NOTE TO USERS

This reproduction is the best copy available.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

INDEXING OF AMERICAN FOOTBALL VIDEO USING MPEG-7 DESCRIPTORS AND MFCC FEATURES

by

Syed G. Quadri
B.Eng., Ryerson University, Toronto, 2002

A thesis
presented to Ryerson University

in partial fulfillment of the
requirement for the degree of
Master of Applied Science
in the Program of
Electrical and Computer Engineering

Toronto, Ontario, Canada, 2004
©Syed G. Quadri 2004

PROPERTY OF
RYERSON UNIVERSITY LIBRARY

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

UMI Number: EC52933

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform EC52933

Copyright 2008 by ProQuest LLC.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Borrower's Page

Ryerson University requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Name	Signature	Address	Date

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

ABSTRACT

Indexing of American Football Video Using MPEG-7 descriptors and MFCC features

©Syed G. Quadri 2004

Master of Applied Science
Department of Electrical and Computer Engineering
Ryerson University

In this work, an application system is proposed to classify American Football Video shots. The application uses MPEG-7 motion and audio descriptors along with Mel Frequency Cepstrum Coefficient features to classify the video shots into 4 categories, namely: Pass plays, Run plays, Field Goal/Extra Point plays and Kickoff/Punt plays.

Fisher's Linear Discriminant Analysis is used to classify the 4 events, using a leave-one-out classification technique in order to minimize the sample set bias. For a database of 200 video shots taken from four different games, an overall system performance of 92.5% was recorded. In comparison to other American Football indexing systems, the proposed system performs 8% to 12% better.

We have also proposed an algorithm that uses MPEG-7 motion activity descriptors and mean of the motion vector magnitudes, in a collaborative manner to detect the starting point of play events within video shots. The algorithm can detect starting points of the play with 83% accuracy.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Acknowledgement

I would like to thank my supervisor Dr. Ling Guan and co-supervisor Dr. Sridhar Krishnan for their encouragement, guidance and continuous support throughout my research work and writing of this manuscript. This work would have been impossible without their feedback, patience and kindness.

I would also like to thank Canada Foundation for Innovation (CFI) and the Department of Electrical and Computer Engineering for providing a very well equipped and technically supported Ryerson Multimedia Laboratory. My thanks are due to the School of Graduate Studies of Ryerson University for providing Graduate Student Scholarship and helping me in securing Ontario Graduate Scholarship (OGS).

I would like to acknowledge my supervisors' funding resources National Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chair Program, for financial support to this research work.

My thanks are due to my manager in IBM Canada Ltd. for his unconditional support during my school years. I would like to thank my colleagues and members of the Ryerson Multimedia Laboratory for creating a friendly and congenial environment in the Lab. It was my pleasure to work with such a great team.

In the end I could not have achieved this goal of mine without the support of my parents, family and wife.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Contents

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Figures

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Tables

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 1

Introduction

1.1 General need for Indexing

THE increase in digital information has created a vital need for development of new techniques that can provide efficient access to information. Recently one of the areas for active research has been Indexing and Retrieval strategies in order to facilitate the access to the vast amount of information. Traditionally textual annotations are used for indexing of digital media, which is an extremely manual process and is prone to subjective bias. In contrast to textual annotations, there are indexing techniques based on content. These systems are known as content based indexing and retrieval systems (CBIR). Users can use query by example or query by sketch to retrieve the information from a multimedia database.

The CBIR systems can form the query based on low level features and retrieve multimedia objects of similar features. However, we as users are more interested in semantics. Low level features are an integral part of any query system, but mostly we are not interested in retrieving information with similar shape, color or texture. We relate objects to its meaning, therefore a modern indexing and retrieval system must provide meaningful or semantically coherent search results.

Appropriate semantic indexing of multimedia is a difficult task due to two main factors. First there is large amount of information present in all types of different modalities, such as audio, image and text. Secondly there are many different levels at which the information can contain semantic information. Additionally there are

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

issues with what to index and what not to index, as different users will search for different information and will have very different criteria definitions.

Design and development of an effective video indexing and retrieval system based on the content of the video poses some interesting challenges. Most people often think of video as a sequence of images, but in reality it is a medium with multiple media. Videos integrate the media presented by the images, graphics, text and audio. Therefore in order to index a video stream requires multiple strategies and a number of processing steps to deal with all the diverse media present. Thus video requires the content to be analyzed at a number of levels, namely lexical, syntactical and semantic. This could potentially mean that video first needs to be segmented into shots, the closed captioned text in the video stream has to be detected and recognized, the audio information needs to be extracted and interpreted, speech recognition may have to be used, visual information needs to be analyzed based on content and finally the shot needs to be annotated semantically [1].

In order to tackle the various issues of indexing a multimedia object, MPEG defined a standard known as "Multimedia Description Framework" or MPEG-7 with the primary goal of defining a scheme to describe a multimedia document by means of its content and the relationship between the different content sets within a multimedia document. MPEG-7 only standardized the description of content and not the algorithms utilized for feature extraction and application development. Therefore MPEG-7 can be used to develop new fast and efficient indexing and retrieval systems by utilizing the descriptors defined by the standard. These descriptors can be used to index a multimedia document at various levels, thus providing semantic retrieval tools.

1.2 Focus of this work

In this thesis a video indexing system is designed to automatically annotate National Football League (NFL) video shots. Every sport has many actions associated with it. But only a few fundamental events can describe the core of the game. The whole

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

game is built on these fundamental events. For example in basketball we can identify jump shots, free throws and turnovers as the three fundamental semantic events (FSE). Likewise in American football (NFL) we have proposed three FSE; namely pass plays, running plays and special team plays.

In the ADVENT project technical report [2], the authors generalize the concept of FSE by calling it Recurrent Visual Semantics (RVS). They define RVS as the repetitive appearance of elements that are visually similar and have common level of meaning within a specific context. The domain in which RVS events most commonly occur are, news video and sports video.

These RVS events or FSE point to the fact that for effective semantic indexing of multimedia content, we need to build a Knowledge Base of domain specific events which has the RVS events at the core. In the thesis we propose a knowledge base for NFL video shots. Figure 1.1 details the proposed NFL Knowledge Base as a hierarchical graph.

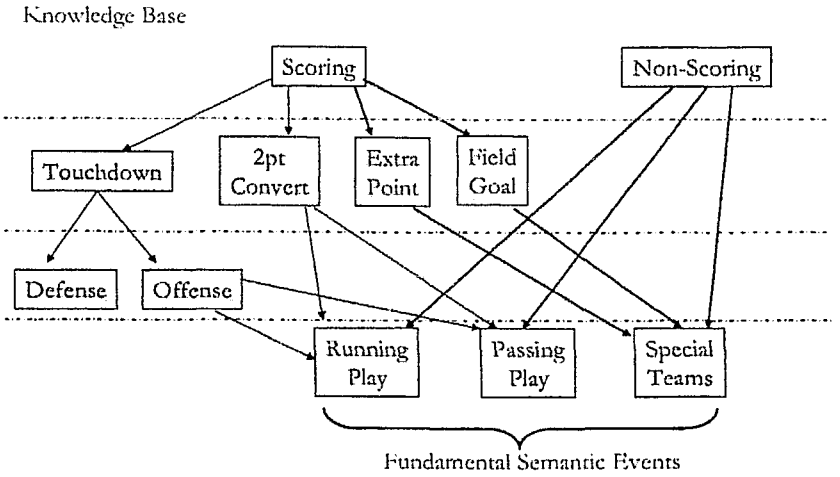

Figure 1.1: Knowledge Base of American Football

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The graph shown in Figure 1.1, is designed in such a way that the outermost nodes carry highest level of semantic value. The deeper the node, the more specific semantic information it carries. For example the events of touchdown and field goals carry the most specific semantic value while the events like scoring or non-scoring carry semantic value at a very high level. Much research has been done in identification of scoring vs non-scoring semantics [3]. In addition, some researchers have built systems to identify touchdowns, field goals and point after attempts [4]. The focus of this research has been to identify the RVS events such as pass, run or kick.

