called a *unit-based* approach. Support Vector Machine (SVM) is used to classify these units into the predefined categories. After classification, the units from one category are actually labeled by the semantic descriptor associated with this category. This descriptor summarizes the semantic meaning of these units. Finally, static parts are merged with these classified units to form a semantic descriptor sequence representing this video stream.

In the frame-based approach, each of 24 blocks of each P frame is classified into one of the four categories (‘audience’, ‘ground’, ‘body’ and ‘other’) by using SVM at the

block level. At the frame level, line detection is applied to search for goal post. Combining the analyses at both block and frame levels, this P frame is labeled with one of the semantic descriptors. Consecutive P frames with the same label is considered as a segment. Then, a buffer-based method is used to look for boundaries for each segment. In the post-processing stage, those static parts are merged with their neighboring segments.

The two proposed approaches are tested on a total of 450 minutes of soccer video without commercial from FIFA World Cup 2002. For the unit-based approach, the highest accuracy is 81.2% for detection of 'audience' while the lowest is 70.9% for detection of 'mid-range body'. The average accuracy is 76.1%. The processing speed is 21 frames per second. For the frame-based approach, the highest accuracy is 87.0% for detection of 'Far view of whole field' while the lowest is 74.0% for detection of 'Player'. The average accuracy for this approach is 81%. The processing speed is 17 frames per second.

Conference Presentation

[1] Haiping Sun, Joo-Hwee Lim, Qi Tian and Mohan S. Kankanhalli, "Semantic Labeling of Soccer Video", Proceeding of Fourth IEEE Pacific-Rim Conference On Multimedia (PCM), Singapore, December 15-18, 2003 (ISBN 0-7803-8185-8), Oral Presentation.

[2] Qing Tang, Haiping Sun, Joo-Hwee Lim, Jesse Jin, Qi Tian, "A Generic Mid-level Representation for Semantic Video Analysis", Accepted by the Eleventh IEEE International Conference on Image Processing (ICIP), Singapore, October 24-27, 2004, Oral Presentation.

Chapter 1

Introduction

1.1 Motivation

Nowadays, with the progress in video compression, storage and communication, we are able to put a large amount of digital videos in database or online for users to perform query for some interesting or meaningful data. While the amount of video data is rapidly increasing, multimedia applications are still very limited in content management capability. Therefore, mining information in video data becomes an increasingly important problem as digital video becomes more and more pervasive.

The ubiquitous consumption of video, however, poses many problems among which the field of multimedia processing focuses on the effective description of video information (video modeling), the relationship between low-level features and semantic meanings of video information (video processing/analysis), and the querying of such information for fast and easy access to the relevant set at a later time (video querying / video search and retrieval).

As the most popular sport, soccer game attracts billions of people. However, even the most faithful fans cannot watch hundreds of games taken on a weekly basis. According to reports in [70], there are over 5,000 official games taken all over the world annually, or at least 13.7 games everyday. How could fans finish watching so many games?

If there is a multimedia analysis tool, which could automatically parse soccer video and output required video clips or the most interesting events such as goals, corner kicks and free kicks, fans could go though many more games without spending much time. This can entertain these funs and in turn further popularize the sport itself. So, soccer video indexing, especially event detection is absolutely necessary.

Event detection in soccer video is a high-level analysis, which needs an effective description of soccer video information and approaches to bridge the gap between low-level features and semantic meanings as its foundations. However, research in this field is far from enough. Shot is commonly used as an intermediate representation, but its propriety for soccer video parsing needs to be further studied and other mid-level representations should be explored. This thesis work has been inspired by this motivation.

1.2 Overview of the Proposed Mid-level Representation

The goal of this research work is to define and realize an appropriate mid-level representation for soccer video analysis.

Based on our study of the soccer video structure, we concluded that shot is not suitable as a mid-level representation for soccer video analysis. Therefore, we provide a new method instead. In this method, a soccer video can be classified into 7 categories associated with 7 semantic descriptors (close-up view, audience, far view of whole field, far view of penalty box, goal post in close-up view, player/players and mid-range view) to limn their semantic meanings respectively. So, a soccer video stream can be divided into segments, each of which belongs to one of these 7 categories. In another word, this video sequence can be delineated by a corresponding semantic descriptor sequence. This sequence is the proposed mid-level representation.

In order to convert a soccer video stream into a semantic descriptor sequence, a computational framework is proposed and two approaches are devised to realize this framework. There are three stages, pre-processing stage, processing stage and post-processing stage in both of the approaches. To reduce the computational complexity, motion magnitude is used to preliminarily segment a soccer video stream into relatively static parts and active parts in the pre-processing stage. Motion features carrying by static parts are ignored and they are processed again in the post-processing stage. Each P frame is divided into a 4 by 6 grid, each of which is called a ***block***.

In one approach, we have defined 4 view types. After pre-processing stage, each active part is divided into segments instead of shots according to the predefined 4 view types. Each of the segments is defined as a *unit* and this approach is called a unit-based approach. Then with help of Support Vector Machine (SVM), motion features are used to classify these units. Finally, static parts are merged with classified units to form a descriptor sequence to represent the video stream. In this approach, we used relative less domain knowledge to do segmentation and classification because we wanted to find an effective generic method which can also be adopted with little modification for the analysis of other types of sports videos.

In the other approach, segmentation and classification are integrated. Combining the analysis at block and frame levels, each P frame is labeled with one of the predefined Semantic Descriptors. Consecutive P frames with a same Descriptor is regarded to belong to the same segment. Then, a buffer-based method is used to look for boundaries for each segment. In the post-processing stage, those static parts are merged with their neighboring segments. We call it a frame-based approach. The purpose of this approach is to use much more domain knowledge to build up a robust system for only soccer video analysis.

After processing by one of the approaches, the input soccer video stream finally becomes a sequence of descriptors.

1.3 Organization of this Thesis

The remaining contents of this thesis are organized as follows. In Chapter 2, previous work on video segmentation, retrieval and those related closely to soccer video analysis are reviewed. In Chapter 3, the proposed mid-level representation is given. Seven semantic descriptors are also introduced in this chapter. From Chapter 4 to 7, we introduce two novel approaches to do semantic soccer video analysis. Because the pre-processing stage and post-processing are the same for both of the two approaches, these two stages, as well as the summary of these two approaches, are presented first in Chapter 4. In Chapter 5 and Chapter 6, the unit-based approach and the frame-based approach are discussed in details, respectively. The experimental results for both of the two approaches are then presented in Chapter 7. The conclusion, the generality of this mid-level representation, its contributions and the future work are discussed in the last chapter.

Chapter 2

Related Work

Multimedia information systems are increasingly important with the advent of broadband networks, high-powered workstations, and compression standards. Compared with still images, videos are dynamic data with the temporal dimensions. That means a video cannot be only regarded as a sequence of still images with information in temporal dimensions ignored. While lots of techniques are developed in image retrieval, unique features of video data give rise to many new challenging issues.

The purpose of this thesis is to discuss semantic soccer video analysis, so the theory and methods used in soccer video analysis need to be carefully studied. In this chapter, existing works on video segmentation and retrieval are surveyed in the first and second sections because it can help us understand commonly used approaches in video analysis. With these understandings, we can better study related work in soccer video analysis, which is discussed and compared in the last section.

2.1 Video Segmentation

Video structure parsing is an initial step to organize the content of videos. Video data are typically organized in a typical hierarchical structure as shown in Figure 2.1. In this step, some elementary units such as scenes, shots, frames, key frame and objects are generated. A successful structure parsing is important for video indexing, classification and retrieval. In the past, many works have been done in video structure parsing, especially in shot detection, motion analysis and video segmentation.

As discussed above, video data are structured into a lot of shot units. Shot changes should be detected before dividing video data into shot units. A shot change can viewed as detection of a camera break. Normally, there are three major editing types of camera breaks: cut, wipe and dissolve. A cut is an immediate change from a shot to another shot; a wipe is a change where first frame of a shot replace with last frame of another shot gradually; a dissolve is a change where one shot gradually appears (fade-in) and another shot slowly disappears (fade-out). A cut can be detected by comparing two adjacent frames. While wipe and dissolve are difficult to detect since they are change gradually. The transition between shots usually corresponds to a change of subject, scene, camera angle, or view. Therefore, it is very natural to use shots as the unit for video indexing and analysis.

There are many works for detection of camera breaks in the past few years. They can be grouped into two categories: uncompressed and compressed domain. Some typical methods for the detection of camera breaks could be found in [7][13][20][44]. Recent published papers for shot change detection could be found in [12][37][39][41] [45][46][55]. Most work has been focusing on pixel difference, intensity statistics comparison, histogram distance, edge difference, and motion information. Among these methods, histogram-based ones have been consistently reliable, while DCT coefficient-based ones give the lowest precision. Motion information based methods are somewhere in between. Some work for performance evaluation of shot detection could be found in [34][57].

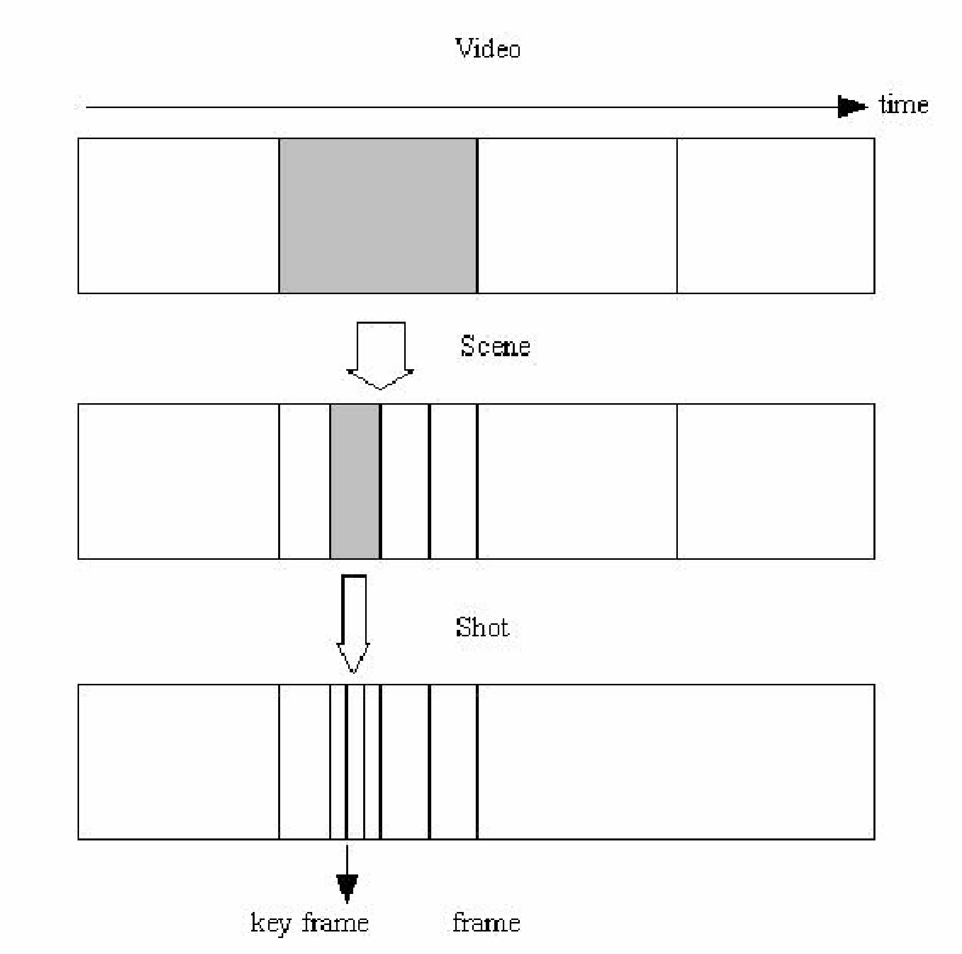


Figure 2.1 Hierarchical structure of video

Some work has been done on detecting these special effects. Related works can be found in [6][22][30][58]. A review and comparison of some of these techniques can be found in [46]. In a recent review paper, Lienhart [46] compares four major shot boundary detection algorithms, which include fade and dissolve detection. Extensive experimental results also favor the color histogram based method [33] for shot boundary detection, instead of the computationally expensive edge-change-ratio method [48].

In [45], a unified framework for semantic shot classification in sports videos is presented. Unlike previous approaches, which focus on clustering by aggregating shots with similar low-level features, the proposed scheme makes use of domain knowledge of a specific sport to perform a top-down video shot classification, including identification of video shot classes for each sport, and supervised learning and classification of the given sports video with low-level and middle-level features extracted from the sports video. This framework looks good but still has some problems: 1) the test data used is not clearly mentioned; 2) methods used to detect flying graphics are too specific; 3) Shot segmentation is finished by some commercial software and if the segmentation meets the require of shot classification.

2.2 Video Retrieval

Video segmentation is not a goal in itself but just a means for further analysis. For example, it can be used in video retrieval. We have already looked at work in video segmentation; from now on, related work in video retrieval will be surveyed.

To date, most video retrieval systems are used to retrieve similar video based on low level features. Video retrieval faces the same problem with image retrieval that it lacks a semantic model and effective representation tool to express human perception.

There exists a gap between high semantic concept and low level features. How to bridge the gap is the most challenging topic in video classification and retrieval research. In this section, we will survey recent work on similarity-based retrieval, clustering-based video retrieval and semantic video retrieval.

2.2.1 Similarity-based Video Retrieval

In current video retrieval system, there are two methods used for retrieval: similarity-based and cluster-based methods [13]. Similarity-based method is employed to retrieval similar video key frame, shot or video scene segment. Similarity matching can be based on the features extracted locally or globally. In a simple way, similarity measure is based on computing the similarity of related key-frame between two videos. More sophisticated methods are employed the spatio-temporal features of video frames between two videos [49][53][68]. Dagtas et al. [49] presented several motion descriptors as intermediate motion model for event-based video retrieval. They retrieved the event videos by computing the similarity of different motion models. Chang et al. [53] proposed a method to retrieval video object by computing similarity of motion trajectories and trails in the spatial and temporal domains. Chang et al. also presented a semantic visual template, which can express the semantic concept [60]. Detailed explanation of the idea will be discussed in later section.

2.2.2 Clustering-based Video Retrieval

Clustering method is introduced as a solution to organize the content of video collections. It provides efficient method to classify and index the video since similar videos are clustered into similar group. Recent work on cluster-based retrieval can be found in [5][19][61]. In [5], Clarkson et al. proposed a framework to find the event by clustering the nature input audio/visual data. They developed a system that can cluster the video data into events such as passing through doors and crossing the street [5]. The clustered events can also be clustered into high-level scene.

2.2.3 Semantic Video Retrieval

Semantic video retrieval of video content is viewed as the promising trend of computer vision and multimedia. Effective semantic retrieval of video is a way to ultimate multimedia understanding. Currently most works are focusing on frame-based structure modeling. Fully automatic multimedia understanding is almost impossible in state-of-the-art. Although it is a very challenging work, there still exist some good research work resid ed on this topic [5][8][9][10][15][16][27][31] [47][50][52][54][60].

In [50], Naphade et al. proposed a probabilistic framework for modeling multimedia object called 'Multiject' and modeling semantic concepts called 'Multinet'. 'Multiject' can represent low-level feature, such as visual features, audio features and textual features. It can also express the intermediate-level meaning such as semantic template [54] and other high-level semantic concepts. The advantages of a multinet are that it provides a framework for support four aspects of semantic indexes. One of its disadvantages is that the complexity of the framework will increase exponentially when the scope of knowledge is increased.

In [54], Chang et al. provide a Semantic Visual Templates (SVT) to modeling the low-level feature and high-level semantic object. They introduced an idea of SVT to bridge the gap between the user's information needs and what the systems can deliver. Although the semantic visual template can express the semantic concept intuitively, however it can only describe some basic and simple semantic concept. It is quite difficult to represent a high-level semantic event concept by sketching an intuitive template.

In the past, a lot of works have been proposed to extract and abstract the video objects in order to model the semantic concepts of objects and events. In [31], Hwang et al. proposed a scheme for object-based abstraction and analysis and semantic event modeling. However, based on the state-of-the-art in computer vision, it is difficult to build such a system since the semantic features modeling depends on domain-specific knowledge.

2.3 Soccer Video Analysis

As the most popular sport, soccer game attracts billions of people. However, even the most faithful fans cannot finish hundreds of games taken on a weekly basis. So, video indexing, especially event detection in soccer videos is absolutely necessary. Methods used in video segmentation and retrieval have already been reviewed above. As a genre of video, soccer video can be analyzed by these methods with some modification. In this section, some important works related to soccer video analysis are reviewed and compared. This can help us have a clear idea about what have been done and what need to be further studied.