In the research community a lot of sports indexing techniques have been explored. Most of the indexing schemes rely on extracting low level visual and audio features which are passed to complex classifiers for identification. Since the finalization of the MPEG-7 standard some of the research activity has focused on using MPEG-7 descriptors as features for indexing news and sports video. For example in [5] MPEG-7 motion and audio features are utilized to summarize news video and sports video which included baseball, tennis and golf. Not much work has been done in the American Football domain. The work done in this domain mainly focuses on retrieval of scoring events from the sports video by extracting low level features such as audio, motion and closed caption text. In one of the works by Terry Caelli [6], spoken commentary along with player movement is utilized to detect different types of formations. All the works conducted in this domain have relied significantly on rule based classification schemes. This work will focus on utilizing MPEG-7 descriptors in order to index NFL video shots using a simple linear classifier.

One of the key issues with indexing of RVS events within a sports domain framework, is the localization of the start and end points of the event within a large collection of videos. The localization of the event helps in reducing the analysis window size and also by eliminating factors effecting the features which are not directly related to the action itself. In this work, I have proposed, MPEG-7 motion descriptor based technique to find the starting point of plays from NFL video shots.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

1.3 Organization of thesis

The remainder of this thesis consists of 5 chapters which are organized as follows:

Chapter 2: *MPEG-7 Concepts and Descriptors*, contains an overview of the MPEG-7 standard and a brief overview of audio and visual descriptors. It also contains a detailed explanation of the motion and audio descriptors that are utilized in this work.

Chapter 3: *Proposed Indexing System*, contains the overview of the proposed indexing system designed to index NFL video shots. Chapter 3 reviews other sports indexing techniques and provides the background for the proposed research activities.

Chapter 4: *Play start Detection*, details the proposed algorithm for localization of RVS action events within NFL video shots. It also presents the results of the algorithm and compares some other event localization efforts in the research community.

Chapter 5: *NFL video shot Indexing System*, details the proposed indexing system using MPEG-7 motion, audio descriptors and Mel Frequency Cepstrum Coefficients (MFCC). It summarizes the results of first using the features independently, and then by combining them together.

Chapter 6: *Conclusions and Future Work*, summarizes the results and discusses the advantages of using motion and audio descriptors. Some consideration is provided on how to enhance the work in the future.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 2

MPEG-7 Concepts and Descriptors

2.1 MPEG-7 Overview

MPEG-7 is a standard developed by the Moving Picture Experts Group (MPEG) via the standardization organizations of ISO/IEC. The formal name of the MPEG-7 standard is "Multimedia Content Description Interface" and is organized in 8 parts [7]. Parts 1 through 5 define the core of MPEG-7 technology, while parts 6 to 8 provide supporting information to the standard. This standard differs from its predecessors, such that the previous standards were concerned with the representation of the content while the objective of MPEG-7 is to standardize the information about the content. Thus MPEG-7 defines Descriptors and Description Schemes to represent the information in the multimedia document.

In MPEG-7 the descriptors represent all types of multimedia documents, such as Images, Graphics, 3D models, audio and video. It does not depend on how the multimedia document is created, coded or stored. Figure 2.1 shows the scope of the MPEG-7 standard.

The MPEG-7 standard consists of three main parts [8].

- **Description Tools:** Descriptors and Description Schemes make up the Description Tools. The Descriptors define the syntax and semantics of a feature, while the description scheme defines the relationship between descriptors and other description schemes.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- **Description Definition Language (DDL):** DDL defines the syntax of Description Tools in textual format. In MPEG-7 DDL is based on XML Schema Language. A detailed introduction on DDL can be found in Chapter 4 of [9]
- **System Tools:** These tools are used for management, synchronization, storage and transmission of descriptions. In MPEG-7 system tools support descriptions in both textual and binary formats.

Detailed introduction on any of the above parts can be found in [9]. Next section will provide more details on the descriptors that were utilized in the scope of this work.

2.2 MPEG-7 Descriptors

2.2.1 Visual Descriptors

These descriptors describe the basic multimedia document content based on visual information only. For example, in an image object, the content can be described based on the shape, texture and color media. Object motion and camera motion along with the above mentioned features can be used to describe video media. The main objective of the visual descriptors is to assist user applications in identification, categorization and filtering of images and videos.

The visual descriptors can be divided into general and domain specific descriptors. General visual descriptors are made up of shape, color, texture and motion features. There is only one domain specific descriptor; the face descriptor. A brief summary

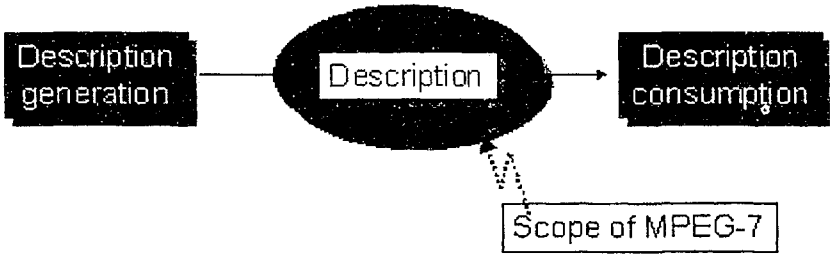

Figure 2.1: Scope of MPEG-7 Standard

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

of the general visual descriptors is given below. The details of only the motion descriptors is included in the document as currently they are the only ones being utilized by the proposed system.

- **Color Descriptors:** In image and video applications, color is the most commonly used feature. It is easily extracted and is some what immune to rotation, translation and viewing angle changes. There are seven (7) color descriptors defined by MPEG-7 as listed below:

  1. Color Space Descriptor
  2. Color Quantization Descriptor
  3. Dominant Color Descriptor
  4. Scalable Color Descriptor
  5. Group of Frames/Group of Pictures Descriptor
  6. Color Structure Descriptor
  7. Color Layout Descriptor

- **Texture Descriptors:** Texture defines the spatial distribution of patterns in an image. There is no universal definition of texture, but patterns in a region of an image can create an appearance of texture. In MPEG-7 three (3) texture descriptors are defined. as listed below:

  1. Homogenous Texture Descriptor
  2. Texture Browsing Descriptor
  3. Edge Histogram Descriptor

- **Shape Descriptors:** Shape is an important feature in image and video retrieval and object identification systems. Humans tend to associate semantics with the shape of objects. In MPEG-7 three (3) shape descriptors are defined as listed below:

  1. Region-based Shape Descriptor

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. Contour-based Shape Descriptor
3. 3-D Spectrum Shape Descriptor

- **Motion Descriptors:** Many different types of motion occur in a video segment. There is motion associated with objects with in the pictures and motion due to camera movements. In MPEG-7 four (4) descriptors are defined, as listed below, which cover all types of motion:

  1. Camera Motion Descriptor
  2. Motion Trajectory Descriptor
  3. Motion Activity Descriptor
  4. Parametric Motion Descriptor

Motion Activity Descriptors

The objective of the motion activity descriptor is to quantify the overall activity or pace of action in a video segment. We tend to perceive sports video segments as fast moving compared to news video segments. The activity descriptor is easily extracted from compressed domain, utilizing the encoded motion vectors. The descriptor utilizes the statistical properties of the motion vector magnitudes to measure intensity of motion activity. Following is the summary of attributes associated with the descriptor.

- **Intensity of Activity:** This attribute contains the global intensity of motion activity on a scale of 1 to 5. A high value indicates high activity.
- **Direction of Activity:** This attributes expresses the dominant direction in the video segment. This attribute classifies the direction into eight equally spaced directions as shown in Figure 2.2.
- **Spatial distribution of activity:** This attribute specifies the number and size of high activity regions within a frame. The attribute gives an indication whether the activity is spread across many regions or one large region.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

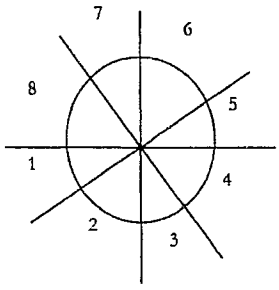

Figure 2.2: Direction of Activity

- **Temporal distribution of activity:** This attribute contains information regarding the duration of motion activity in a video segment. It specifies if the activity is confined to one part of the video or sustained throughout the segment.

The intensity of motion activity attribute is most commonly associated with the motion activity descriptor. The other three descriptors, dominant direction, spatial distribution and temporal distribution are optional attributes. In this work we will utilize both the intensity of motion and dominant direction attributes of the motion activity descriptor.

2.2.2 Audio Descriptors

In the standard, audio descriptors can be divided into two categories. One based on low level audio features designed for general use, while the other designed for application specific tasks. In the standard the generic audio tools are also known as Audio Description Framework. A detailed summary of the Audio description framework is given in the next sub section. Following is a brief summary of the low level and high level audio descriptors and description schemes.

- **Low Level Audio Descriptors:** There are eighteen (18) low level spectral

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and temporal audio descriptors defined in the standard. These descriptors can be categorized into the following seven groups:

1. Basic Descriptors
2. Basic Spectral Descriptors
3. Basic Signal parameter Descriptors
4. Temporal timbral Descriptors
5. Spectral timbral Descriptors
6. Spectral basis Descriptors
7. Silence Segment Descriptor

Figure 2.3 shows the summary of the groups along with the audio descriptors in the respective categories [8].

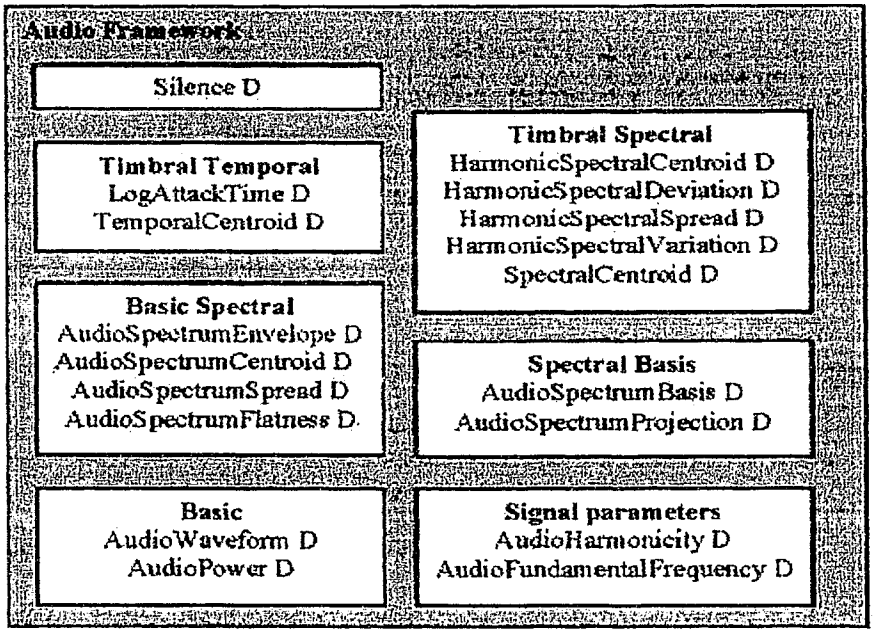

Figure 2.3: Summary of Low level Audio Descriptors

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- **High level Audio Descriptions:** There are four description schemes defined in the standard for application specific tasks. The list of the descriptors is as follows:

    1. General sound recognition and indexing tools
    2. Spoken Content description tools
    3. Musical instrument timbre description tool
    4. Melody description tool

Audio Description Framework

Audio description framework consists of eighteen (18) low level descriptors. The objective of the framework is to define a wide range of features that can be readily used to build general audio based applications. A brief summary of the audio descriptors in the framework is given below. Details on how to compute these descriptors is not included here, but Section 5.2.2 provides details on computing the descriptors utilized in this work.

1. **Audio Waveform:** This descriptor provides information about the envelope of the signal. It contains the minimum and maximum values of the signal within a specified window.

2. **Audio Power:** This descriptor provides information about the power of the signal. It describes the instantaneous power of the samples in the specified window.

3. **Audio Spectrum Envelope:** This descriptor provides information about spectral resolution of the logarithmic bands. The resolution can be controlled between 1/16 of an octave to 8 octaves. The descriptor can be computed by the Fast Fourier Transform of the specified window.

4. **Audio Spectrum Centroid:** This descriptor contains information about the distribution of the power spectrum. The descriptor can be regarded as an

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

approximation of the perceptual sharpness of the signal. It is calculated by adding all the energy within the frequency bins and dividing it by the total energy in the audio frame.

5. **Audio Spectrum Spread:** This descriptor contains information about the shape of the power spectrum. It indicates if the spectral content is concentrated at the centroid or if it is distributed across a range of the spectrum. It also has the property of discriminating between noise like and tonal sounds. It is calculated by taking the second moment of the log-frequency power spectrum of the signal.

6. **Audio Spectrum Flatness:** This descriptor contains information about the tonal components in each band of the signal. It is calculated by taking the ratio of the geometric mean over the arithmetic mean of spectral power within a band.

7. **Audio Fundamental Frequency:** This descriptor contains information regarding the musical pitch and periodic content of speech signals. Since calculation of fundamental frequency is not an exact science, the standard does not specify how the descriptor is calculated.

8. **Audio Harmonicity:** This descriptor contains information about the harmonic nature of signal spectrum. It contains two measures: the first, called the Harmonic Ratio, gives a measure of the proportion of harmonic components in the spectrum. The other measure called the Upper Limit of Harmonicity specifies the point in the spectrum beyond which no harmonic content is present.

9. **Log Attack Time:** This descriptor defines the time it takes a signal within a window to start and reach a sustained level or a maximum. The time is represented in logarithmic scale. This descriptor can be used to differentiate between a suddenly rising sound and a smoothly increasing sound.

10. **Temporal Centroid:** This descriptor calculates a time based centroid of the signal envelope for a particular segment. This descriptor provides time resolu-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

tion of the signal energy, in other words it provides information about the signal energy and where it is focused in time domain.

11. **Harmonic Spectral Centroid:** This descriptor is calculated by taking the amplitude weighted means of the harmonic peaks in a power spectrum. This descriptor is similar to other centroid descriptors, but most commonly used for musical tones.

12. **Harmonic Spectral Deviation:** This descriptor is defined as the spectral deviation from a spectral envelope.

13. **Harmonic Spectral Spread:** This descriptor is calculated by taking the power weighted RMS deviation from the Harmonic Spectral Centroid.

14. **Harmonic Spectral Variation:** This descriptor defines the spectral variation between adjacent frames. It provides the normalized correlation between the amplitudes of two subsequent frames.

15. **Spectral Centroid:** This descriptor is calculated by taking the power weighted average of the frequency in a linear power spectrum. This descriptor is similar to Audio spectrum centroid, but most commonly used for musical instrument signals.

16. **Audio Spectrum Basis:** This descriptor is calculated by taking a series of basis function derived from the Singular Value Decomposition (SVD) of a normalized power spectrum.

17. **Audio Spectrum Projection:** This descriptor represents the projections of the basis function calculated with the Audio Spectrum Basis descriptor.

18. **Silence Segment:** This descriptor indicates if the audio segment has significant sound or not. It can also include an indicator for different level of silence based on a threshold.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The low level descriptors defined in the Audio Description Framework are very useful for building high level audio description tools. For the purpose of this work 3 low level basic spectral descriptors were used, as we were only interested in getting the general sound characteristics from the video shots. The descriptors used are as follows:

- Audio Spectrum Envelope
- Audio Spectrum Centroid
- Audio Spectrum Flatness

2.3 Applications of MPEG-7

Due to the enormous growth in the production of digital content, there is now a critical need for sophisticated applications to manage the content. These applications must be able to provide extremely efficient methods of managing, searching and retrieving the digital content. MPEG-7 has taken the step in providing a standard format for describing the content in order for different types of applications to use the descriptions to produce fast and effective retrieval of multimedia data.