Y.H. Gong et al. in [65] proposed a system that can automatically parse soccer video programs using domain knowledge. The parsing process was mainly built upon line mark recognition and motion detection. They categorized the position of the play into several predefined classes by recognizing the compound line pattern with signature method. The motion vectors field is used to infer the play positions for those scenes without line marks. Despite the strong semantic indexes from the categorization of play positions, they have yet to address these two problems: 1) how to identify different camera angle and shooting scale, otherwise the line mark recognition cannot be robust; 2) how to determine reasonable segment for processing.

D. Yow et al. in [14] presents techniques to automatically detect and extract the soccer highlights by analyzing the image contents, and to present these shots of action by the panoramic reconstruction of selected events. The analyses include the recognition of prominent features of the game, tracking of ball, camera movement compensation for effective recognition, and construction of the panoramic views. The authors pointed out a direction for application of soccer video analysis.

V. Tovinkere et al. in [59] present an effective data mining framework for automatic extraction of goal events in soccer videos. The extracted goal events can be used for high-level indexing and selective browsing of soccer videos. The proposed multimedia data mining framework first analyzes the soccer videos by using joint multimedia features (visual and audio features). Then the data pre-filtering step is performed on raw video features with aid of domain knowledge, and the pre-filtered data are used as the input data in the data mining process using classification rules. The proposed framework fully exploits the rich semantic information contained in visual and audio features for soccer video data, and incorporates the data mining process for effective detection of soccer goal events. This framework has been tested using soccer videos with different styles as produced by different broadcasters. The results are promising and can provide a good basis for analyzing the high-level structure of video content.

O. Utsumi et al. in [40] proposed a novel object detecting and tracking method in order to detect and track objects necessary to describe contents of a soccer game. On the contrary to intensity oriented conventional object detection methods, the proposed method refers to color rarity and local edge property, and integrally evaluate them by a fuzzy function to achieve better detection quality. These image features were chosen considering the characteristics of soccer video images, that most non-object regions are roughly single colored (green) and most objects tend to have locally strong edges. We also propose a simple object tracking method, which could track objects with occlusion with other objects using a color based template matching. The result of an evaluation experiment applied to actual soccer video showed very high detection rate in detecting player regions without occlusion, and promising ability for regions with occlusion.

P. Xu et al. in [42] introduced a framework for play / break events detection in soccer video. In this paper, three kinds of views in soccer video, global, zoom-in and close-up, are predefined. The counterparts' terms of these views are long shot, medium shot, and close-up, respectively. Here the grass value and classification rules are learned and automatically adjusted to each new clip. Then heuristic rules are used in processing the view label sequence, and obtain play/break status of the game. The system is novel, but it is just a good start for further event detection in soccer video.

A. Ekin et al. in [1] presented a fully automatic and computationally efficient framework for analysis and summarization of soccer videos using cinematic and object-based features. In this paper, algorithms of dominant color region detection, robust shot boundary detection and shot classification, as well as goal detection, referee detection, and penalty-box detection, are discussed. Three types of summaries can be automatically produced: i) all slow-motion segments in a game, ii) all goals in a game, and iii) slow-motion segments classified according to object-based features. The algorithm of dominant color region detection is very impressive, but the methods used in goal detection and referee detection depend heavily on man-made rules.

L.Y. Duan et al. in [35] presented a unified framework for semantic shot classification in sports videos. Unlike previous approaches, which focus on clustering by aggregating shots with similar low-level features, the proposed scheme makes use of domain knowledge of a specific sport to perform a top-down video shot classification, including identification of video shot classes for each sport, and supervised learning and classification of the given sports video with low-level and middle-level features extracted from the sports video. This framework looks good but still has some problems: 1) where the test data came from is not clearly mentioned; 2) methods used to detect flying graphics are too specific; 3) their methods for shot classification is mainly based on shot segmentation, which is done by some commercial software.

Other works such as [2][24][26][62] are also related to soccer video analysis. With consideration of our research work, a comparison among [1][35][42] is given in Table 2.1.

Table 2.1 Comparison of research work in soccer video analysis

Paper	Function of System	Feature used	Classes of Shot	Result	Contribution	Comment
P. Xu	View classification Grass Orientation Classification Play/Break Segmentation	Color (hue) Texture Rules	Long Shot Medium Shot Close-up	For view classification, 85% For Play / Break Segmentation, 75%	Color-based grass detector Play/Break Segmentation	The thresholds for different games in view classification are different Using simple rules to do Play/Break Segmentation
L.Y. Duan	Shot Classification	Color Motion Texture Camera motion pattern Dominant Object Motion Homogeneous Regions Rules	Close-up Field View, Following, Player Medium Still, Audience, Corner Kick, Goal View, Replay	85% -95%	Relationship between shot and semantic meanings Mapping from low-level features to mid-level features Mid-level features representation Fusion of valid mid-level features at shot level Real-time Performance	The method for Flying Graphics detection is too specific Method for detection of Field-Players Interaction Curve (FPIC) may be heavily affected by bad light conditions Where their testing data came from is not clearly mentioned
A. Ekin	Shot classification Slow-motion detection Event detection	Color Motion Rules	Long Shot In-field media shot Close-up / Out-field shot Goal Referee Penalty Box	For cut detection, 97.3% recall and 91.7% precision. For Gradual transitions, 85.3% recall and 86.6% precision. short classification 88%.	Field color detection Novel features for shot classification	Make full use of color information But haven't put much effort on how to use motion features

Chapter 3

A New Mid-level Representation for Soccer Video Analysis

Video processing and computer vision communities usually employ shot-based structural video models, and associate low-level descriptors such as color, texture, shape and motion, and semantic descriptions in the form of textual annotations, with these structural elements. But there are very little work that aims to bridge the gap between the low-level features and semantic descriptions to arrive at a well-integrated structural-semantic video model. To this effect, we propose a novel mid-level representation in this chapter for efficient soccer video parsing.

The drawbacks of using shot as a mid-level representation for soccer video analysis are first reviewed in this chapter; and then the proposed method is introduced. In this method, 7 categories are defined for soccer video classification, and 7 semantic words are selected to name these categories respectively to show their semantic meanings. Thus, a soccer video stream can be represented by a sequence of descriptors after segmentation and classification processing – this is the mid-level representation for this soccer video stream. Relevant definition and illustration are given in Section 3.2.

3.1 Drawbacks of Shot-based Mid-level Representation

Traditionally, structural video analysis represents video as a union of smaller coherent shots that are obtained by a temporal or a spatio-temporal segmentation process. The boundaries of these temporal shots correspond to large differences in some feature space while a temporal shot has similar features within itself. These features are usually a combination of color, texture, shape, and motion, which are commonly referred to as low-level features.

A shot can be defined as a collection of frames recorded during a continuous motion of the camera. There are two main reasons of doing this: 1) to simplify computational complexity in video processing; and 2) the assumption that shots in a video stream can be regarded as a natural segmentation. Hence, the frames within a shot represent a continuous action in time and space, and share the same high-level features as well as similar low-level features. Thus, the frame-to-frame similarity within a shot is exploited to generate compact video representations by *key frames*, which refer to one or more frames in a shot that best represent its content [24].

Can shots elucidate and highlight both the temporal and the spatial information of the soccer video? Can shots represent the corresponding semantics for soccer video analysis? These are some of the questions addressed in this section. In the following, we summarized two drawbacks based on the video parsing results using traditional shot-based approach:

- Shot-based representation can only describe video at a coarse level
- Too many shot transitions make shot boundary detection difficult

3.1.1 Shot-based Coarse Level Representation

“Can shot be good enough as a mid-level representation for soccer video analysis?” The answer is “NO”. We are going to explain our reasons from two aspects: *soccer video structure* and *event detection requirement*.

We first consider the structure and components of soccer video from the angle of shot. A soccer video comprises around 600 shots. If we use shot as the mid-level representation for a soccer video, we can represent this video as a sequence of shots, each of which can be represented by one or more key frames. That means we finally get a sequence of key frames and the work of parsing this video converts to that of analyzing the frame sequence.

But in fact, a shot may not correspond well to some semantic meaning. This impropriety in soccer video can be found in Figure 3.1.

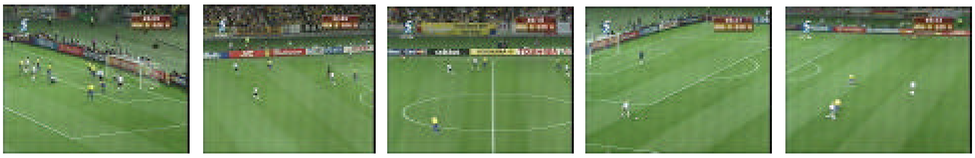


Figure 3.1 Frames selected from a long shot to present three different field views

The frame sequence shown in Figure 3.1 is selected from a long shot (more than 10 seconds in length). The camera presented what happened first in the right penalty box, then in the middle field, and then in the left penalty boxes and finally again in the field between the two penalty boxes. This process is depicted in Table 3.1 according to the order of occurrence.

Table 3.1 Explanation of the actions captured in a long shot as shown in Figure 3.1

Location	*Events*	*Camera Actions*
Right Penalty Box	White side tried to score but failed; The goalkeeper of yellow side initiated a beat-back	Focusing on the right penalty box to show this attacking
Between Two Pena lty Box	Yellow side passed ball toward the left penalty box.	Moving toward the left penalty box
Left Penalty Box	The goalkeeper of white side got the ball and passed it to his teammates.	Focusing on the left penalty box and then moving following the ball
Between Two Penalty Box	White side passed ball toward the right penalty box.	Moving toward the right penalty box

According to our observation over 30 soccer games, we found out that:

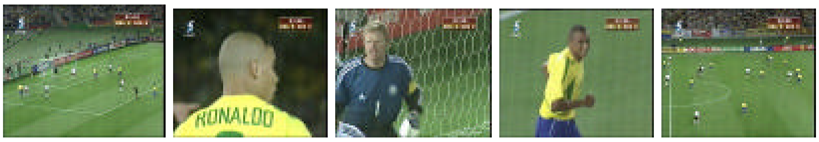
- One penalty boxappears in a far view when goals or corner kicks occurred;
- The field area appears in a far view when the players are fighting for ball possession, or the attacking-defending procedure.

Based on the statistic s over those games, we know that this kind of shots occupies more than 40% of total time in each game. Much information will be lost if we only give a semantic meaning to such kind of whole long shots. The error rate resulting from this approach varies with the frequency of such far view shot that depends on the broadcasting style, and it may reach intolerable levels for the employment of higher level algorithms for certain analysis. In these cases, we may predefine two categories: one indicating one of penalty boxes and another indicating the area in between the penalty boxes, and label them with semantic meaningful words such as ‘far view of penalty (FP)’ and ‘far view of middle field (FM)’ respectively. According to the two observation results listed above, this division could provide us more accurate information in high-level soccer video analysis such as goal detection, estimation of ball possession and so on.

So, according to this rationalization, we believe that shot represents semantics only at a coarse level, such as the name of an event or the name of the leading object in the scene, e.g. goal and corner kick are treated as shots in [35]. It is not a good representation for soccer video analysis.

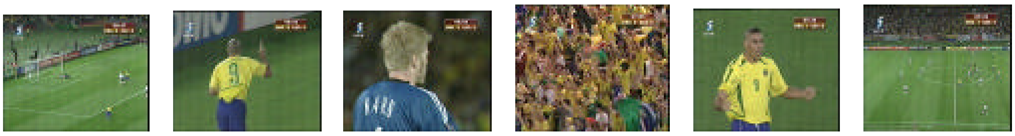
Then we study this viewpoint from the angle of what we need in event detection. After the observation of more than 30 soccer games broadcasted by different TV stations, we concluded that a corner kick event or a goal event could always be claimed detected when we find a frame sequence containing sub-sequences like the following in Figure 3.2 or in Figure 3.3.

Each frame in Figure 3.2 and Figure 3.3 represents a sub-sequence, and in each sub-sequence, a frame is similar to others in both content and camera capture positions. We have to emphasize that the sequence here is not equal to a shot. For example, the one labeled by (FP) is only a part of a long shot sometimes as presented above. The meanings of the words in the two figures are explained between Figure 3.3 and 3.4. These words are used to depict the semantic meanings of the sub-sequence. So, a goal event or a corner kick event is actually decomposed to a semantic word sequence as shown in Figure 3.4.



(FP) (CP) (Net) (Player) (FM)

Figure 3.2 A frame sequence from a typical corner kick event



(FP) (Player) (CP) (AD) (Player) (FM)

Figure 3.3 A frame sequence from a typical goal event

FP: Far view of penalty box

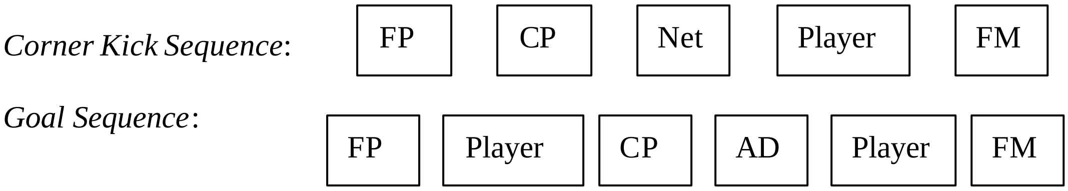
CP: Close up view

Net: Goal Net

Player: A player in mid-range view

FM: Far view of middle field, in which the penalty box is invisible

AD: Audience



Corner Kick Sequence: FP CP Net Player FM

Goal Sequence: FP Player CP AD Player FM

Figure 3.4 Two sequences representing corner kick and goal events respectively

This gives us three suggestions:

1) In a soccer video stream, an event can be decomposed as a sequence of semantically meaningful segments, (we call them sub-sequences in above figures) and they do not have one-to-one correspondence to a shot.
2) By assigning each segment a semantic and yet atomic label, an event can be represented by a sequence of semantically meaningful labels.
3) How to divide a video stream into these semantically meaningful segments should be considered.

The first two points are related to our proposed novel mid-level representation and are introduced in Section 3.2, while the last one is relevant to the realization of this representation and will be discussed in the following chapters.

In practice, we designed two different segmentation methods. In one method, we first defined 4 view types for segmentation, whose names, sample image, type models and definitions are shown in Figure 3.5. In the preprocessing stage, a video stream is preliminarily segmented into relatively active parts and static parts according to motion magnitude of every frame. The static parts will be processed in the post-processing stage while each of active parts is segmented according to the 4 view types. Each segment from each active part is defined as a ***unit***. So, this method is called unit-based method.

In the other method, called frame-based method, analysis of soccer video is based on frame and segmentation and classification are integrated into one stage. Segments are the outcome of the integrated stage.

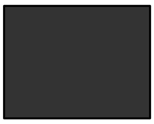
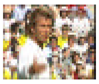
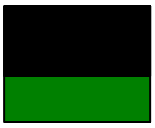
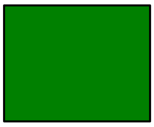
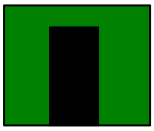
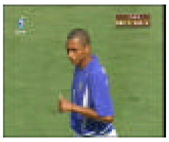
Type	Type Name	Type Models	Sample Frames	Comments
I	No Field			Almost no field appears
II	Part Field			The field appears at the lower part of a frame
III	Full Field			The field almost occupies the whole frame
IV	Field With Player(s)			Player(s) standing within the field

Figure 3.5 Definitions of four view types for segmentation in the unit-based approach

3.1.2 Transitions in a soccer video stream

According to [11], there are three major types of camera breaks: cut, wipe and dissolve. A camera cut is an instantaneous change from one shot to another; a wipe is a moving boundary line crossing the screen such that one shot gradually replaces another; a dissolve superimposes two shots where one shot gradually lighten while the other fade out slowly. Wipe and dissolve are normally referred to as gradual transitions.

According to statistic data in [11], more than 70% of all kinds of transitions are cut and less than 30% are other kinds of transitions; a sports video clip almost always contains both cuts and gradual transitions. So, the detection of these kinds of transitions except for cut should be more important if we insist to do shot segmentation. But the accuracy rate of transition detection is not very good, around 85%; and transition detection alone is still a difficult research topic.

We experimented with shot segmentation at the beginning of our research work. We applied two relatively simple methods to do shot segmentation. In the color histogram based method, a threshold is set and color histogram of each frame is calculated and compared with that of its neighbors to detect shot boundaries. In the motion based method, the ratio of the number of macroblocks predicting from the upcoming picture to that from the preceding picture for each B frame is compared with a threshold to detect the boundaries of shot. The results are shown in Table 3.2. They are not satisfactory.

Table 3.2 Test results of using common methods to do shot segmentation in soccer video

	Ground Truth (Shots)	Missed (Shots)	False Alarm (Shots)	Output (Shots)
Motion-based method	231	107	4	128
Color histogram-based method	231	12	98	317

The frame sequence shown in Figure 3.6 indicates an example which may be mis-segmented as two shots: after fighting for the control of the ball, the player just runs back to a position he is expected to be as a defender. This is a long shot; but possibly, this sequence is divided into at least two segments due to the significant changes in the backgrounds.