MPEG-7 provides a vast array of descriptors so that digital archives, databases and libraries can be queried using not only text but also queries made of spoken words, images, melodies and video. Following are a few examples of the applications that have been developed using MPEG-7.

1. **Article Based News Browser:** The application is developed to group related news articles. As detailed in [10] the application provides the users with four key frames which summarize the events. These key frames are anchor key frame, episode key frame, news icon and synthesized text.

2. **Music Browser:** This application provides users to search a music database based on sound, music similarities as well as keyword search capabilities. This application is detailed in [11]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3. **Query by Humming:** In this application the user hums a tune for retrieval from a database [12]. The application takes into consideration the general shape of melody from note to note, that is if the note is higher or lower than the previous note.

4. **Real-time video identification:** In this application the video clips are analyzed in real-time in order to identify broadcasted content. Details can be found in [13]

In this thesis we are developing an application utilizing MPEG-7 descriptors of motion and audio for sports domain. The application will provide user with an interface from which they can choose a clip of interest and retrieve similar clips that have been annotated with high level semantics.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 3

Proposed Indexing System

3.1 Introduction

IN the entertainment and sports industry of today, video is extensively used. In sports it is being utilized to train athletes, scout for new talent, prepare strategy for the opposing teams and also for self-analysis. In the entertainment industry, video on demand is becoming very common. People are not looking for programming designed by some studio producer, but they are asking for programming depending on their mood and surrounding environment. Broadcasters are providing video for entertainment not only over the television but also over the Internet. Video is more a part of the sports and entertainment industry than most people ever imagined. As the technology advances and bandwidth constraints become less of an issue, the demand for video will only increase.

The design of any video indexing and retrieval system relies on the specific users of the system. Different users put different demands on the system based on their needs. For example in the sports video domain, currently two types of video logging approaches exist. First is the production logging, where the producer annotates live feeds or recorded footage to be used shortly, as an example the sports highlights program. Second is the posterity logging, where librarians add detailed and standardized annotation to archived material [14]. This is used by statisticians and other people involved with the industry.

Effective video indexing requires a multi-modal approach, the modes being visual,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

auditory and textual. Efficient algorithms need to be devised so that either the most effective modality is selected or multiple modalities are used in collaborative fashion [15]. Not only do we need to fuse the modalities to effectively index a video based on context but we also need to include domain specific information to provide higher-level semantic annotation. The domain knowledge can be used in both the indexing and querying aspects of the system [16].

3.2 Review of Sports Indexing Systems

Popularity of sports and general interest of people in sports means that in every part of the world sports video is being recorded and annotated for future use. Indexing and annotation are mostly done manually. Recently a lot of research has been conducted on automating the process of indexing and annotating the video streams. Nearly all the major sports have been used to test the indexing and retrieval systems. But the vast majority of the systems designed rely on visual information to index the sports video. Only a few systems use both the audio and visual information. And some utilize visual, textual and audio information to segment particular events within a sports video. In this section we review some of the sports indexing systems.

One of the major projects working in generating semantic sports video annotations is the ASSAVID project. As detailed in [17], this project focuses on developing a system that can categorize different types of sports and provide users with an interface to query events in a particular sport.

A summary of recent sports indexing systems is given below. These systems utilize low level features to index semantics in a particular sport. As evident from the research activity, an indexing system has been developed for nearly every popular sport in the western world.

Basketball

In the paper by Zhou et al. [18], basketball game is classified using a rule-based approach. The rules were calculated using an inductive decision tree learning approach

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

applied to low level image features such as motion, color and edge. The system was used to classify the basketball video into 9 major events shown in Table 3.1;

Features	Events
dominant direction	left/right Offense
motion magnitude	left/right Fast break
Color	left/right Dunk
Edge	left/right Score
	Close up

Table 3.1: Summary table of features used and semantic events retrieved in Basketball

Tennis

In the paper by Miyamori and Iisaku [19], particular tennis events are classified using visual models of the court and the players. First the court and net lines are extracted using a court model and Hough transforms. Then player position is extracted and tracked. Then ball position is tracked using special prediction modes. Lastly player behavior is identified using player shape changes. The system was used to classify the 5 events shown in Table 3.2;

Features	Events
court edges and net	forehand stroke
player position	backhand stroke
ball position	forehand volley
player behavior	backhand volley
	service

Table 3.2: Summary table of features used and semantic events retrieved in Tennis

In the paper by Lu and Tan [20], color features are first used to segment the video and then camera motion is used to identify the volleyball or tennis serve events by different teams/players.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Formula 1 Racing

In the paper by Petkovic et al.[21], some of the events that occur repeatedly during a race are classified by using audio and visual features. Bayesian Network and Dynamic Bayesian Networks are used to classify events like fly out, passing and starts. Table 3.3 summarizes the features and events;

Features	Events
Short term energy	starts
Pitch	passing
MFCC	fly outs
Pause rate	
color	
shape	
motion	

Table 3.3: Summary table of features used and semantic events retrieved in F1 racing

Track and Field

In the paper by Wu et al. [22], track and field events are classified using a three layer inference scheme. Initially low level features like global motion, color and texture are extracted and used by the system to segment the clips into semantic units. Then semantic concepts are extracted using learning RBF neural networks and decision tree classifier. Finally, a rule based finite state machine is designed for event inference. Table 3.4 summarizes the features and events;

Features	Events
Global Motion	High Jump
Color	Long Jump
Texture	Javelin
	Weight throwing
	Dash

Table 3.4: Summary table of features used and semantic events retrieved in track and field

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Soccer

In the paper by Assfalg et al. [23], two models are devised. The first model relies on motion features only while the second model also uses the location of players on the field. Hidden Markov model is used to classify the shots into 3 events. The features and events are summarized in Table 3.5.

Features	Events
Motion	Penalty
Player position	Free kick
	Corner

Table 3.5: Summary table of features used and semantic events retrieved in soccer

Baseball

In the paper by Han, Chang and Gong [24], baseball highlights are classified into views, which are then used in a Hidden Markhov Model to detect 4 types of play events. The features extracted are primarily based on camera motion, color of grass or field, edge detection for player height and texture analysis of the field and shape analysis of the field. Table 3.6 summarizes the features and the classified views and events;

Features	Events
Camera Motion	Home run
Player height	Catch
Shape	Hit
Texture	Infield play
Color	

Table 3.6: Summary table of features used and semantic events retrieved in baseball

In the paper by Han, Hua, Xu and Gong [25], an entropy based model is used to classify events in baseball. Closed caption text, audio features such as Mel cepstral coefficients and visual features such as color distribution, edge distribution, camera

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

motion and player tracking are used as features. Table 3.7 summarizes the features extracted and events classified by the authors.

Features	Events
Camera Motion	Home run
Player tracking	Outfield hit
Edge	Outfield out
Color	Infield hit
MFCC	Infield out
Closed Caption	Strike out
Texture	Walk

Table 3.7: Summary table of features used and semantic events retrieved in baseball

Football

In the paper by Miyauchi [26], audio, textual and visual information is used to classify American football video. Touchdowns and field goals are detected. The paper is an extension of previous work where only textual and visual information was utilized. In this paper the audio energy of a particular shot is also extracted and it is shown that the precision rate of the system is increased by adding this feature. Table 3.8 summarizes the features extracted and events classified by the authors.