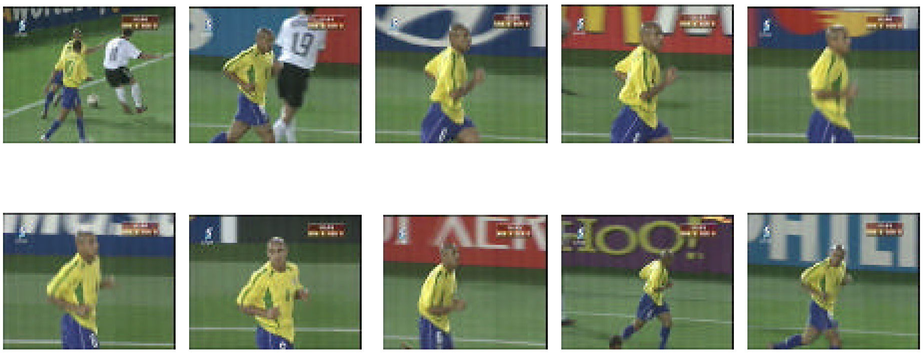


Figure 3.6 A frame sequence showing different backgrounds with the same player

With the above discussion, we believe that shot is not a suitable mid-level representation for soccer video parsing.

3.2 A Mid-Level Representation for Soccer Video Analysis

Just as mentioned in the last section, the study illustrated from Figure 3.2 to Figure 3.4 hints us a way of segmenting a soccer video into units and classifying them into several predefined categories. If we set up a suitable relationship between these categories and some simple and atomic words like those used in Figure 3.4 to indicate their semantic meanings, we can use these words to represent a whole soccer video stream. This is a method to form a mid-level representation for a soccer video, which can bridge the semantic gap between low-level features and semantic understanding.

3.2.1 Definitions of Descriptors and their Illustrations

Following this inspiration, we predefined 7 categories according our observation to soccer video and selected 7 atomic words to indicate their semantic meanings respectively. We call them ***semantic descriptors***. Their definitions are listed in Table 3.3. The word 'atomic' means each one of them cannot be further separated and also cannot be simply mapped to certain shot. The difference between 'FMA' and 'FMS' can be explained in this way: for example, given two frames, the first frame is labeled as 'FMA' and the second is labeled as 'FMS'. That means the motion magnitude of the first frame is higher than a certain threshold while that of the second frame is lower than the threshold. This explanation is also suitable to 'FPA' and 'FPS' as well as to 'MBA' and 'MBS'. This division will be helpful in detection of play / break in soccer video, which is very useful for free kick detection. The illustrations of these descriptors are given in Figure 3.7.

3.2.2 Could this Representation Work Properly

Normally, there are two steps in event detection in soccer video: mapping from extraction of low-level features to the mid-level representation in the first step and detection from the representation.

Table 3.3 The descriptors and their semantic meanings

Descriptors		*Semantic meanings*	*Description*
CP		Close up	Close-up of a player, referee, coach, goalkeeper with no field color
AD		Audience	Far view of audience
FM	FMA	Fast movement toward a penalty box or Fight for ball control	Far view of whole field (goal post not visible)
	FMS	A break happens between two penalty boxes	
FP	FPA	Move inside or outside a penalty box	Far view of half field (goal post visible)
	FPS	Players are waiting for free kick or corner kick or Break	
GP		Free kick, Corner kick, Goal, Shot or Goal Kick	Goal post in close-up view
Player(s)		Player who fouled, missed a change or is to take a free kick.	Mid-range or close-up of a player
MB	MBA	Players are fighting for controlling ball.	Mid-range view (whole body visible)
	MBS		

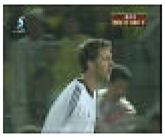
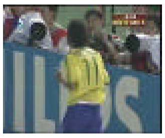
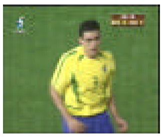
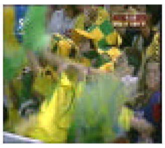
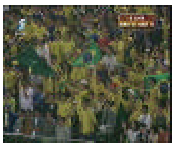
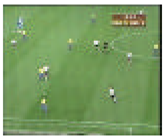
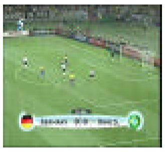
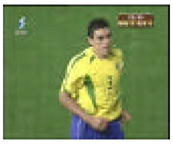
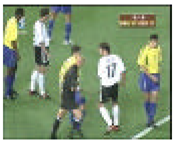
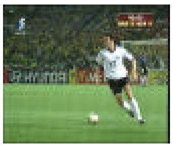
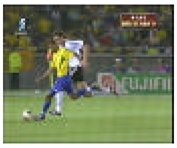
Descriptor	Sample	False Sample (what is not) & Reasons	
CP	 		According to definition, the field should not be visible
AD	 		
FM			
FP			The two goal posts should be visible
GP			
Player(s)	 		
MB	 		

Figure 3.7 Illustration of the defined seven Semantic Descriptors in Table 3.3

Y. L. Kang et al. in [67] presented a system for event detection. In this system, grammars are computed from observations to detect goal events and corner kick events in 4 FIFA2002 games. They also defined seven semantic words (Far view of whole field, Far view of half field, Mid range view (whole body of player visible), Close up view of multiple players, Close up view of single player and Goal Post), similar to ours. But their focus was to detect events from a mid-level representation of soccer video. So, they just manually segmented soccer videos according to their seven semantic words to test their grammar.

We used 4 soccer games, (a total of 450 minutes of soccer video different from what Y. L. Kang et al. used) from FIFA2002 to test our mid-level representations to see if it is effective with the new edition of the system described in [67]. The accuracies are 80% for the frame-based approach and 79.1% for the unit-based approach respectively. This confirms that our proposed mid-level representation is effective.

Chapter 4

Semantic Soccer Video Analysis

We have already outlined a novel method for mid-level representation in soccer video parsing. As mentioned before, we designed two approaches to realize this representation. In this chapter, the summary of the two approaches is presented first. Each approach contains three processing stages: pre-processing, processing and post-processing stages. Because both of the approaches are the same in the first and the last stages, these two stages are also introduced in this chapter and will not be repeated in later chapters. Also, the differences between the two approaches in the processing stage are discussed in this chapter.

4.1 Introduction to the Frame-based and the Unit-based Approaches

Recently, looking for mid-level representations for soccer video analysis becomes more and more popular, such as L.Y Duan et al. [35]. A method called *shot labeling* method is used to parse video. There are three main steps in this method: label set definition, shot segmentation and shot labeling (classification) with the predefined label set. In this method, shot is used as a mid-level representation.

We have already discussed the shortcomings of using shot as the intermediate representation for high-level soccer video parsing. Thus, we provided our proposed method for bridging the gap between low-level features and semantic meanings by using predefined Semantic Descriptors to represent soccer videos. Two approaches, the unit-based and the frame-based, are designed to realize this representation. The main stages of the two methods are similar except for the processing stage. Steps in common are shown in Figure 4.1 and summarized below.

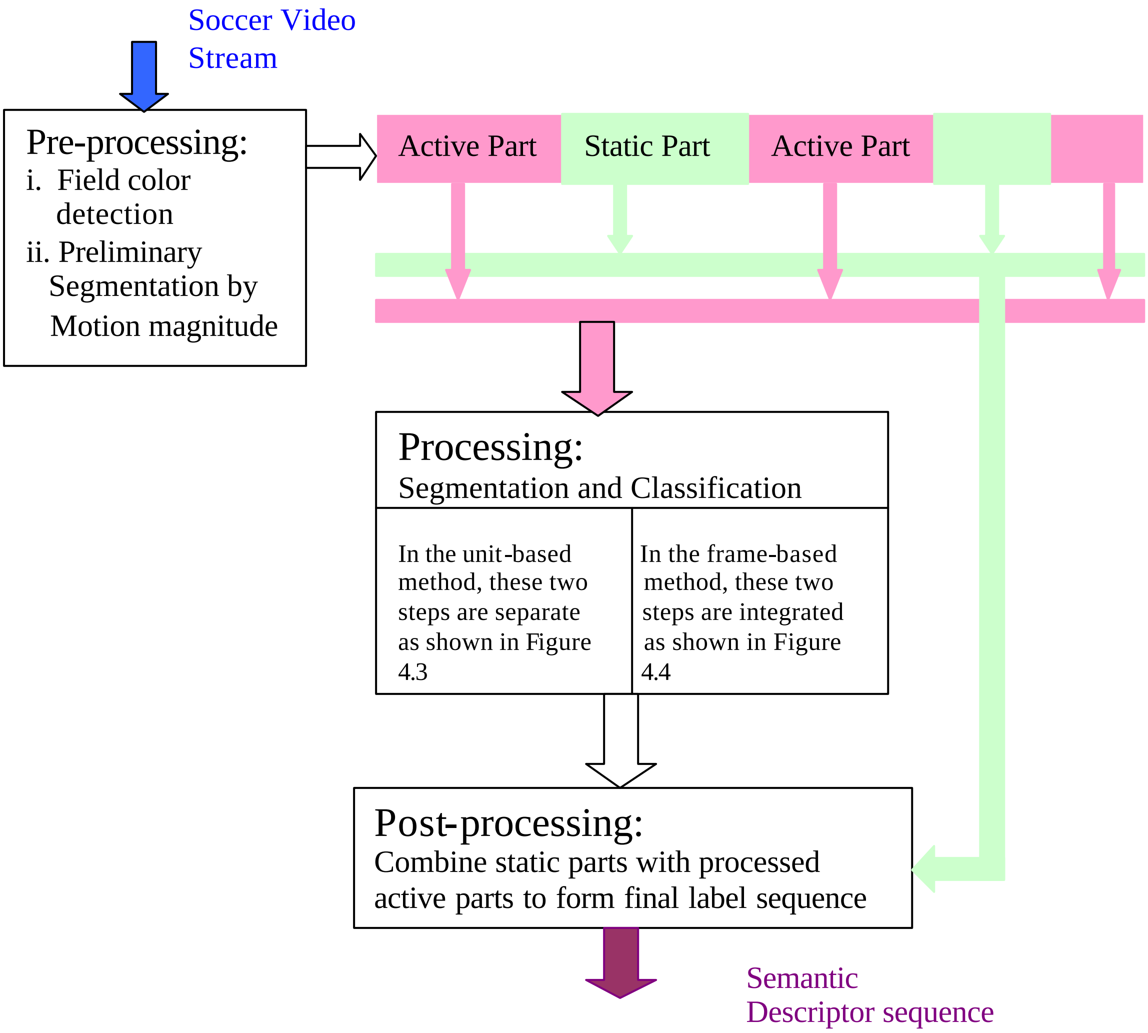

Soccer Video Stream
Pre-processing:
i. Field color detection
ii. Preliminary Segmentation by Motion magnitude
Active Part
Static Part
Active Part
Processing:
Segmentation and Classification
In the unit-based method, these two steps are separate as shown in Figure 4.3
In the frame-based method, these two steps are integrated as shown in Figure 4.4
Post-processing:
Combine static parts with processed active parts to form final label sequence
Semantic Descriptor sequence

Figure 4.1 This flow chart illustrates the common processing steps in the two approaches

1. Preprocessing: some short training clips are used to automatically compute field colors. A video stream is divided into relatively static parts and active parts. This

is a preliminary segmentation. For static parts, motion features are ignored and key frames are saved; all active parts will be processed in the next step.

2. Segmentation and Classification: Then segmentation and classification are done to each active part to form the coarse semantic descriptors sequence for the video stream. Here SVM is used as the classifier.
3. Post-Processing: static parts are merged with adjacent segments to form the final semantic descriptors sequence representing this soccer video.

4.2 Preprocessing Stage for the Two Approaches

The main purposes of this stage are to obtain field color and preliminarily segmented video streams for the consideration of speed and computational complex. They are introduced separately in this section.

4.2.1 Obtaining Field Color

Distinguishing field colors from others is not as easy as one may think because the RGB values may change under different lighting and field conditions or different camera shooting positions. Authors in [2] used a self-adapted method to detect field color in HIS color space. It is effective but too complex. P. Xu et al. in [42] set two thresholds to detect field color. With consideration of accuracy and complexity, we designed a method to solve this problem by using three tables. Their names, illustrations and functions are listed in Table 4.1.

Firstly, a table called Green Color Table (GCT) is built up manually. All colors perceived by people as field colors are recorded in this table. It is possible that some colors that are actually not field green colors are also kept in the GCT. Then some short sample clips (from view Type II, III, IV as shown in Figure3.5) from a soccer video are fed to automatically obtain field colors for this video. For the color of a block (this color should be in GCT), it is kept in the Upper Half Green Table (UHGT) if this block is believed to be colored by one of the field colors and is within the upper half of a P frame; or keeps it in the Lower Half Green Table (LHGT) if it is colored with a green color and is within the lower half of a P frame. In order to reduce effects coming from noise (field green colors could be found in audience too; also one field green color appears differently under different camera shooting positions), the size of UGT (*m*) is set to be larger than that of LGT (*n*). In our experiments, *m = 11* and *n = 6*.

Table 4.1 Three tables used for field color look-up

Table Name	Examples	Functions
GCT	□	All colors considered as field colors are saved there. It's a field color RGB value database for all games
UHCT	🟩 □	All field colors appeared in upper half of a frame in a game are saved for look-up when judging if a color of a pixel from the upper half of a P frame is a field color. It is a subset of GCT
LHCT	□ 🟩	All field colors appeared in lower half of a frame in a game are saved for look-up when judging if a color of a pixel from the lower half of a P frame is a field color. It is a subset of GCT

4.2.2. Preliminary Segmentation

As shown in Figure 4.2, a soccer video stream consists of two kinds of frame sequences. We call a frame sequence a relatively static part if the magnitudes of motion vectors of most frames in this sequence are so low that motion features can be ignored during processing. On the contrary, we call a frame sequence a relatively active part if the magnitudes of motion vectors of almost all frames are high and cannot be ignored. The motivation of doing this is to filter input video stream to speed up the process as well as to reduce computational complexity before further processing.

As summarized in [25], motion features can be extracted for block, regions, objects, and whole video frames for motion-based query and object / region tracking. Currently, they are largely used in fields such as trajectories estimation, camera operation estimation for foreground / background objects segmentation as well as video indexing.

In particular, motion features are used to improve the accuracy of shot segmentation or for classification in video parsing and indexing as described in [17], [18], [25], [41], [49] and other research work. But the proposed motion-based segmentation method is different. It is used as a preliminary segmentation. What we are concerned is only the motion magnitude and we do not care the relationship between the segmented parts and shot or other meaningful unit.

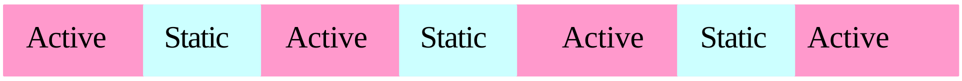


Figure 4.2 A video stream can be divided into relatively active parts and static parts by motion magnitude of a frame

For each P frame, the sum of all motion vectors' magnitudes, *Mag*, is calculated by adopting the equation shown below in Equation 4.1:

$$Mag = \sum_{j=1}^{k} \left| Mag_{mb}(i) \right| \tag{4.1}$$

Where $Mag_{mb}(i)$ means the motion magnitude of the i^{th} macroblock in current P frame.

Setting a certain threshold, a video stream can be divided into relatively static parts and active parts (shown in Figure 4.1 and Figure 4.2). The motion features in a static part are ignored and the key frames extracted are considered as its representative. The threshold is determined empirically (in our approaches, we set the threshold to be 60). Static parts are processed again in the post-processing phase.

4.3 Processing Stages in the Two Approaches

In this stage, each P frame is divided into 4 by 6 square regions, each of which is called a ***block***. In the unit-based method, analysis is mainly based on block level; while in the frame-based method, that is based on both block level and frame level. The main processing steps of the two approaches are the same except for the processing stage. In this section, the summary of this stage for each approach is discussed.

4.3.1 Processing Stage in the Unit-based Approach

The processing steps in this stage are shown in Figure 4.3. As shown, each relatively active part coming from the pre-processing stage are further segmented according to the predefined 4 view types introduced in Section 3.1 and illustrated in Figure 3.5. Each segment after segmentation is called a *Unit*. More details can be found in Chapter 5.

4.3.2 Processing Stage in the Frame-based Approach

M. Szummer et al. in [36] presented a method to classify images into two categories: indoor and outdoor. In this method, each image is divided into 4 by 4 blocks. Each block is labeled with either 'in' or 'out' with respect to its features analysis. Then the image is labeled 'indoor' or 'outdoor' according to all the 16 labels of this image.

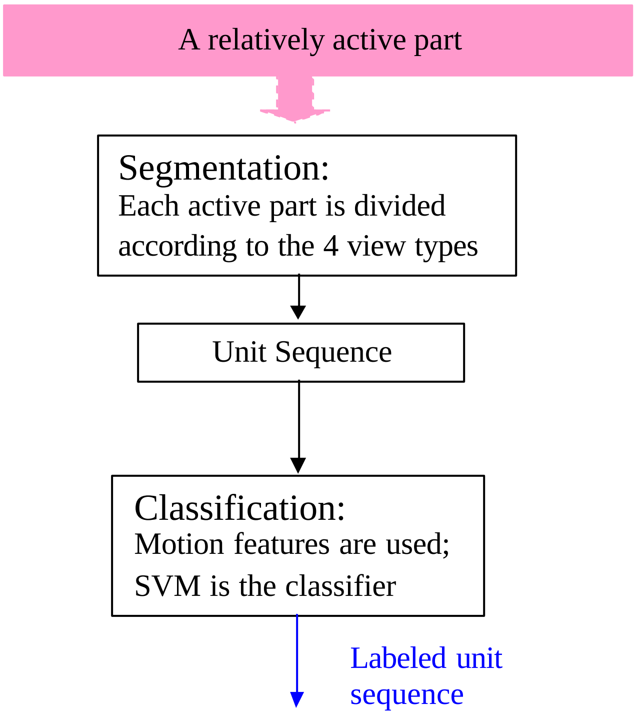


Figure 4.3 Flow chart of the processing stage in the unit-based approach

Inspired by this local classification approach, our analysis is based on both block level and frame level in the frame-based approach. 'Audience', 'Ground', 'Body' and

‘Other’ form our ***Block Label Set***. At the block level, color, motion and texture features are extracted, and SVM is used to label each block of a P frame. At the frame level, Hough Transform line detector is used to detect goal post in far view in each P frame.

The processing steps for each P frame in this approach are illustrated in Figure 4.4. Each P frame will be labeled by one of the 7 semantic descriptors whe n applying this procedure to a whole relatively active part. After that, the whole active part is consequently segmented and classified. In Chapter 6, detailed procedure is given.

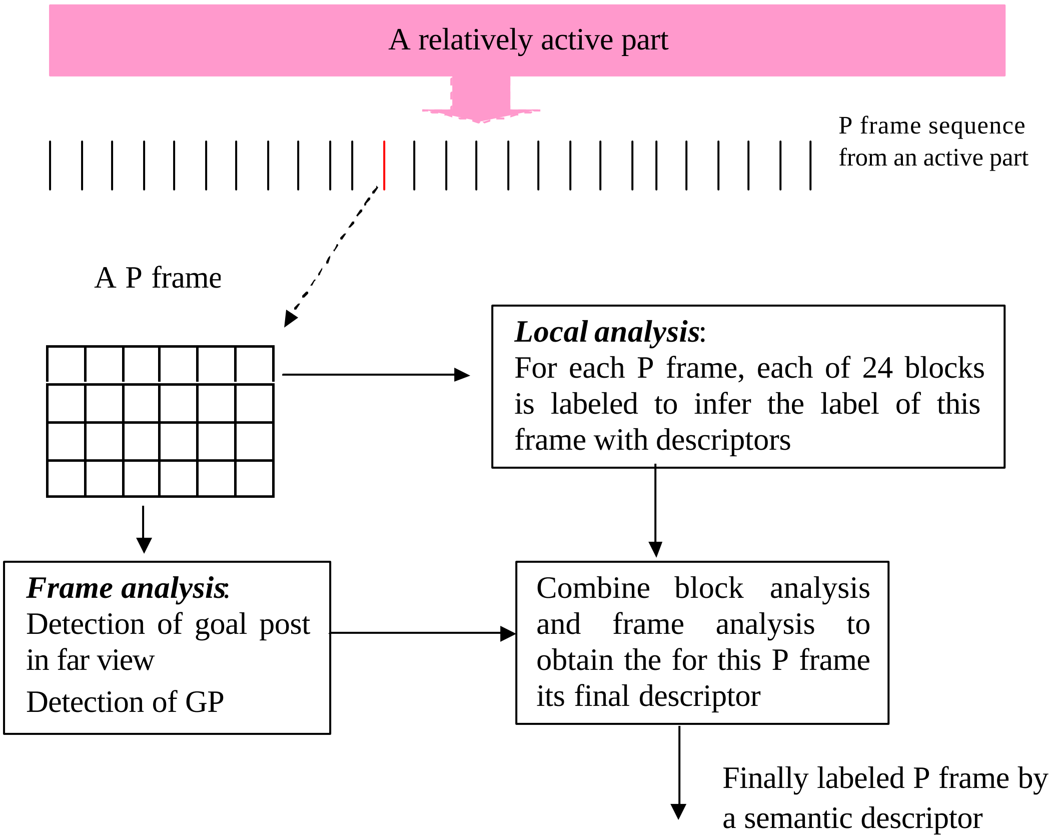

A relatively active part
P frame sequence from an active part
A P frame
Local analysis: For each P frame, each of 24 blocks is labeled to infer the label of this frame with descriptors
Frame analysis: Detection of goal post in far view
Detection of GP
Combine block analysis and frame analysis to obtain the for this P frame its final descriptor
Finally labeled P frame by a semantic descriptor

Figure 4.4 The procedure of classification and segmentation to each P frame in the frame-based approach

4.4. Post-processing Stage for the two Approaches

In this stage, static parts are combined with their neighbors. Generally speaking, a static part may contain several meaningful sub-parts. So, a static part should first be divided into sub-parts by color histogram. In practice, we adjusted the threshold so that a static part contains no more than two sub-parts. The last 5th frame of its left neighbor and the 5th frame of its right neighbor are selected as their comparable references. The fifth frame of a static sub-part is extracted as its key frame. The differences between the key frame and comparable references are computed to decide which neighbor a sub-part is to merge with, if the difference is below a threshold. Otherwise, the sub-part is to be abandoned. As a result, segments labeled by 'FM', 'FP', or 'MB', are divided into relatively active and static sub-segments by a threshold set manually.

Chapter 5

Unit-based Semantic Soccer Video Analysis

We have already provided our method to represent soccer video in Chapter 3, and outlined the two approaches in Chapter 4. In this chapter, the unit-based approach for semantic soccer video parsing will be presented in detail.

In this approach, we defined 4 view types. Each relatively active part is segmented into units according to the predefined view types. Their definitions are introduced in Figure 3.5 in Section 3.1. Here we explain the judgment rules to distinguish one view type from others in the first section. In order to classify these units, the mapping relationship between 7 semantically labeled categories and these 4 view types ought to be given. It is also discussed in this section of this chapter.

Because we used SVM as the classifier, we briefly go through its basic theory. Finally, the procedures of segmentation and classification are addressed in the second section. As said in Chapter 4, the post-processing stage is discussed in section 4.4 and will not be repeated in this chapter.

5.1 View Types and its Mapping Relationship with Semantic Descriptors

Segmentation is a good way to simplify the procedure of soccer video processing; but as we argued in Chapter 3, commonly used shot segmentation is not suitable for soccer video parsing, we therefore designed another new segmentation method instead.

P. Xu et al in [42] defined 3 types of views: global, zoom-in and close-up views. And also, they used a color-based detector to classify video. The counterparts' terms of these views are long shot, medium shot, and close-up, respectively. We feel that they are not adequate for soccer video segmentation. Just as shown in Figure 5.1, (a) and (b) are difficult to be distinguished only by the ratio of field color to pixel number, not to mention the accuracy of the algorithm for field color detection the authors used in [42].

(a) Close-up of a player (b) Medium range

Figure 5.1 These two kinds of frames represent two kinds of view types

Here, we defined 4 view types according to camera shooting positions and ratio of field colors to non-field colors within one frame. The definitions of these view types were illustrated in Figure 3.5. Here this figure is again listed in Figure 5.2 for the convenience to introduce other relevant contents.

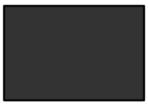
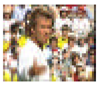
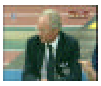
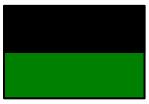
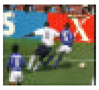
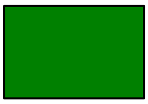
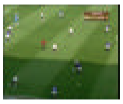
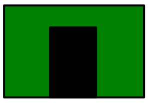
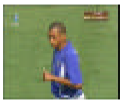
Type	Type Name	Type Models	Sample Frames	Comments
I	No Field		 	Almost no field appears
II	Part Field			The field appears at the lower part of a frame
III	Full Field			The field almost occupies the whole frame
IV	Field With Player(s)			Player(s) standing within the field

Figure 5.2 Definitions of four view types for segmentation in the unit-based approach

We used 4 kinds of green / non-green (field / non-field) templates to model and binarize the 4 view types as shown in the third column in Figure 5.2. When processing video stream in the unit-based approach, each P frame is binarized to one of these 4 templates. The judgment rules to discriminate one type from others are illustrated in Figure 5.3.

These 4 kinds of view types are defined for video segmentation. Each segment resulting from the segmentation is called a ***unit***. The relationship between a unit and a view type is illustrated in Figure 5.4.

We have already introduced the definitions of Semantic Descriptors in Section 3.2. They are used to label a soccer video to form its final analysis results. Since unit is just a transitional step during the soccer video processing, the mapping relationship between them must be given. Here, we illustrate the mapping relationship in Figure 5.5.

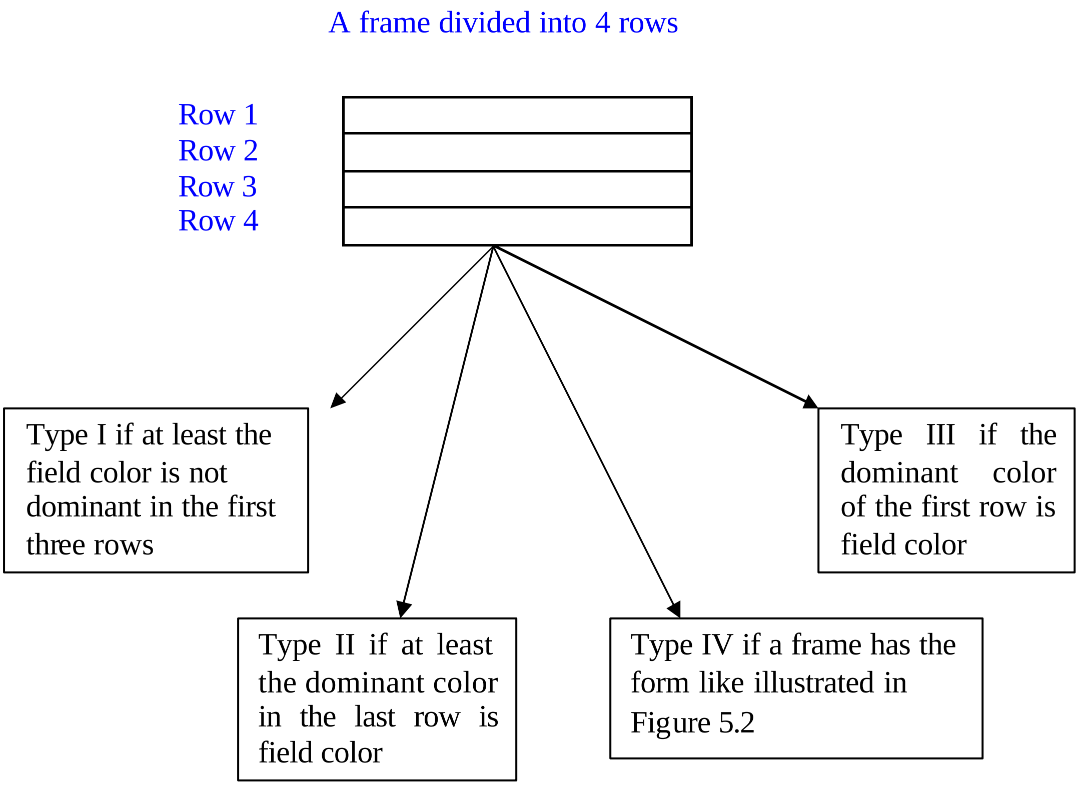

A frame divided into 4 rows
Row 1
Row 2
Row 3
Row 4
Type I if at least the field color is not dominant in the first three rows
Type II if at least the dominant color in the last row is field color
Type IV if a frame has the form like illustrated in Figure 5.2
Type III if the dominant color of the first row is field color

Figure 5.3 The judgement rules for distinguishing types from each other

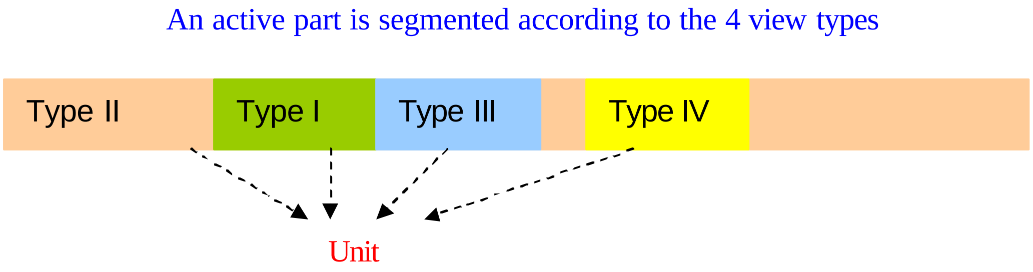


Figure 5.4 The relationships between view types and units

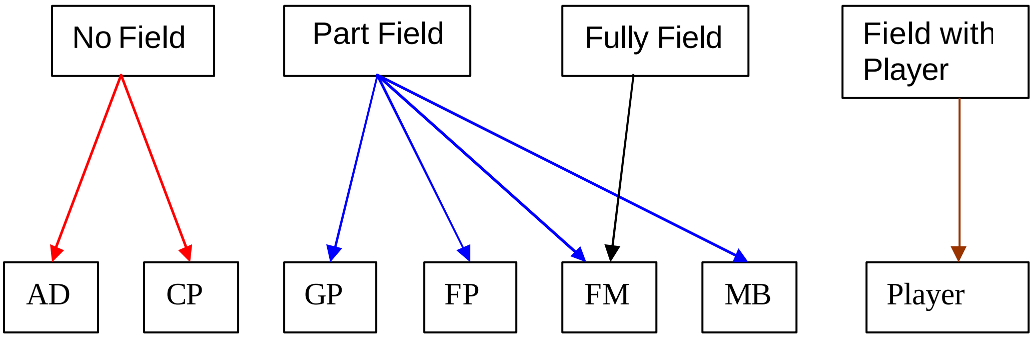


Figure 5.5 The mapping relationship s between the four view types and the Semantic Descriptors

This mapping relationship is just an ideal one. In practice, one unit belonging to type *A* can be recognized unsuccessfully as one belonging to type *B*.

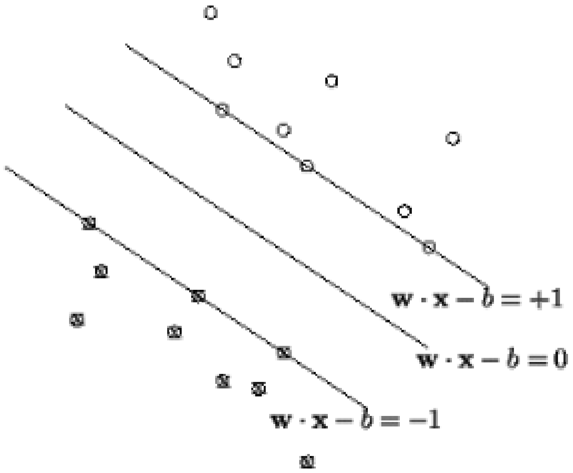
5.2 Introduction to Support Vector Machine (SVM)

SVM is used as our classifier in the processing stage, so a brief introduction is helpful to understand this method.

Support Vector Machine (SVM) is formulated based on statistical learning theory. SVM claims to guarantee generalization, i.e. the decision rules reflects the regularities of the training data rather than the incapabilities of the learning machine. It also allows various other learning machines to be constructed under a unified framework, hence simplifying comparisons and promoting understanding.

5.2.1 Linear Support Vector Classifier

First let us look at the linear support vector machine. It is based on the idea of hyperplane classifier, or linearly separability.