Features	Events
Closed Caption text	Touch down
Short term audio signal energy	Extra Point
Dominant Color	Field Goal

Table 3.8: Summary table of features used and semantic events retrieved in football

In the paper by Lazarescu [6], American football games are classified into events using the natural language commentary from the game, the geometrical information about the play and the domain knowledge. Only five formations are classified with results shown for four. Table 3.9 summarizes the features and events classified.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Features	Events
Natural Language Commentary	Pro formation
Player tracking	I formation
	Single back formation
	Goal line formation
	Far formation

Table 3.9: Summary table of features used and semantic events retrieved in football

As evident from the above review of previous works, the area of sports video indexing is a very active research area. Nearly every sport has been investigated and a wide variety of features have been utilized to classify semantic events. The next section will provide details on the motivating factors for developing an indexing system in American football domain and also highlight the contribution made in this thesis.

3.3 Motivation and Contribution of the proposed system

The concept of "On Demand" entertainment and programming is fast becoming a reality with the popularity of digital TV channels. Now nearly every professional sports league and team in North America has a digital channel boasting of on demand programming and statistics. But the reality is that it takes nearly three to four hours in post production work to prepare the highlights for a game. For example, on NFL Sunday Ticket you get Highlights-On-Demand on Monday morning for the games played on Sunday. In order to minimize the delay between the live broadcast to "On Demand" programming we need technological advancements that can analyze the contents of the broadcast and derive the semantics from the input. These semantics can be made available to the users for querying in order to create a true "On Demand" experience.

The primary motivation is to develop an indexing system for American football that could be easily implemented in a hardware device, thus providing "On Demand"

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

indexing to the users. In order to accomplish this goal we had to utilize data that was readily available and also use features that did not require complex computations but were able to discriminate between the different plays effectively.

Keeping these objectives in mind, we decided on using motion vectors, which are already encoded in the MPEG bitstream. Thus only partial decoding of the bitstream is required to extract the motion vector information. From the discussion in Section 3.2 we can see that motion is a popular feature utilized in the sports indexing community. This is due to the fact that the sports domain consists of Recurrent Visual Semantic (RVS) events. For RVS events the color and texture information usually stays the same but motion varies. Thus motion carries a wide variety of information that can be utilized to discriminate between the different RVS events.

We also utilized audio information from the MPEG files. The audio track can be easily de-multiplexed from the MPEG bitstream without much computation, as most of the MPEG computational complexity is in the encoding of the image frames. The audio processing has long been established and is mature in the research community. The audio features provide complimentary support to the motion features, as using only audio information to discriminate between RVS events is very difficult if simple classification techniques are to be utilized.

Also seen from the section on review of sports indexing techniques, most of the indexing schemes utilize some sort of domain knowledge to fine tune the system for a particular sport. In the paper by [6], the authors propose a domain knowledge system based on sets and subsets concepts in the game. In this thesis we are also proposing an American football knowledge base, but it differs from the one proposed, such that fundamental set proposed in this work consists of RVS events and not concepts of the game.

In addition this thesis work focuses on utilizing existing standard MPEG-7 descriptors as the basic features in order to index events in American football. Some of the researchers such as Divakaran [5], have proposed applications for generation of summary highlights in sports domain, but no one has yet to our knowledge used the MPEG-7 descriptors to index RVS events in the American football domain.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3.4 Proposed System Overview

The proposed system consists of three stages. The first stage is responsible for localization of action within the video shots. The second stage extracts the MPEG-7 descriptors and the audio Mel Frequency Cepstrum features. These features are then passed to the third stage, the classification stage. The system is detailed in Figure 3.1.

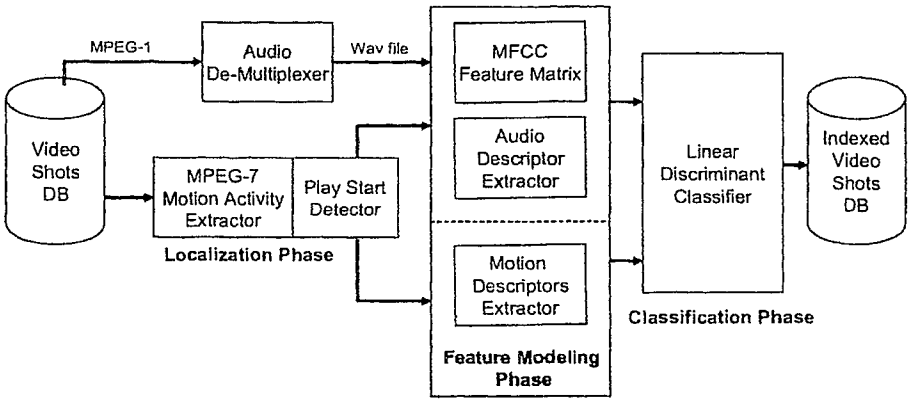

Figure 3.1: Proposed System overview

3.4.1 Stage 1: Localization Phase

The primary motivation of designing this stage was to reduce the analysis window size and secondly to remove the features that do not directly contribute to the action. The issue of localization of action in the sports domain is extremely important. In many sports like golf, American football, tennis, bowling and baseball, the players come to a certain position before starting the play. Then the play is followed by a delay before the next action takes place. This gap between the plays contain information that is not directly related to the semantics of the game.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Figure 3.2 shows the details of the first phase. This phase can be considered as the pre-processing stage of the indexing system.

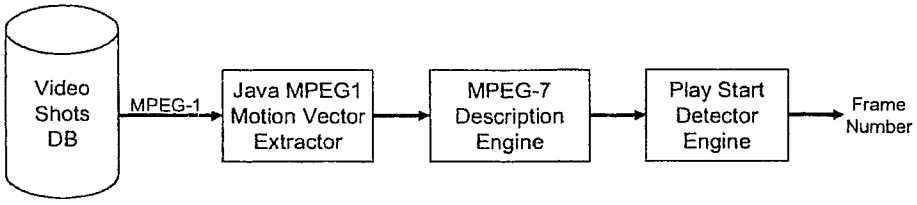

Figure 3.2: Play localization phase overview

First the motion vectors from the MPEG-1 video shots are passed through an MPEG-7 engine to extract the motion activity descriptors. This stage calculates the mean and standard deviation of the intensity of motion activity for each frame. These two components of MPEG-7 motion activity descriptor are then passed to the detection engine.

3.4.2 Stage 2: Feature Modeling Phase

The feature extraction and modeling phase is the heart of every indexing and retrieval systems. At this phase most of the important decisions are made regarding the features that can optimize the performance of the system. In this stage the features are also normalized in order to minimize the bias.

In the proposed system, three types of features are extracted from two groups of modalities. The first group is based on the audio content of the video shots. The second group is based on the visual content or more specifically, the motion content of the video shots.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Audio Features

Every sport has a language associated with it. This specific language is used by most of the commentators to describe the action. In American football there are many different events that take place, each play has its own specific words and its own rhythm, which are mostly spoken by the commentators to describe the play. In theory most of the similar types of plays will have similar sounding words spoken by the commentator. Therefore we want to extract the general sound characteristics from the audio information and use this to classify the video shots into different categories.

The first set of features were extracted by getting three (3) spectral MPEG-7 audio descriptors. The descriptors used were:

1. Audio spectrum envelope
2. Audio spectrum centroid
3. Audio spectrum flatness

These descriptors were chosen since each one defines a specific property of the audio signal. The first descriptor, audio spectrum envelope, represents the log-frequency nature of the audio signal. The second descriptor, audio spectrum centroid, represents the sharpness of the audio signal and the last descriptor, audio spectrum flatness, represents the tonal component in the audio signal. Therefore these three descriptors in combination provides details about the spectral characteristics of the audio signal.