$\mathbf{w} \cdot \mathbf{x} - b = +1$
$\mathbf{w} \cdot \mathbf{x} - b = 0$
$\mathbf{w} \cdot \mathbf{x} - b = -1$

Figure 5.6 Illustration of linear SVM

Suppose we have ***N*** training data points {(x1,y1),(x2,y2),‥,(xn,yn)} where $x_i \in \Re^d$ and $y_i \in \{\pm 1\}$. We would like to learn a linear separating hyperplane classifier:

$$f(x) = \mathrm{sgn}(w \cdot x - b) \qquad (5.1)$$

Furthermore, we want this hyperplane to have the maximum separating margin with respect to the two classes. Specifically, we want to find this hyperplane ***H*** and two hyperplanes parallel to it and with equal distances to it,

$$H_1 : y = w \cdot x - b = +1 \qquad (5.2)$$
$$H_2 : y = w \cdot x - b = -1 \qquad (5.3)$$

with the condition that there are no data points between ***H1*** and ***H2***, and the distance between ***H1*** and ***H2*** is maximized. There will be some positive examples on ***H1*** and some negative examples on ***H2***. These examples are called *support vectors* because only they participate in the definition of the separating hyperplane, and other examples can be removed and/or moved around as long as they do not cross the planes ***H1*** and ***H2***. So, in order to maximize the distance, we should minimize the first formula below with the condition that there are no data points between ***H1*** and ***H2***:

$$\| w \| = w^T w \tag{5.4}$$

$$w \cdot x - b \geq +1, \text{ for positive examples } y_i = +1 \tag{5.5}$$

$$w \cdot x - b \leq -1, \text{ for negative examples } y_i = -1 \tag{5.6}$$

These two conditions can be combined into

$$y_i (w \cdot x_i - b) \geq 1 \tag{5.7}$$

So our problem can be formulated as

$$\min_{\text{w,b}} \frac{1}{2} w^T w \quad \text{subject to } y_i (w \cdot x_i - b) \geq 1 \tag{5.8}$$

Introducing Lagrange multipliers $a1, a2, ..., an \geq 0$, we have the following Lagrangian:

$$L(w, b, a) = \frac{1}{2} w^T w - \sum_{i=1}^{N} a_i y_i (w \cdot x_i - b) + \sum_{i=1}^{N} a_i \tag{5.9}$$

5.2.2 Non-linear Support Vector Classifier

What if the surface separating the two classes is not linear? Well, we can transform the data points to another high dimensional space such that the data points will be linearly separable. Let the transformation be $\Phi(\bullet)$. In the high dimensional space, we solve

$$L_D \equiv \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j \Phi(x_i)\cdot\Phi(x_j) \tag{5.10}$$

Suppose, in addition,

$$\Phi(x_i)\cdot\Phi(x_j) = k(x_i, x_j) \tag{5.11}$$

That is, the dot product in that high dimensional space is equivalent to a *kernel* function of the input space. There are many kernel functions that can be used this way, for example, the radial basis function (Gaussian kernel)

$$K(x_i, x_j) = e^{-\|x_i - x_j\|^2/2s^2} \tag{5.12}$$

5.3 Segmentation and Classification for the Unit-based Approach

In the unit-based approach, a video stream is first divided into relatively static parts and active parts. For active parts, they are further segmented into units according to the 4 view types. When extracting features, each P frame is divided into a 4 (rows) by 6 (columns) grid, each of which is called a block as shown in Figure 5.7. At the same time, the P frame is converted to a green/non-green frame as shown in Figure 5.8.

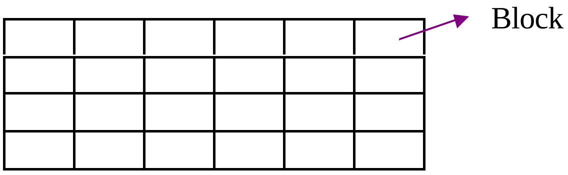


Figure 5.7 Each P frame is divided into 24 blocks averagely

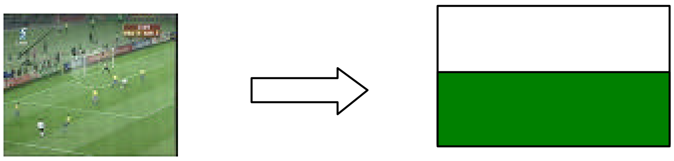


Figure 5.8 Each P frame is converted to a green / non-green frame

5.3.1 Unit-based Segmentation

As mentioned above, each P frame is divided into 24 blocks. The unit-based approach binarizes each P frame according to the following method for all the 24 blocks:

1. Get the dominant color (C_d) of a block;
2. If C_d is in the upper half of a P frame, the block is converted to non-green unless its C_d is in the UGT. If so, it is converted to green color.
3. If C_d is in the lower half of a P frame, the block is converted to non-green unless its C_d is in the LGT. If so, it is converted to green color.

Then for each of the four rows of a P frame, the number of colors (except colors in UGT or LGT) is computed and the decision rules shown in Figure 5.9 are used to do segmentation.

In Figure 5.9, 'Blk_row (i -j)' is the number of blocks considered as colored with field colors from i^th^ row to j^th^ row; 'Clr-Row (m-n)' is the number of colors from m^th^ row to n^th^ row. Ps are parameters obtained from experiments (P1=16, P2=9, P3=6).

We manually segmented 4 soccer games into units and labeled them according to the Semantic Descriptors to test this segmentation method. The results which are listed in Table 7.2 indicate that this method is robust and effective.

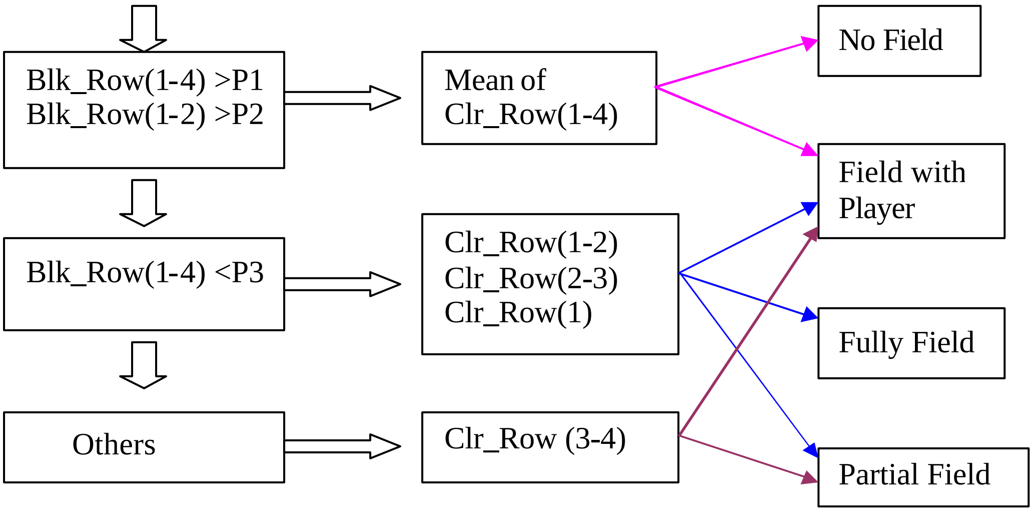


Figure 5.9 Segmentation rules for the 4 view types

5.3.2 Classification by Motion features

The dominant motion based method was adopted in many video retrieval systems, such as E. Ardizzone, et al did in [18]. However, it cannot provide sufficient motion information for users, since it is only a coarse description of motion intensity and direction between frames. Moreover, it is impossible to discriminate the object motion from camera motion in dominant motion. The parametric global motion estimation was used to extract object motion from background by neutralizing global motion as described by J. R. Ohm, et al in [32]. But unreliability and time consuming are the main drawbacks of this approach.

Camera motion based method is an alternative for video indexing. As presented by E. Ardizzone, et al in [17], the qualitative descriptions about camera motion models, such as panning, tracking, zooming, are used as motion features for video retrieval. Although the camera motion is useful for filmmakers or other professional users, it could be meaningless to the general users because they may pay no attention to camera operations when they enjoy video content. The object-based video retrieval as introduced by S. F. Chang, et al in [51] is a much better method for users. However, semantic object segmentation still needs human interaction at the current stage. Also, in most of previous works, the visual features were extracted based on the key-frame as done by H. J. Zhang, et al in [23]. Such representation may not be complete due to the sparse nature of key-frames, especially for motions in video. To overcome this shortcoming, some shot-based feature extraction methods have been proposed for color description recently by T. Lin, et al. in [56] and A. M. Ferman, et al. in [3].

With the consideration of the speed and complexity, we just used some simple yet effective motion operation.

After the unit-based segmentation, each active part is partitioned into segments, each of which belongs to one of the four view types. Motion features are extracted to do classification with help of Support Vector Machine (linear SVM). Each P frame in a unit is converted to a green/non-green frame as shown in Figure 5.8. Extracting motion features is done mainly at block level, but only one feature extracted at frame level.

At the block level, the magnitudes of motion vectors are first mapped into 3 values if the magnitudes of a motion vector are non-zero and the value of the angle of each micro-block is mapped into 8 directions as shown in Figure 5.10. Next, the means and standard deviations of magnitudes and angles of motion vectors of a block are extracted.

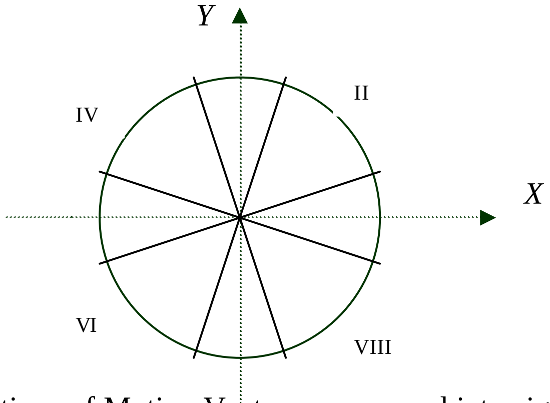
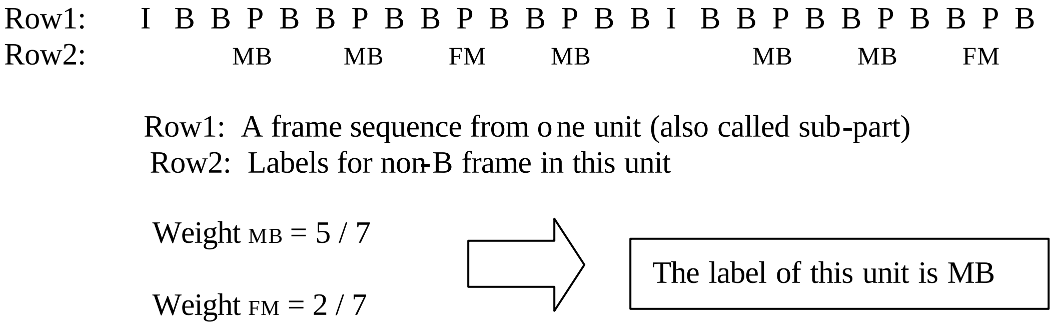


Figure 5.10 Directions of Motion Vectors are mapped into eight directions

At the frame level, only a direction frequency feature is extracted. That is, in order to describe motion, the direction distribution of all motion vectors within a frame is counted and kept. Motion texture proposed by Y.F. Ma et al. in [64] is a compact representation for motion. It can characterize 6 motions. In our system, the motion features used realize the same effect partially.

Statistic method is then used to judge the unit's semantic label. In the other word, how many frames for each of 7 kinds of Semantic Descriptors are calculated. The judgment formula (equation 5.13) is given below and an example is given in Figure 5.11.

$$\text{Weight}_k = \frac{\text{Appearance times of a label}}{\text{Total frame number of a sub active part}} \qquad (5.13)$$



Row1: I B B P B B P B B P B B P B B I B B P B B P B B P B

Row2: MB MB FM MB MB MB FM

Row1: A frame sequence from one unit (also called sub-part)

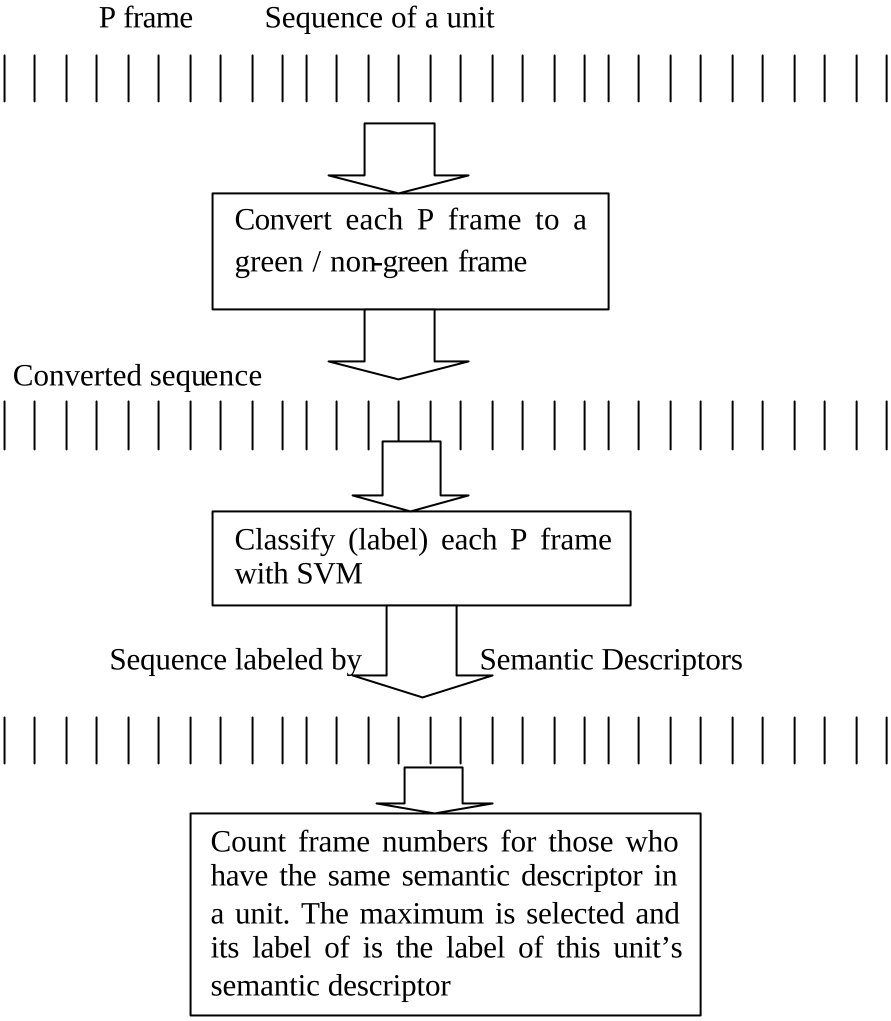
Row2: Labels for non-B frame in this unit

Weight $_{MB}$ = 5 / 7

Weight $_{FM}$ = 2 / 7

The label of this unit is MB

Figure 5.11 An example to show how to label a unit with a semantic descriptor



P frame
Sequence of a unit
Convert each P frame to a green / non-green frame
Converted sequence
Classify (label) each P frame with SVM
Sequence labeled by
Semantic Descriptors
Count frame numbers for those who have the same semantic descriptor in a unit. The maximum is selected and its label of is the label of this unit's semantic descriptor

Figure 5.12 The procedure for a unit to obtain a semantic descriptor

The semantic descriptor with the highest weight is the one to be selected for this unit. The classification so far is finished; but we need to combine the labeled units with static part in the post-processing stage to obtain the final descriptor sequence for the video stream. The steps for a unit to obtain a preliminary semantic descriptor are summarized in Figure 5.12.

In order to test the motion-based classification method, we divided the data set into training data and test data. The results presented in Table 7.3 in Chapter 7 indicate that the method is effective.

Chapter 6

Frame-based Semantic Soccer Video Analysis

The differences between this frame-based approach and the unit-based one are: 1) the classification and segmentation are integrated in this approach; 2) analysis to each relatively active part is performed based on both the block level and frame level. That is, 4 semantic words are selected to form a ***Block Label Set*** in advance. Each block of a P frame is labeled with a *Block Label* to indicate its semantic meaning. For the whole frame, it is to be labeled by a Semantic Descriptor according to its 24 block labels.

In this chapter, the definition of Block Label Set is introduced first in Section 6.1. The mapping relationship between the Block Label Set and the semantic descriptors has to be given because Block Labels and Semantic Descriptors are designed for different level analysis. It is presented in Section 6.2. Then, the integrated classification and segmentation are discussed in the last section.

The post-processing stage is necessary for the approaches to eventually represent a soccer video by a semantic descriptor sequence. This stage was discussed in Section 4.4 and is not to be repeated in this chapter.

6.1 Definition of Block Label Set

A video stream cannot be regarded as just a frame sequence, which means we cannot just analyze each frame as we process each image along for the consideration that the computational speed must be acceptable, and more importantly, there are some valuable relations in video analysis between consecutive frames in a video stream. The motivation of dividing frames into a 4 by 6 grid (totally 24 blocks) is to analyze frames locally without too much lost in processing speed.

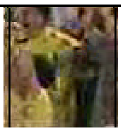
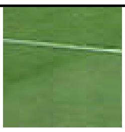
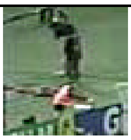
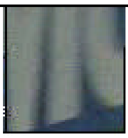
M. Szummer et, al in [36] divided each image into a grid and label each block to judge the class of the whole image in image analysis. A. Ekin et al. n [1] divided up the screen in the 3:5:3 proportion in both directions, and positioning the main subjects on the intersection of these lines according to suggestions from G. Millerson in [21] and A. M. Ferman et al. in [4]. It is used just for referee detection.

We used Semantic Descriptors to label frame and further to label units in the unit-based approach. Inspired by this idea, we can also define some basic and meaningful words to label each block of a frame so that we can deduce the semantic meaning of this frame. The Block Label Set contains four elements as shown in Table 6.1.

6.2 Relationships between Block Labels and Semantic Descriptors

The Block Label Set is just auxiliary for labeling of frames. So, the relationships between the Block Label Set and the Semantic Descriptor Set must be clarified. The relationships are illustrated in Figure 6.1. The detection of GP (close-up view of goal post) does not use this method to realize, so it doesn't present in this figure. Thus, only six relationships are listed in this figure. How to detect GP is discussed later in Section 6.3.2.

Table 6.1 Definition of Block Label Set

Component	*Abbreviation*	*Samples*
Audience	A	
Ground	G	
Body	B	
Other	O	 

6.3 Integrated Classification and Segmentation in the frame based Approach

As known, a video stream is first divided into relatively active parts and static parts in the frame-based approach as the unit-based approach does. Then, without further segmentation, all P frames belonging to the same active part are parsed and assigned with descriptors; this is a preliminary labeling. For each active part, the number of the P frames labeled by the same descriptors is counted for segmentation. The processing steps were shown in Figure 4.4 in Section 4.3.2. For the convenience of readers to understanding this approach, it is presented again in Figure 6.2.

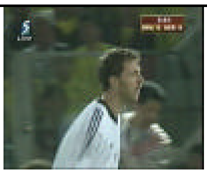
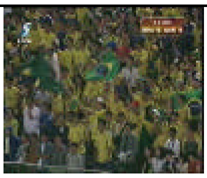
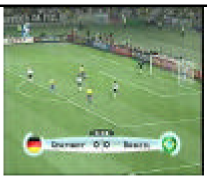
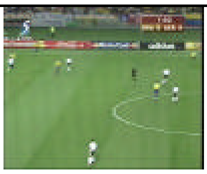
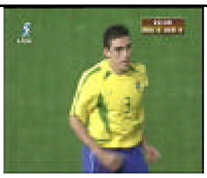
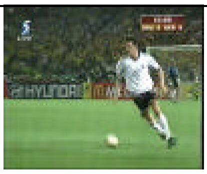
	Sample Frame	Ideal Labels
Close up (CP)		O O O O O O O O B B O O O O B B O O O O B B O O
Audience (AD)		A A A A A A A A A A A A A A A A A A A A A A A A
Far Penalty (FP)		A A A A A A A G G O O A G G G G G A G G G G G G
Far Middle (FM)		A A A A A A O O O O O O G G G G G G G G G B G G
Player		G G G G G G G G B B G G G G B B G G G G B B G G
Mid Body (MB)		A A A A A A A A A B B O O G G B B O G G G B B G

Figure 6.1 The relationships between the Block Label set and the Semantic Descriptors

In this section, we will separately introduce the local analysis and frame analysis algorithms for each P frame adopted by this approach, followed by the rules used to combine the analysis results to infer the final Semantic Descriptors for each P frame. Finally, algorithms about segmentation in each active part are discussed.

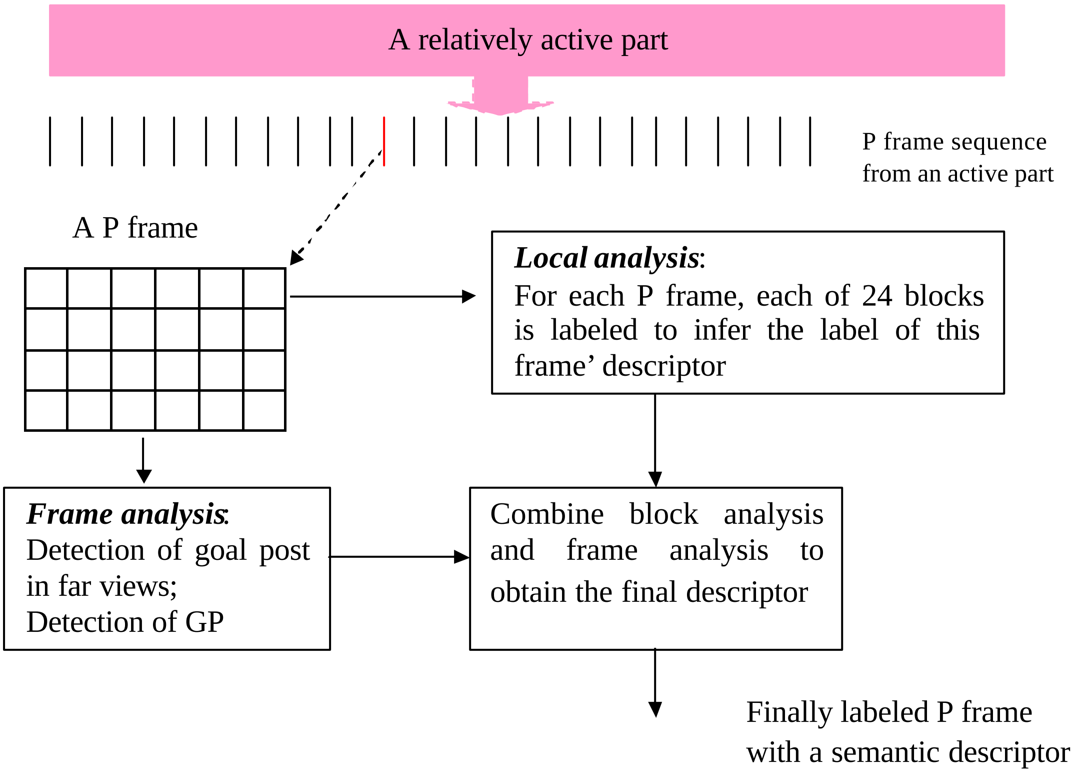

A relatively active part
P frame sequence from an active part
A P frame
Local analysis: For each P frame, each of 24 blocks is labeled to infer the label of this frame' descriptor
Frame analysis: Detection of goal post in far views; Detection of GP
Combine block analysis and frame analysis to obtain the final descriptor
Finally labeled P frame with a semantic descriptor

Figure 6.2 Classification steps to each P frame

6.3.1 Block Level Analysis

As we introduced before, there are four components in our Block Label Set: *A (Audience), G (Ground), B (Body)* and *O (Other)*. This approach extracts features such color, motion and texture from each of 24 blocks of each P frame. So, this is a local or regional processing. Before we discuss how to use these features, we review the extraction of edge and texture features first.

I Edge Detection --- SOBLE Edge Detector

Edge detection is a problem of fundamental importance in image analysis. In typical images, edges characterize object boundaries and are therefore useful for segmentation, registration, and identification of objects in a scene. In this section, the construction, characteristics, and performance of a number of gradient and SOBEL edge operator will be presented.

An edge is a jump in intensity. The cross section of an edge has the shape of a ramp. An ideal edge is a discontinuity (*i.e.*, a ramp with an infinite slope). The first derivative assumes a local maximum at an edge. For a continuous image $f(x, y)$, where *x* and *y* are the row and column coordinates respectively, we typically consider the two directional derivatives $\partial_x f(x, y)$ and $\partial_y f(x, y)$. Of particular interest in edge detection are two functions that can be expressed in terms of these directional derivatives: the gradient magnitude and the gradient orientation. The gradient magnitude is defined as

$$|\nabla f(x,y)| = \sqrt{(\partial_x f(x,y))^2 + (\partial_y f(x,y))^2} \qquad (6.1)$$

and the gradient orientation is given by

$$\angle \nabla f(x,y) = ArcTan(\partial_y f(x,y) / \partial_x f(x,y)) \qquad (6.2)$$

In theory at least, the operator consists of a pair of 3× 3 convolution kernels as shown in Figure 6.3. One kernel is simply the other rotated by 90°.

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Figure 6.3 Sobel convolution kernels for edge detection

These kernels are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid, one kernel for each of the two perpendicular orientations. The kernels can be applied separately to the input image, to produce separate measurements of the gradient component in each orientation (call these *Gx* and *Gy*). These can then be combined together to find the absolute magnitude of the gradient at each point and the orientation of that gradient. The gradient magnitude is given by:

$$|G| = \sqrt{G_x^2 + G_y^2} \qquad (6.3)$$

Typically, an approximate magnitude is computed using:

$$|G| = |Gx| + |Gy| \qquad (6.4)$$

which is much faster to compute. The angle of orientation of the edge (relative to the pixel grid) giving rise to the spatial gradient is given by:

$$q = \arctan(G_y / G_x) \qquad (6.5)$$

II Texture ---- Edge Density and Direction

Since edge detection is a well-known and simple-to-apply feature detection scheme, it is natural to try to use an edge detector as the first step in texture analysis. The number of edge pixels in a given fixed-size region gives some indication of the busyness of that region. Support that a gradient-based edge detector is applied to a region of N pixels, in which producing two outputs for each pixel. We use two formulae to measure the gradient magnitude *Mag* (p) and the gradient direction *Dir*(p):

$$F_{edgeness} = \frac{|\{ P \mid Mag(p) >= T \}|}{N} \tag{6.6}$$

$$F_{magdir} = (H_{mag}(R), H_{dir}(R)) \tag{6.7}$$

We also used SVM as our classifier. Introduction to SVM can be found in Appendix.

III Procedure of Block analysis

For each block of a P frame, the system extracts the following features as shown in Table 6.2. According to what are described in this table, the approach can automatically extract the following features:

$F_{edgeness}$, F_{magdir}, Motion $_{mag}$, Motion $_{dir}$, Colors, field color /non-field color

In practice, we both used a decision tree and SVM (linear) to label blocks. The steps are shown below:

i) In a block, if the colors are not rich & Field color is dominant, then it's labeled as 'Ground';

ii) else if its 'TD' is high & 'Motion' shows either stillness or movement in one direction, then it's labeled as 'Audience';

iii) else if its colors are not rich & non-field colors are dominant & motion indicates movement in different directions, it's labeled as 'Body';

iv) otherwise, it's labeled as 'Other', which means we cannot make sure its semantic meaning.

SVM is again used to decide the motion model in this approach. In order to reach this purpose, the motion magnitude and direction of each macro-block within a block is calculated. Also, the mean and standard deviations of all macroblocks from a block are computed. Just as we did in the unit-based approach, the direction of a motion vector is quantified to one of 8 directions as shown in Figure 5.10.

Table 6.2 Feature descriptions for block labels

Block Labels	*Feature Description*		
Audience	Texture Density		High
	Color		Colors are rich
	Motion	Magnitude	Large or small
		Direction	In one direction or still
Ground	Texture Density (TD)		Low or high
	Color		Colors are not rich and field colors are dominant
	Motion	Magnitude	Large or small
		Direction	In one direction or still
Body	Texture Density		Low
	Color		Colors are not rich and non-field colors
	Motion	Magnitude	Large
		Direction	Random
Other	Means the type of this block can not be clearly decided		

8-neighborhood is used in computation of motion vector direction frequency (MVD Frequency). That is, as shown in Figure 6.4, we want to calculate the MVD Frequency of a block marked as 1; the system computes the distribution of all motion vectors in this block as well as its 8 neighbors.

	2	2	2		
	2	1	2		
	2	2	2		

Figure 6.4 An example to show how to compute MVD Frequency

6.3.2 Frame Level Analysis

Figure 6.2 shows that the classification is based on both local analysis and frame analysis. The former has just been discussed in Section 6.3.1; from now on, we start to introduce the frame analysis. Since Hough Transform line detection is used in the frame analysis, its theory is briefly reviewed firstly, then followed by the method for Close-up view of Goal Post (GP).

I Hough Transform Line Detection

The Hough technique is particularly useful for computing a global description of a feature(s) (where the number of solution classes need not be known *a priori*), given (possibly noisy) local measurements. The motivating idea behind the Hough technique for line detection is that each input measurement (*e.g.* coordinate point) indicates its contribution to a globally consistent solution (*e.g.* the physical line which gave rise to that image point).

We can analytically describe a line segment in a number of forms. However, a convenient equation for describing a set of lines uses *parametric* or *normal* notion:

$$x\cos q + y\sin q = r \qquad (6.8)$$

where $\boldsymbol{r}$ is the length of a normal from the origin to this line and q is the orientation of $\boldsymbol{r}$ with respect to the X-axis (See Figure 6.5). For any point (x, y) on this line, the $\boldsymbol{r}$ and q are constants.

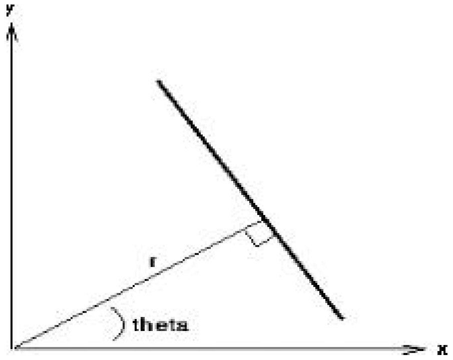


Figure 6.5 Parameter description of a straight line for Hough Transform Line Detection

In an image analysis context, the coordinates of the point(s) of edge segments (*i.e.* $(\mathbf{x_i}, \mathbf{y_i})$) in the image are known and therefore serve as constants in the parametric line equation, while $\boldsymbol{r}$ and q are the unknown variables we seek. If we plot the possible $(\boldsymbol{r}, q)$ values defined by each $(\mathbf{x_i}, \mathbf{y_i})$, points in Cartesian image space map to curves (*i.e.* sinusoids) in the polar Hough parameter space. This *point-to-curve* transformation is the Hough transformation for straight lines. When viewed in Hough parameter space, points, which are collinear in the Cartesian image space, become readily apparent as they yield curves, which intersect at a common $(\boldsymbol{r}, q)$ point.

The transform is implemented by quantizing the Hough parameter space into finite intervals or *accumulator cells*. As the algorithm runs, each $(\mathbf{x_i}, \mathbf{y_i})$ is transformed into a discretized $(\boldsymbol{r}, q)$ curve and the accumulator cells, which lie along this curve, are incremented. Resulting peaks in the accumulator array represent strong evidence that a corresponding straight line exists in the image.

II Procedure of frame analysis

We hope we can detect the far view of goal post in a P frame to pick FP (far view of penalty) frame sequences from others by using line detection. The steps for detection are presented below.

i) Use Hough Transform to detect 3 white parallel lines in a P frame; if failed, then there is no goal post in far view in this frame;

ii) Otherwise, detect the two white posts above the leftmost or rightmost white line; if failed, then there is no goal post in far view in this frame;

iii) Otherwise, we can claim a success of finding a goal post in far view in this frame

We also wish to know if a P frame contains a close-up view of a goal post (GP). The method used to detect GP is also given below:

i) Use domain knowledge to detect goal post or cross bar in close-up view in each P frame. The result is A;

ii) Use edge detector to detect goal net in each P frame with help of SVM. The result is B;

iii) If A or B is ‘Yes’, then we can claim we find a goal post or cross bar in close-up view;

iv) Otherwise, there is no a goal post or cross bar in close-up view detected in this frame.

Y. Gong et al. in [65] defined some patterns to detect penalty box. In [1], A. Ekin et al. detected the three parallel field lines to detect penalty box. The results are better than those in [14] by D. Yow et al., who used the similar method to do this task. Our method is similar to theirs and has satisfying results (94.7% on average) as shown in Table 7.5.

As for detection of goal post in close-up view, goal net and goal post or cross bar are detected. In order to detect goal net in P frames, edge directions and magnitude are extracted and feed into SVM (Linear), which is again used as our classifier in this method. For detection of goal post or cross bar, domain knowledge is used. That is, if the width of a white line is more than 15 pixels with length of more than 30 pixels, it is considered as a cross bar. A similar rule is used to detect goal post in close-up view. The results are also shown in Table 7.4 in Section 7.3.

6.3.3 Issues about Segmentation for Each Active Part

As we have already known that the classification is performed on each P frame of each active part. After doing this, each active part is actually segmented by P frames. This is illustrated in Figure 6.6.

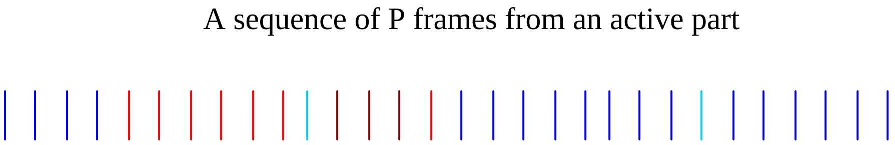


Figure 6.6 A typical P frame sequence from an active part

In Figure 6.6, frames with the same color are labeled with the same Semantic Descriptor.

In order to locate each segment's boundary, a buffer array is used in practice. In this array, the labels or descriptors of four consecutive P frames are kept; a flag is used to indicate the current label type. Only if at least three of the four elements in the buffer are different from the flag, a boundary change is claimed to be successfully detected; otherwise, the label type of the next P frame is used under consideration with the labels in the buffer. We have to mention that we do not need exact boundaries because it is not necessary.