Figure 3.3 shows the details of feature extraction for the first set of features. First the de-multiplexed audio file is passed to the MPEG-7 engine to extract the above mentioned audio descriptors, which are then normalized and quantized into 10 bins. This provides us with 30 features related to the spectral audio descriptors.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

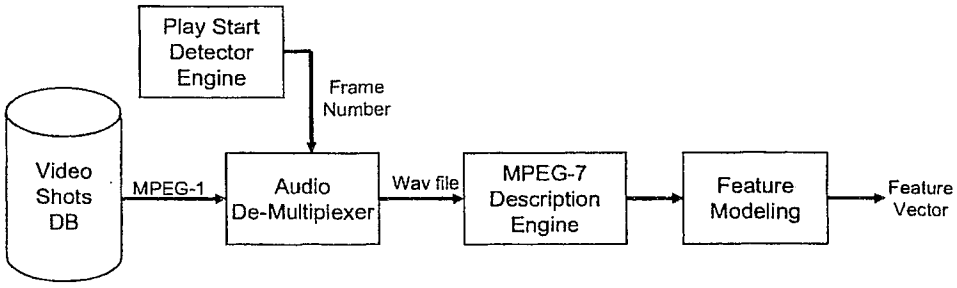

Figure 3.3: Audio feature modeling phase

The second set of features consist of Mel Frequency cepstrum Coefficients (MFCC). Due to the fact that most of the video shots contain a lot of crowd noise, and we want to extract the perceived rhythm and sound of the spoken content, we needed a feature that can model the human hearing and also works well under noisy conditions. MFCC has been used extensively in the speech recognition systems as it tries to emphasize the frequencies that are more perceptive to the human ear. Figure 3.4 shows the details of MFCC extraction.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

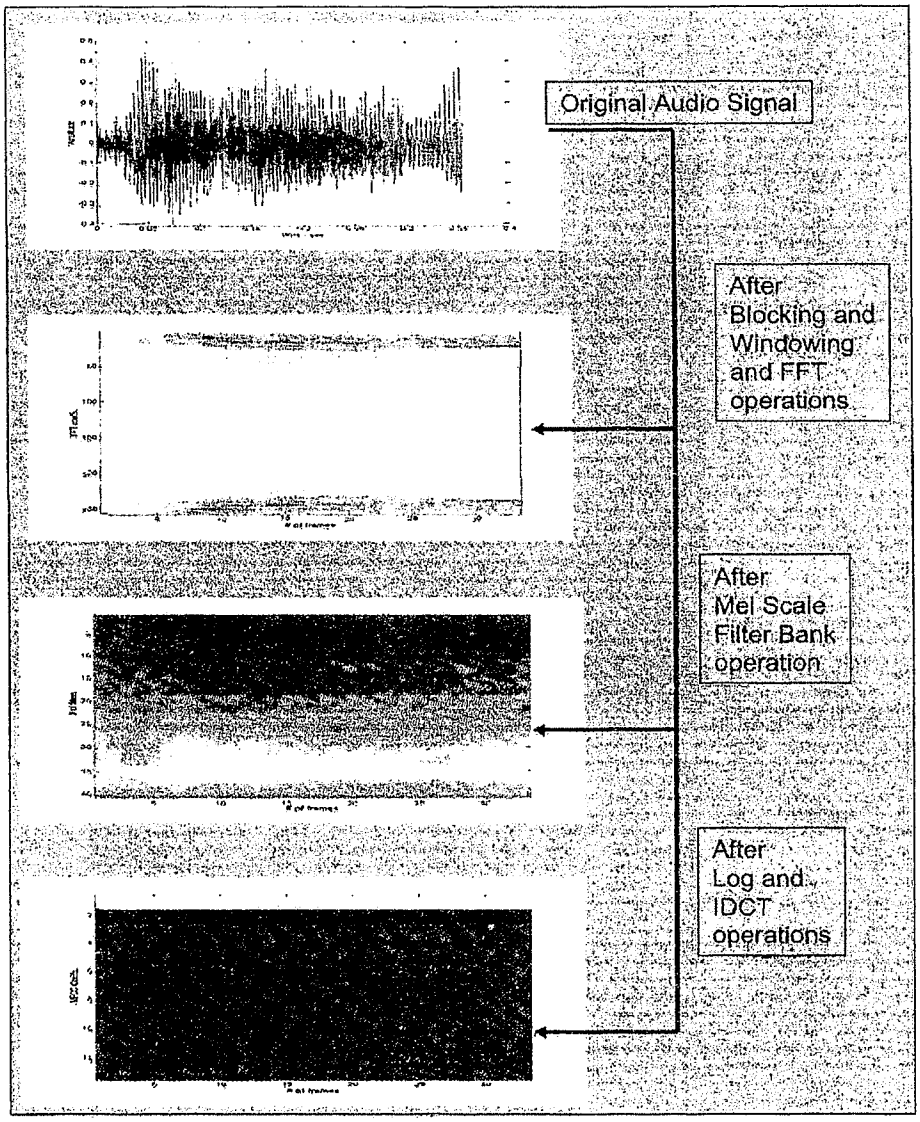

Figure 3.4: MFCC extraction process

First the audio file is pre-processed in order to remove the silent segments. Then 13 MFCC coefficients are extracted for each segment. Each of the segments have 50% overlap and thus there is lot of redundancy between adjacent MFCC values. This can be seen in the Figure 3.5. The blue colour represents low values and the red colours

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

represent high values. In order to reduce the dimension of the matrix, the MFCC values are passed to a feature reduction stage.

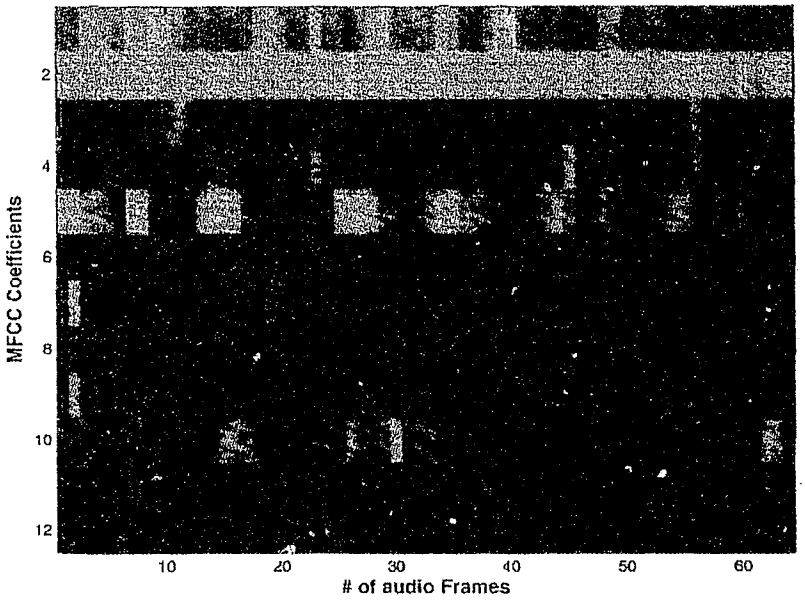

Figure 3.5: MFCC feature redundancy

The MFCC features are reduced to a 12×64 matrix. The first MFCC feature in every audio frame represents the average energy of the audio frame, therefore this feature is discarded. The other 12 coefficients are retained for each segment. 64 is the number of segments remaining after redundant feature reduction.

Motion Features

Motion plays an integral part in many sports indexing and retrieval systems. In this stage the MPEG-7 motion descriptor of motion activity is extracted with an optional descriptor of dominant direction of motion. Figure 3.6 shows the details of the feature extraction of motion.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

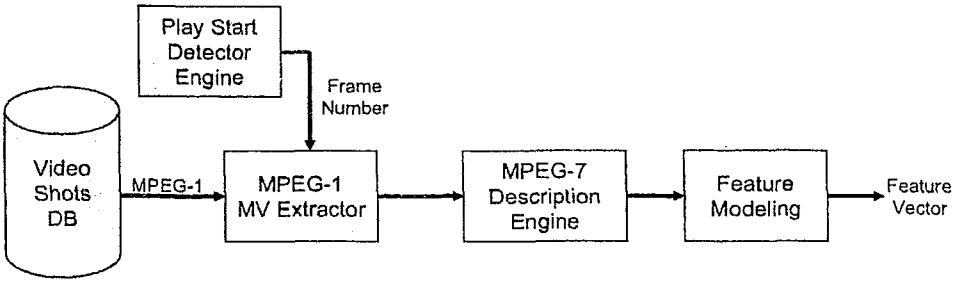

Figure 3.6: Motion feature modeling phase

First the MPEG-1 motion vectors are passed to an MPEG-7 engine which extracts the statistical properties of mean and standard deviation. It also calculates the dominant direction of motion. The intensity of motion activity descriptor and the dominant direction descriptor are quantized into a two dimensional matrix as shown in Figure 3.7.