After the integrated classification and segmentation step, every static part needs to be combined with their neighbors to complete the analysis procedure in the post-processing stage, as introduced in Section 4.4. After the post-processing stage, a soccer video is represented by a sequence of semantic descriptors.

Chapter 7

Experimental Results and the Evaluation

The methodology we designed for semantic soccer video analysis has been completely introduced in previous chapters. In this chapter, we discuss the experimental results from the tests of the two approaches. We first describe the data set in Section 7.1. In practice, we have always utilized part of data set to test algorithms to have a direct idea if they are effective. These algorithms will be adopted only if the results are promising. In Section 7.2, experimental results from evaluation of unit segmentation and classification algorithms in the unit-base approach are given, followed by the results from classification stage in the frame-based approach in Section 7.3. Then, we explain how ground truth is defined followed by the testing results and their evaluation. Finally, the uniqueness of this proposed representation among other research work is discussed.

7.1 Data Set

We chose two games from FIFA World 98 and two from FIFA World Cup 2002 as our training data set. For evaluation of an algorithm, both the training data and the testing data were selected from this set. For example when evaluating the effectiveness of the unit segmentation algorithm in the unit-based approach, all games in this set were manually segmented and used as the training data. Also 4 10-minute video clips were selected from it as the testing data to evaluate this algorithm.

5 soccer games, a total of 450 minutes of soccer video without commercial, from the FIFA World Cup 2002 were used as our testing data set to test the implemented approaches, including two games played in the afternoon and the other three in the evening. They are listed in Table 7.1. We chose the games played at different time because we want to see if our method to detect field colors is effective and robust when fields are under different light conditions.

Table 7.1 Training data (a) and testing data (b)

a. Four games as the training data set

FIFA 2002	Mexico VS USA (afternoon)
	Spain VS Iran (evening)
FIFA 1998	Paraguay VS France (afternoon)
	Netherlands VS Brazil (evening)

b. Five games selected from FIFA 2002 as our test set

Played In the Afternoon	*Played In the Evening*
England VS Brazil	Brazil VS Germany
Korea VS Spain	USA VS Germany
	Korea VS Turkey

7.2 Test Results for Algorithms Used in the Unit-based Approach

In the unit-based approach, the unit segmentation and classification were performed in two separate steps. Before integrating them, we wish to know if the algorithms for each stage can work well. Here, we present test results to support our segmentation and classification methods.

7.2.1 Results for Unit Segmentation

We mentioned in section 4.2 and Table 4.1 that we used three tables, namely Green Color Table (GCT), Upper Green Table (UGT) and Lower Green Table (LGT), as references to decide if a pixel is a field color. We defined 4 view types according to camera shooting positions and ratio of field colors to non-field colors within one frame in Section 3.1 (Figure 3.5) to segment all active parts. Please see Section 5.3.1 for details.

In order to know if this segmentation algorithm for the unit-based method is effective, we manually segmented all the 4 soccer games from the training data set, a total of 360 minutes, into units and labeled each unit according to the defined semantic descriptors in Table 3.3. The results are shown in Table 7.2 a.

Then 4 10-minute video clips (one clip from each game) were utilized to test the algorithm to evaluate its robustness, and the results are shown in Table 7.2, b.

Table 7.2 The testing results from our segmentation algorithm

a. Testing results by using manually segmented video clips

	No Field	Part Field	Full Field	Field with Player
Test Samples	110	211	95	133
Correct	103	193	88	122
Percent (%)	93.6	91.5	92.6	91.7

b. Testing results by using 4 10-minute long video clips

	Ground Truth	Output	Correct	Missed	Accuracy (%)
No Field	27	29	26	1	89.7
Part Field	59	59	52	7	88.1
Full Field	32	33	29	3	87.9
Field with Player	46	43	38	8	88.4

In the unit-based approach, this segmentation step is the foundation of next step, classification. Our experimental results (Table 7.2) show that green / non-green frames and color numbers are adequate to obtain satisfactory segmentation result.

7.2.2 Results for Unit Classification

In order to have a clear idea if our classification method to be used in the unit-based approach can work properly, we manually segmented all the games in the evaluation data set and used the first halves of the four soccer videos as the training data and the second halves as the test data. Support vector machine ([69], linear with the 'multi-classify' option) was adopted as the classifier.

Table 7.3 Test results from the unit classification algorithm

View Types		Accuracy
No Field	AD / CP	85.7%
Part Field	FM / Others	79.1%
	FP / Others	81.2%
	MB / Others	70.1%
	CP / Others	73.0%
Full Field	FM / MB	93.1%
Field with Player	MB / other	78.4%

Means and standard deviations of motion vectors' magnitudes and angles of motion vectors as well as the direction frequency are extracted as the features (Please see Section 5.3.2 for details).

From Table 7.3, we can see that it is not easy to recognize MB segments from others, because the motion patterns between each two of ‘FM / MB’ and ‘FP / MB’ are not discriminative enough. And also, replays may also affect the results. For example, given a frame showing a standing player in the field with lots of other players’ legs at the upper part of this frame, it is possible to be labeled as FP. Although there are some shortcomings, our results on 333 test segments (Table 7.3) are acceptable and thus indicate that the method is promising.

Since the classification is done after the unit segmentation, which means its accuracy depends largely on the segmentation results, here we have used manually segmented video to evaluate this classification method. The results can reflect the effectiveness of this method.

7.3 Results for Algorithms used in the Frame-based Approach

In order to test our method for goal post detection in close-up view, 79 video segments, including 22 among them as ground truth and 57 from other types of segments, were manually segmented from the evaluation data set to test the adopted method. Also, we segmented 236 clips from the same data set, which comprise 102 for goal post in far view and 134 for other types of clips, to test our algorithm for goal post detection in far view in the frame-based approach. As we always did, to see if the analysis on block level can work well, we manually extracted features from 2400 blocks. Here one thing must be pointed out: in the frame-based approach, the analysis is based on every P frame or block, so, using manually segmented data to evaluate the algorithms is enough. Linear SVM is used as the classifier. Please refer to Section 6.3.2 II for goal post detection in both close-up view and far view and Section 6.3.1 III for labeling of blocks.

The experimental results are respectively presented in Table 7.4, Table 7.5 and Table 7.6. From these results we can conclude that the algorithms can work effectively.

Table 7.4 Experimental results for GP detection in close-up view

	Clips	Misclassified	Accuracy	Overall accuracy
Close-up of goal post	22	3	86.4%	65/77=84.4%
Others	57	11	80.7%	

Table 7.5 Experimental results for detection of goal post in far view

	Clips	Misclassified	Accuracy	Overall accuracy
Goal post in far view	102	6	94.1%	221/236=93.6%
Others	134	9	93.3%	

Table 7.6 Results for block classification in the frame-based approach

	Ground Truth	Output	Correct	Missed	Accuracy (%)
Audience	530	472	417	113	88.3
Ground	960	926	843	117	91.0
Body	910	885	776	134	87.7

7.4 Results for the two Approaches and their Evaluation

7.4.1 Definition of Ground Truth

How to define the ground truth for testing is a tough question we had faced because we did not use physically segmented shot as the unit in both of our approaches. We manually segmented all the five games according to the seven semantic descriptors as our ground truth. The experimental results are shown in Table 7.8 and Table 7.9. In order to compare the output with the ground truth, we defined some rules:

i. The descriptor for the output segment is the same as that for the segment as ground truth

ii. $(L_{outputseg} – L_{groundtruthseg}) / L_{groundtruthseg} < 10\%$

iii. $(S_{outputseg} – S_{groundtruthseg}) / S_{groundtruthseg} < 15\%$

iv. $(E_{outputseg} – E_{groundtruthseg}) / E_{groundtruthseg} < 15\%$

where for two segments (one is our ground truth segment and the other is output segment), the meanings of the terms in these rules are explained in Table 7.7.

Table 7.7 Explanation of terms in the rules to define ground truth

	Meanings
$L_{outputseg}$	The Length of the output segment
$L_{groundtruthseg}$	The Length of the relevant ground truth segment
$S_{outputseg}$	The start frame number of the output segment
$S_{groundtruthseg}$	The start frame number of the relevant ground truth segment
$E_{outputseg}$	The end frame number of the output segment
$E_{groundtruthseg}$	The end frame number of the relevant ground truth segment

7.4.2 Test Results

The test results for both of the approaches are listed in Table 7.8 and Table 7.9. Data in ‘Output’ are the detection results. Column ’Correct’ shows cases that are both detected and classified successfully.

Table 7.8 Experimental results from the unit-based approach

	Ground Truth	Output	Correct	Missed	Accuracy (%)
AD	62	69	56	6	81.2
FM	659	684	526	133	76.9
FP	507	482	349	158	72.4
MB	314	285	202	112	70.9
CP	345	414	335	10	80.9
GP	53	59	45	8	76.3
Player	296	247	182	114	73.7

The results for recognition of AD and CP are 81.1% on average in the unit-based approach, while the accuracy of detection of AD, CP, FP and FM are 84.3% on average in the frame-based approach. This shows that the methods for recognizing AD, CP, FP and FM in soccer video are stable and effective. Of course, the percentage of ‘MB’, ‘GP’ and ‘Player’ are not as outstanding as others in both of the two tables. The method for detection of GP depends heavily on color, which is not stable sometimes when under different lighting or weather conditions. For MB and Players, the methods used in the two approaches to extract motion features are not effective enough. So, we ought to find other ways to improve the accuracy of detection ‘MB’, ‘GP’ and ‘Player’.

Table 7.9 Experimental results from the frame-based approach

	Ground Truth	Output	Correct	Missed	Accuracy (%)
AD	62	68	56	6	82.4
FM	659	671	584	75	87.0
FP	507	497	422	85	83.2
MB	314	286	218	96	76.2
CP	345	394	334	11	84.7
GP	53	59	46	7	77.9
Player	296	273	202	94	74.0

Since the frame-based approach adopts more complex and specific algorithms to analyze video streams, its processing is around 17 frames per second; on the contrary, the unit-based approach uses relatively simple algorithms, its processing speed is about 21 frames per second, which is nearly real-time.

Table 7.10 Processing speeds of the two approaches

	Unit based Method	Frame based method
Processing Speed (frames/sec)	21	17

As we mentioned, the unit-based approach adopts less domain knowledge while the frame-based one use much more domain knowledge. The intention is to see how far a generic approach can go and how much domain knowledge can help. From the experimental results we can see that a system using only basic domain knowledge (e.g. field color) can obtain a good result, e.g. the results for the unit-base approach is good. To obtain better performance, much more domain knowledge is necessary. This can be seen in the frame-based approach.

7.5 Uniqueness of the Proposed Representation

In Section 2.3, a brief review of existing work on soccer video analysis was given. From the review, we can see that each research work was done to solve certain analysis problem in soccer videos: e.g. P. Xu in [42] tried to do Play / Break detection; L.Y. Duan in [35] classified shots, which includes certain events (e.g. corner kick detection); A. Ekin in [1] provided an effective system to do summarization.

However, their efforts have focused on predefined structural events in soccer videos and ignore the importance of intermediate representation. On the contrary, since our research purpose is to find an effective mid-level representation for high-level soccer video analysis, the semantic gap between low-level features and semantic meanings can be bridged more easily based on our representation. For instance, we can use this mid-level representation for a soccer game to detect play / break in this game. Moreover, we have demonstrated that semantic units are more appropriate segment representation for soccer video analysis. Last but not least, this proposed framework can be utilized to analyze other kinds of sports videos as presented in [43].

Chapter 8

Conclusion and Future Work

8.1 Conclusion

In this thesis, a novel mid-level representation for soccer video parsing was introduced for bridging the gap between low-level features such as color, motion and texture and semantic meanings. Two computational approaches (the unit-based and the frame-based) to realize this method were discussed in detail.

The main idea of our task is to segment, classify, label a soccer video stream and therefore convert it into a sequence of well-defined semantically meaningful descriptors as the representation of this video stream for the further analysis such as event detection.

In order to introduce our method, we first discussed the shortcomings of using shot as an intermediate representation for soccer video analysis in Chapter 3. Then, the definitions of 7 semantic descriptors (close-up view, audience, far view of whole field, far view of penalty box, goal post in close-up view, player/players and mid-range view) were given. With these definitions, a soccer video stream can be represented by a descriptor sequence after segmentation and classification.

In Chapter 4, we summarized the two proposed unit-based and frame-based approaches. There were three stages in each approach, namely pre-processing stage, processing stage and post-processing stage. Because the pre-processing and post-processing stages in both of the approaches were very similar, they were also introduced in this chapter and were not repeated in the latter chapters.

The unit-based method was presented in detail in Chapter 5. In this method, video stream was first divided into static parts and active parts by motion magnitude, and then dominant color was used to segment each active part. SVM acted as the classifier to classify segments before the mergence of static parts with classified segment.

In the frame-base approach, SVM was again used to classify each of 24 blocks of one P frame coming from active parts. This process was based on a block level. In the meanwhile, domain knowledge was used to detect goal post in the whole P frame. Then, combining the analysis on both block level and frame level, the P frame was labeled with one of predefined semantic descriptors. Consecutive P frames with the same label was considered as a segment, and a buffer-based method was applied to look for boundaries for each segment. The representation was finally finished after post-processing stage. Details can be found in Chapter 6.

In Chapter 7, the test results were listed. We selected five games (or, 450 minutes) from the FIFA World Cup 2002 as our experimental data to test both of the approaches. For the first approach, the average accuracy is 76.1% while for the second one, that is 81%. Especially, detection in ‘AD’, ‘FM’, ‘FP’ and ‘CP’ in both of the approaches got better results than that in ‘GP’, ‘MB’ and ‘Player’. This means further study is necessary for detecting these three descriptors. The reasonable results show that our proposed method for mid-level representation of soccer video is effective.

We conclude that while the unit-based approach adopts less domain knowledge for exploring a relatively generic method which can be used to analyze other types of sports video, the frame-based one uses much more domain knowledge to provide an effective analysis for only soccer video. As we discussed in Section7.4.2, using only basic domain knowledge (e.g. the unit-based approach uses the field color) can obtain a good result. To obtain better performance, much more domain knowledge is necessary, as demonstrated by the frame-based approach.

8.2 Generality of the Proposed Mid-level Representation

We have discussed the uniqueness of the proposed mid-level representation among other research work in Section 7.4. Here the generality of this representation is discussed.

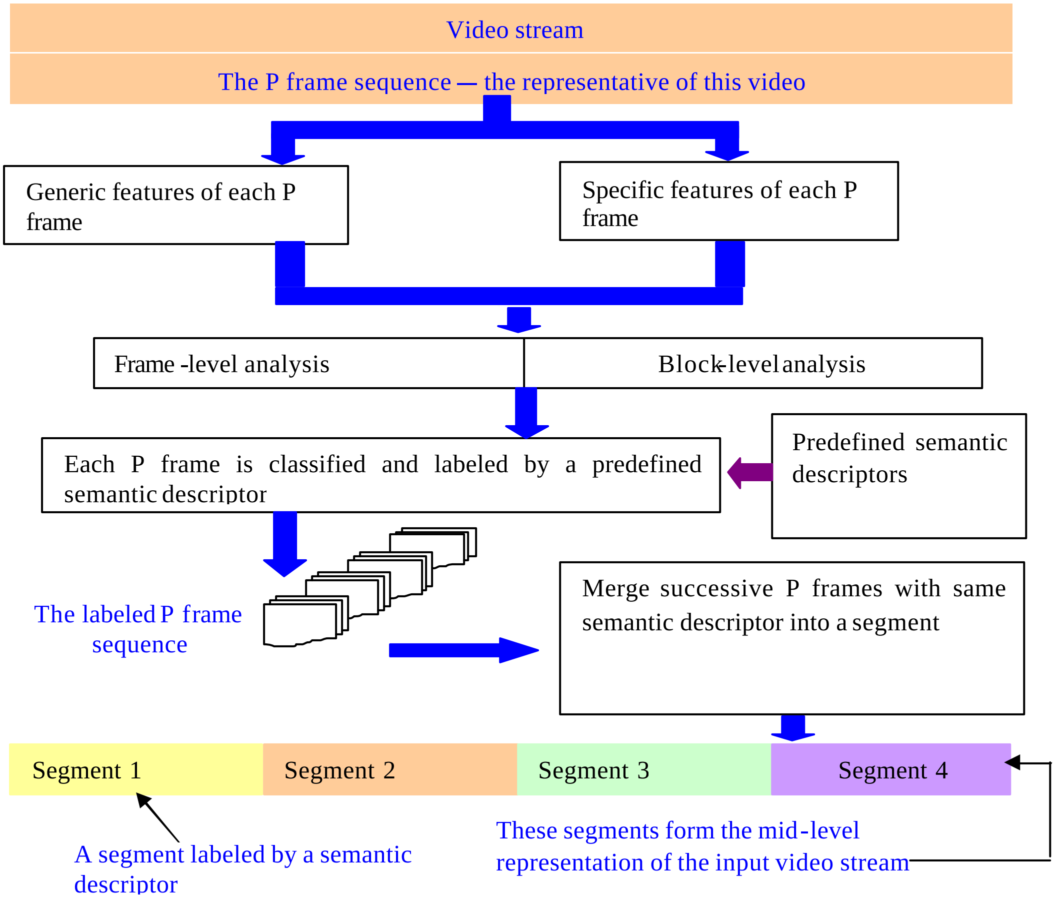

Video stream
The P frame sequence — the representative of this video
Generic features of each P frame
Specific features of each P frame
Frame -level analysis
Block-level analysis
Each P frame is classified and labeled by a predefined semantic descriptor
Predefined semantic descriptors
The labeled P frame sequence
Merge successive P frames with same semantic descriptor into a segment
Segment 1
Segment 2
Segment 3
Segment 4
A segment labeled by a semantic descriptor
These segments form the mid-level representation of the input video stream

Figure 8.1 A generic mid-level representation for efficient semantic video analysis

Although the mid-level representation was mainly introduced for soccer video analysis, this intermediate representation scheme can be used to parse other video streams. In [43], the authors have adopted this method for the analysis of tennis videos and news videos. The steps they used are illustrated and explained below. Figure 8.1 is reproduced from [43].

i. Predefined semantic descriptors (SD) for each kind of videos are defined according to their structure and the needs for analysis such as event detection in these videos;

ii. P frame sequence is the representative of an input video and the analysis shall be based on P frames rather than shots;

iii. Each P frame is partitioned into a *m* (row) by *n* (column) grid, and analyses based on both frame and block basis are performed on all P frames of the input video stream;

iv. Each P frame is classified into predefined categories, each of which is labeled by a SD to indicate its semantic meaning. Hence, the video stream is represented by a set of labeled P frames;

v. Merging process is performed in this set so that successive P frames with the same SD are gathered into the same segment. Hence, the video stream is converted into a set of semantically labeled segments.