The intensity of motion descriptor is quantized into 12 bins while the dominant direction descriptor is quantized into 8 bins. This gave us a 12×8 matrix feature which simultaneously represents the motion intensity and direction of motion in a video shot.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

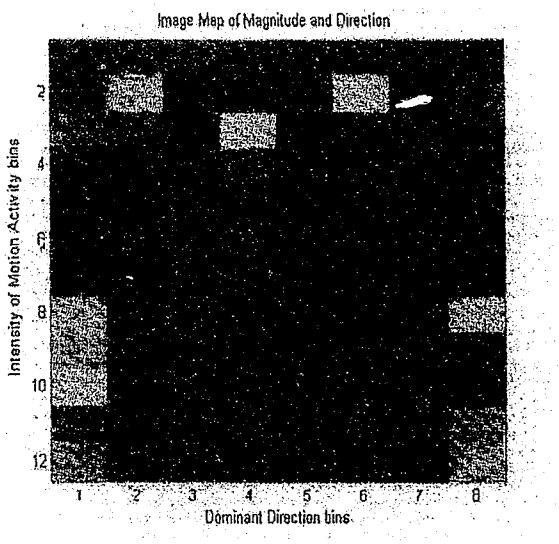

Figure 3.7: Motion feature quantization

Another set of motion features is extracted by calculating the mean and standard deviation of the magnitude of motion vectors with in a specified window. First the highest peak of the magnitude is detected and then the adjacent motion vector magnitudes are included in the window. This provides us with 2 more features.

3.4.3 Stage 3: Classification Phase

The primary objective in designing this phase was to utilize classification schemes that were simple and efficient. A decision had to be made on what type of classification scheme can achieve the goals of this phase. First the decision was made to utilize a supervised classification scheme, as unsupervised classification scheme use iterative algorithms which can be computationally expensive in a large data set. Secondly a decision was made to utilize linear classifiers rather than non-linear classifiers. The reason for this was to evaluate the performance of the system by using a simple classification scheme first and then evaluate the need to utilize more complex classifiers such as Neural Networks or Radial Basis Functions.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Linear Discriminant Analysis (LDA) generally refers to techniques that output a discriminant function that take linear inputs. In a specific sense LDA also commonly refers to techniques in which a transformation is done in order to maximize between-class separability and minimize with-in class variability.

LDA works on the feature set with no prior assumptions about the nature of the data set. It tries to compute a weight vector $\boldsymbol{w}$, which when multiplied by the input feature vector $\boldsymbol{x}$ would generate discriminant functions $g_i(\boldsymbol{x})$. For C classes problem we define C discriminant functions $g_1(\boldsymbol{x})...g_C(\boldsymbol{x})$. The feature vector $\boldsymbol{x}$ is assigned to a class whose discriminant function is the largest value of $\boldsymbol{x}$, as given by the following equation.

$$g_i(\boldsymbol{x}) = max_j g_j(\boldsymbol{x}) \ . \tag{3.1}$$

LDA has been a proven classification scheme. Therefore in order to perform the classification in the proposed system we utilized software package SPSS. Figure 3.8 shows the details of the classification phase.

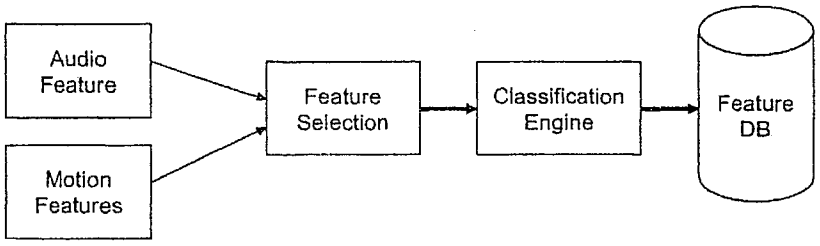

Figure 3.8: Classification phase overview

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The first part of the classification phase performs feature selection using a scatter matrix of within class spread and between-class spread. The criteria maximized is given by Equation 3.2.

$$Tr\{\mathbf{S}_W^{-1}\mathbf{S}_B\}\,, \tag{3.2}$$

where $\mathbf{S}_W$ is the within class spread and $\mathbf{S}_B$ is between-class spread and Tr is the trace function.

After feature selection, the features are classified using Fisher's criterion for LDA. This generates discriminant functions for each class and the selected features are classified in a class that has the highest discrimination value.

3.5 Test Database of American Football Video Shots

The test database was created by recording some of the National Football League games broadcasted during the 2003-2004 season. The games were recorded from all the four major networks, namely: ABC, CBS, ESPN and FOX. The recorded video was manually cut into shots containing all the details of a game. The shots were indexed into three categories namely: pass plays, run plays and kicking plays.

The database consists of 200 video shots with durations varying from 5 seconds to about 25 seconds. In the database there are 88 pass plays, 67 run plays and 45 kicking plays. A total of 8 different teams were used to create the database from 4 different networks. This variety in the database ensured that the sample space of our work was diverse and included different types of production styles.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 4

Semantic Localization

4.1 Introduction

THE concept of segmentation or localization of objects within a multimedia document has been an area of active research for quite some time. For example speech recognition relies on good localization of phonemes in order to perform similarity matching. Object segmentation in images has been implemented in MPEG-4, which has the option of coding the entire video frame or arbitrary objects within a frame.

Segmentation of objects from multimedia documents can be done in both the temporal domain or the spatial domain. Speech recognition is an example of temporal segmentation while segmentation of objects from images is an example of spatial segmentation. Likewise object tracking within a video is a combination of temporal and spatial segmentation.

The philosophy behind segmentation is to extract meaningful information from the multimedia object. The segmented object must contain semantic information so that it can be easily mapped into human perception. For example, we localize words from spoken content, since humans can relate to words and not to the signal energy. We segment objects like sun, tree or house from the images and not color or texture.

Building on the above mentioned philosophy, our objective was to localize play events from American football video shots. This localization will provide us with a semantic unit which can then be classified into a specific category. In this chapter, we

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

will review some of the related works within a sports video context. Then, we propose a novel algorithm to segment plays, followed by the experimental results detailing the performance of this algorithm.

4.2 Related Works

Sport events have very well defined structures. They have a set of rules that must be followed in order for the game to be played properly. This definite structure provides a variety of clues which can be used to segment the sports video. Therefore most of the segmentation and localization algorithms within the sports domain rely heavily on specific sport knowledge base.

In the paper by Nitta et al. [29], they proposed a scheme to localize semantic events by using closed caption (CC) text. The objective was to create semantic story events by evaluating CC text for keywords. The authors first segment the (CC) text based on pauses between dialogue or speaker changes. These segments are then input into a Bayesian network to evaluate the probability of the segment containing semantic information about American football.

In the paper [30], Li and Sezan proposed a sports highlight generation scheme. The authors first segment video in two categories, namely: play and non-play events. The play events are detected by using low level features such as color, texture, motion and player shape and movement. Based on these features two models are developed to categorize the video segments. The first is a rule based inference model and the second is the probabilistic inference model based on HMM. The authors experimented this scheme with baseball, football and sumo wrestling.

4.3 Proposed Algorithm

Many sports such as golf, baseball, bowling and American football have a requirement that the team or players must be in a distinctive position before each play. In golf the player positions himself by the ball in order to hit it in a certain direction. In baseball the batter awaits for the pitcher to go through its motions and deliver the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

pitch. Likewise in American football the two teams first line up face to face before the ball is snapped to begin the play.