The experimental results in [43] indicate that our proposed method is a generic and effective method for analysis of different kinds of sports videos, not limited to soccer videos.

The contributions of our research work are listed below:

- A new method for mid-level representation in soccer video analysis to bridge the gap between low-level features and semantic understanding instead of using shot as the intermediate representation has been developed and tested. This method can also be used to analyze other kinds of sports videos.

8.3 Future Work

Several areas that may be promising for future research is described in the following:

- How to further improve the performances for both the approaches by using more generic features for the unit-based approach and by using more domain knowledge and features for the frame-based approach.
- In chapter 5, the computation of the initial dominant color statistics was based on the ratio of dominant color pixels in the training set that was input by a human operator. Although this did not pose any problem for the applications because it can be completed before the start of the game, automatic computation of thresholds for dominant color region detection should be considered for our future study.

  Similarly, the thresholds used in the two approaches are set manually. Approaches using automatically adjustable thresholds need to be explored.
- Yu [61] proposed a novel trajectory-based algorithm for automatically detecting and tracking the ball in broadcast soccer video. We aim to extend our method

using motion trajectories and shapes as low-level evidence in addition to color and texture.

- The research on finding relationships between audio and video features for soccer video analysis is a promising avenue. For example, [37] developed an automatic whistling detection approach for the soccer video. Furthermore, the relationship between an audio peak and semantic features will be investigated.
- Extension of the proposed approach to different sports, such as American football, basketball, and baseball, which require different event and object detection modules, will be addressed in the future.

References

[1] A. Ekin, A. M. Tekalp, and R. Mehrotra, “Automatic Soccer Video Analysis and Summarization”, IEEE Trans. on Image Processing, Vol.12, pp796-807, July 2003

[2] A. Ekin and A. M. Tekalp, ”Robust Dominant Color Region Detection with Applications to Sports Video”, submitted to Comp. Vision & Image Understanding.

[3] A. M. Ferman, S. Krishnamachari, A. Murat Tekalp, “Group-of-frames/pictures Color Histogram Descriptors for Multimedia Applications”, IEEE International Conf. On Image Processing, Oct. 2000

[4] A. M. Ferman and A. M. Tekalp, “ A Fuzzy Framework for Unsupervised Video Content Characterization and Shot Classification,” International Journal of Electronic Imaging, vol. 10, no. 4, pp. 917-929, Oct. 2001

[5] B. Clarkson and *A.* Pentland, “Unsupervised Clustering of Ambulatory Audio and Video”, International Conf. On Acoustics, Speech, and Signal Processing, vol. 6, Page(s): 3037 -3040,1999

[6] B. L. Yeo and B. Liu, “Unified Approach to Temporal Segmentation of Motion JPEG and MPEG video,” International Conf. on Multimedia Computing and systems, pp. 2-13, 1995

[7] C.G.M. Snoek and M. Worring, “A State-of-the-art Review on Multimodal Video Indexing”, Proc. of the 8th Annual Conf. of the Advanced School for Computing and Imaging pages 194--202, Lochem, Netherlands, 2002

[8] C. Kim and J.N. Hwang, "Fast and Robust Moving Object Segmentation in Video Sequences", IEEE international Conf. on Image Processing, Kobe, Japan, Oct. 1999

[9] C. Kim and J-N Hwang, "Object-Based Video Abstraction and an Integrated Scheme for On-line Processing", IEEE Trans. on Circuits and Systems for Video Technology (CSVT). Oct. 2000

[10] C. Kim and J. N. Hwang, "Object-Based Video Abstraction Using Cluster Analysis", IEEE International Conf. on Image Processing, Greece, Oct. 2001

[11] C. W. Ngo, T. C. Pong and R. T. Chin, "Detection of Gradual Transitions through Temporal Slice Analysis", IEEE Conf. on Computer Vision and Pattern Recognition, June, 1999

[12] C. W. Ngo, T. C. Pong, and R. T. Chin. "A Robust Wipe Detection Algorithm", Asian Conference on Computer Vision, 1:246-251, 2000

[13] C.W. Ngo, T.C. Pong and H. J. Zhang, "Recent Advances in Content based Video Analysis", International Journal of Image and Graphics, 1(3):445-468, 2001

[14] D. Yow, B. L. Yeo, M. Yeung and B. Liu, "Analysis and Presentation of Soccer HighLights from Digital Video", Asian Conference on Computer Vision, pp. 499-503, 1995

[15] D. Zhong and S. F. Chang, "An Integrated Approach for Content-Based Video Object Segmentation and Retrieval", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1259-1268, Dec. 1999

[16] D. Zhong, and S. F. Chang, "Region Feature-based Similarity Searching of Semantic Video Objects", IEEE International Conf. on Image Processing, Volume: 2, Page(s): 111 -115 vol.2 , 1999

[17] E. Ardizzone, M. La Casica, D. Molinelli, “Motion and Color-Based Video Indexing and Retrieval”, International Conf. on Pattern Recognition, pp135-139, 1996

[18] E. Ardizzone, M. La Cascia, A. Avanzato, and A. Bruna, “Video Indexing Using MPEG Motion Compensation Vectors”, IEEE Conf. on Multimedia Computing and Systems pp.725-729, June 1999

[19] E. Sahouria and A. Zakhor, “Content Analysis of Video Using Principal Components”, IEEE Trans. on Circuits and Systems for Video Technology, 9(8):1290--1298, 1999

[20] G.Ahanger and T. D. C. Little, “A Survey of Technologies for Parsing and Indexing Digital Video”, International Journal of Visual Communication and Image Representation, 7(1):28-43, March 1996.

[21] G. Millerson, “The technique of television production”, 12th Ed., Focal Publishers, March 1990

[22] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, “Automatic Partitioning of Full-motion Video”, Multimedia Systems Journal, Vol. 1, No. 1, pp. 10-28, 1993

[23] H. J. Zhang, J.Y.A. Wang, and Y. Altunbasak, “Content-based Video Retrieval and Compression: A Unified Solution”, IEEE International Conf. on Image Processing, pp.13-16, 1997

[24] H. J. Zhang, C. Y. Low, S. W. Smoliar, and J. H. Wu, "Video Parsing, Retrieval and Browsing: An Integrated and Content-based Solution", ACM Multimedia'95, Nov 1995

[25] H. l. Wang, A. Divakaran, A. Vetro, S. F. Chang, H. F. Sun, “Survey of Compressed-Domain Features Used in Audio-Visual Indexing and Analysis”,

Journal of Visual Communication and Image Representation, 14(2):150-183, June 2003

[26] H. w. Kim and K. S. Hong, "Soccer Video Mosaicing using Self-Calibration and Line Tracking", International Conf. on Pattern Recognition, Barcelona, Spain, pp. 592-595, Sept., 2000

[27] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, "Extracting Semantic Information from News and Sport Video", Image and Signal Processing and Analysis, 2001

[28] J.H. Lim. "Building Visual Vocabulary for Image Indexation and Query Formulation", Pattern Analysis and Applications (Special Issue on Image Indexation), 4(2/3), 125-139, 2001

[29] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas "On Combining Classifiers", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 20(3), pp. 226-239,1998

[30] J. Meng, Y. Juan, and S. F. Chang, "Scene Change Detection in an MPEG Compressed Video Sequence," IS&T/SPIE Symposium Proceedings, Vol. 2419, Feb. 1995

[31] J. N Hwang, Y. Luo, "Automatic Object based Video Analysis and Interpretation: A Step toward systematic video understanding," invited special session talk in ICASSP, Orlando FL, May 2002

[32] J. R. Ohm, et al, " A Multi-Feature Description Scheme for Image and Video Database Retrieval", IEEE Workshop on Multimedia Signal Processing, pp.123-128, 1999

[33] J. S. Boreczky and L. A. Rowe, "Comparison of Video Shot Boundary Detection Techniques," SPIE Conf. on Storage and Retrieval for Image and Video Databases IV, pp. 170-179, Jan. 1996

[34] L. F. Cheong, "Scene-based Shot Change Detection and Comparative Evaluation", International Journal on Computer Vision and Image Understanding, 79(2):224-235, Aug., 2000

[35] L. Y. Duan, M. Xu, X. D. Yu, Q. Tian, "A Unified Framework for Semantic Shot Classification in Sports Video", ACM, Juan-les-Pins, France, Dec.2002

[36] M. Szummer, and R. W. Picard, 'Indoor-outdoor Image Classification' IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98

[37] M. S. Drew, S. N. Li and X. Zhong. "Video Dissolve and Wipe Detection via spatio-temporal Images of Chromatic Histogram Differences", International Conf. on Image Processing, 3:929-932, 2000

[38] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian, "Creating Audio Keywords for Event Detection in Soccer Video", IEEE International Conf. on Multimedia and Expo Vol II, 281-284, July, 2003

[39] N. Vasconcelos and A. Lippman. "Statistical Models of Video Structure for Content Analysis and Characterization", IEEE Trans. on Image Processing, 9(1):3-19, Jan. 2000

[40] O. Utsumi, K. Miura, I. Ide, S. Sakai, and H. Tanaka, "An Object Detection Method for Describing Soccer Game from Video", IEEE International Conf. on Multimedia and Expo, Aug., 2002

[41] P. Bouthemy, M. Gelgon, and F. Ganansia, “A Unified Approach to Shot Change Detection and Camera Motion Characterization”, IEEE Trans. on Circuits and Systems for Video Technology, 9(7):1030-1044, Oct. 1999

[42] P. Xu et al., “Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video”, IEEE International Conf. on Multimedia and Expo, Tokyo, Japan, Aug, 2001

[43] Q. Tang, H.P. Sun, J.H. Lim, Jesse Jin, Q. Tian, “ A Generic Mid-level Representation for Semantic Video Analysis”, Accepted by the Eleventh IEEE International Conf. on Image Processing (ICIP), Singapore, October 24-27, 2004.

[44] R. Brunelli, O. Mich, and C. M. Modena. “A Survey on the Automatic Indexing of Video Data”, International Journal on Visual Communication and Image Representation, 10:78-112, 1999

[45] R.A. Joyce and B. Liu. “Temporal Segmentation of Video Using Frame and Histogram Space”, International Conf. on Image Processing, 2000

[46] R. Lienhart, “Comparison of Automatic Shout Boundary Detection Algorithms,” SPIE Conf. on Storage and Retrieval for Image and Video Databases VII, pp. 290-301, Jan. 1999

[47] R. *Qian, N. Haering,* and I. *Sezan,* “ A Computational Approach to Semantic Event Detection”, IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, volume 1, pages 200--206, June 1999

[48] R. Zabih, J. Miller, and K. Mai, “ A Feature-based Algorithm for Detecting and Classifying Scene Breaks,” ACM Multimedia 95, pp. 189-200, Nov. 1995

[49] S. Dagtas, W. A1-Khatib, A. Ghafoor, and R.L. Kashyap, "Models for Motion-based Video Indexing and Retrieva l," IEEE Trans. on Image Processing, vol. 9, no. 1, pp. 88-101, Jan. 2000.

[50] M.R. Naphade and T.S. Huang, "Extracting Semantics from Audiovisual Content: The Final Frontier in Multimedia Retrieval", IEEE Trans. on Neural Networks, Vol. 13, Num 4, July 2002

[51] S. F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, "VideoQ: An Automated Content Based Video Search System Using Visual Cues", Proc. ACM Multimedia, pp.313-324, 1997

[52] S. F. Chang, and H. Sundaram, "Structural and Semantic Analysis of Video", IEEE International Conf. on Multimedia and Expo, Volume: 2, Page(s): 687 -690 vol.2, 2000

[53] S. F. Chang, W. chen, H. J. Meng, H. Sundaram and D. zhong. " A fully Automatic Content-based Video Search Engine Supporting Multi-object spatio-temporal Queries", IEEE Trans. on Circuits and Systems for Video Technology, 1998

[54] S. F. Chang, W. Chen, and H. Sundaram, "Semantic Visual Templates—Linking Features to Semantics", IEEE International Conf. on Image Processing, vol. 3, Chicago, IL, pp. 531–535., Oct 1998

[55] S. W. Lee, Y. M. Kim, and S. W. Choi, "Fast Scene Change Detection Using Direct Feature Extraction from MPEG Compressed Videos", IEEE Trans. on Multimedia, 2(4): 240-254, 2000

[56] T. Lin, and H.J. Zhang, "Automatic Video Scene Extraction by Shot Grouping", International Conf. on Pattern Recognition, 2000

[57] U. Gargi, R. Kasturi, and S.H. Strayer. "Performance Characterization of Video-shot-change Detection Methods", IEEE Trans. on Circuits and Systems for Video Technology, 10(1):1-13, Feb. 2000

[58] V. Kobla, D. DeMenthon, and D. Doermann, "Special Effect Edit Detection Using VideoTrails: a Comparison with Existing Techniques," SPIE Conf. on Storage and Retrieval for Image and Video Databases VII, 1999.

[59] V. Tovinkere and R. J. Qian, "Detecting Semantic Events in Soccer Games: Towards a Complete Solution", IEEE International Conf. on Multimedia and Expo, 2001

[60] W. Chen and S. Chang, "Generating Semantic Visual Templates for Video Databases," IEEE International Conf. on Multimedia Expo, vol. 3, pp. 1337–1340., July 2000

[61] W. Zhou, A. Vellaika, and C. C. Jay Kuo. "Rule-based Video Classification System for Basketball Video Indexing", International workshop on Multimedia Information Retrieval, 2000

[62] X. G. Yu, C. S. Xu, H. W. Leong, Q. Tian, Q. Tang and K. W. Wan, "Trajectory-Based Ball Detection and Tracking with Applications to Semantic Analysis of Broadcast Soccer Video", ACM Multimedia2003, Nov., 2003

[63] X.G. Yu, C.S. Xu, Q. Tian and H. W. Leong, "A Ball Tracking Framework for Broadcast Soccer Video", IEEE International Conf. on Multimedia and Expo, Vol II, 273-276, 2003

[64] Y. F. Ma and H.J. Zhang, "Motion Texture: A New Motion Based Video Representation", IEEE Conf. on Pattern Recognition, 2002

[65] Y. H. Gong, L. T. Sin, C. H. Chuan, H. J. Zhang, and M. Sakauchi, “Automatic Parsing of TV Soccer Programs”, IEEE International Conf. on Multimedia Computing and Systems, pp. 167-174, 1995

[66] Y. Iwai, J. Maruo, M. Yachida, T. Echigo and S. Iisaku, “A Framework of Visual Event Extraction from Soccer Games”, Asian Conf. on Computer Vision, vol. 1, pp. 222–227, 2000.

[67] 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, Dec., 2003

[68] Y. Wu, Y. Zhuang and Y. Pan. “Content-based Video Similarity Model”, ACM Multimedia, 2000

[69] LibSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[70] http://www.zc668.com/zcgnys.htm