The common theme among all these sports is the perceived motion activity, before and after the play starts. This distinction in the motion activity is utilized in the proposed algorithm in order to divide the video into non-play event segments and play event segments, as seen in Figure 4.1.

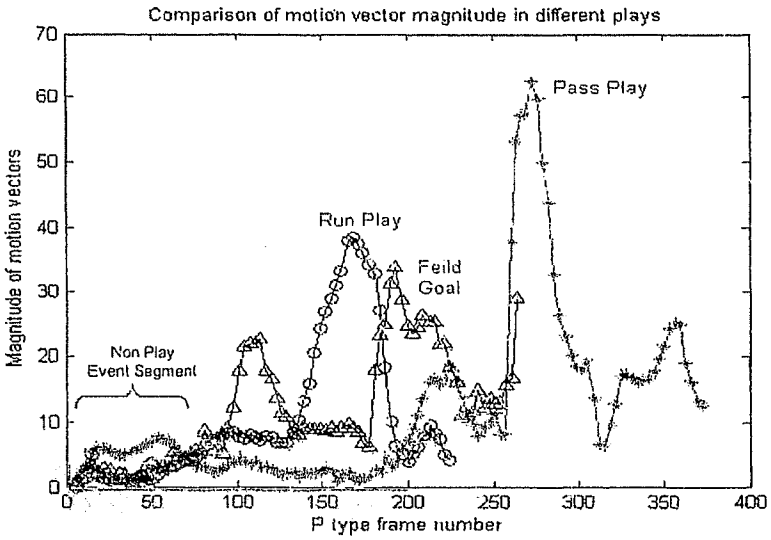

Figure 4.1: Mean of Motion Vectors for different types of plays

The primary objective of the algorithm is to detect the key frame that can be used as the starting point of the play event in the shot. The end point of the play event is not extracted, as in most American football video shots containing play events, the shot usually terminates at the end of the play. Therefore, the algorithm is designed to detect the instance where motion activity in the video shot is sustained at a certain level.

In order to extract the intensity of motion descriptor, MPEG-1 video motion vectors are used. Only the motion vectors from the P frames are analyzed in order to

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

speed up the processing time. In MPEG-7 the motion activity descriptor represents the standard deviation of motion vector magnitudes within a frame. This is given by the following equation.

$$\sigma_{mv} = \sqrt{\frac{\sum_1^N (MAG_{MV} - \mu_{mv})^2}{N}}, \tag{4.1}$$

where MAG_{MV} is the magnitude of motion vector with coordinates (x, y), and is calculated by $MAG_{MV} = \sqrt{x^2 + y^2}$. μ_{mv} is the mean of the motion vectors and is defined as:

$$\mu_{mv} = \frac{\sum_1^N MAG_{MV}}{N}, \tag{4.2}$$

where N is the number of macro-blocks that have a motion vector coded in the MPEG-1 stream. The number N varies from frame to frame as not all the macro-blocks are coded with a motion vector. The two features (μ_{mv} *and* σ_{mv}) are used collaboratively in the algorithm to detect the start point of the play. Figure 4.2 shows the plot of the mean and standard deviation of the magnitudes of motion vectors in a video shot.

An analysis with 20 video shots selected from each category was conducted to estimate the thresholds for the mean and standard deviation of motion vectors. From the analysis it was found that at the starting point of the play the mean was consistently within a range of 3 and 4.5, while the standard deviation of the motion vectors in the frame ranged from 1.2 till 4.2. This large variation between the standard deviation was due to the fact that in some video shots the play started as the camera was zooming in. Therefore based on the results from the above analysis, the threshold for mean was set at 4. In the MPEG-7 standard the motion activity descriptor goes from level 1 to level 2 when the standard deviation of the motion vector magnitude reaches 3.9. Therefore in our algorithm the standard deviation threshold corresponds to this change in level.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

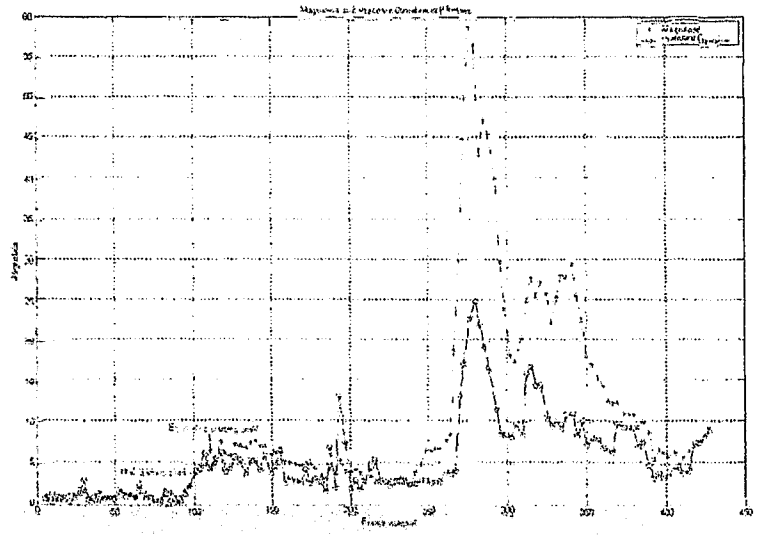

Figure 4.2: Mean and Standard deviation of Motion Vectors

Figure 4.3 shows the flow chart of the proposed algorithm to estimate the frame which represents the starting point of the play event. The steps of the algorithm are explained below:

- **Step1:** Find a P frame with a mean value of 4 or higher
- **Step2:** Determine the gradient of the mean values within a window (3 or 4 adjacent frames)
- **Step3:** If gradients are all positive mark the frame as possible starting point, else go back to Step 1.
- **Step4:** If the intensity of motion descriptor has a value of 2 or higher, return frame number as the starting point
- **Step5:** If the intensity of motion descriptor has a value of 1, determine the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

gradient of the standard deviation values within a window (3 or 4 adjacent frames)

- **Step6**: If the gradients are all positive return the frame number as the starting point, else go back to step 1.

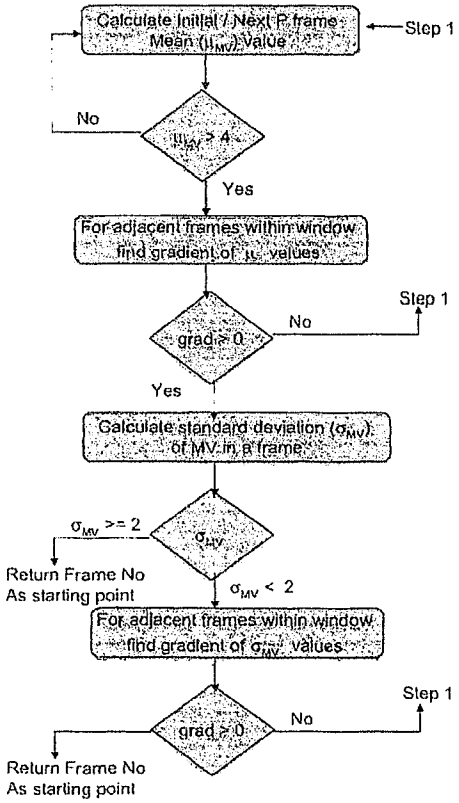

Figure 4.3: Flow chart of proposed algorithm

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

4.4 Play start detection results

The above algorithm was tested on the American football video shot database which consists of 200 video shots taken from 4 different games and 4 different networks, as detailed in Section 3.5. In order to measure the performance of the algorithm, we had to establish the ground truth about the starting point of the play event within each video shot. To establish the ground truth an observer manually indexed the frame number within a video shot which best represented the start point of the play event.

Comparison of results was done by calculating the difference between the ground truth frame number and the frame number estimated by the algorithm. Figure 4.4, shows the deviation of the estimated frame numbers from the ground truth.

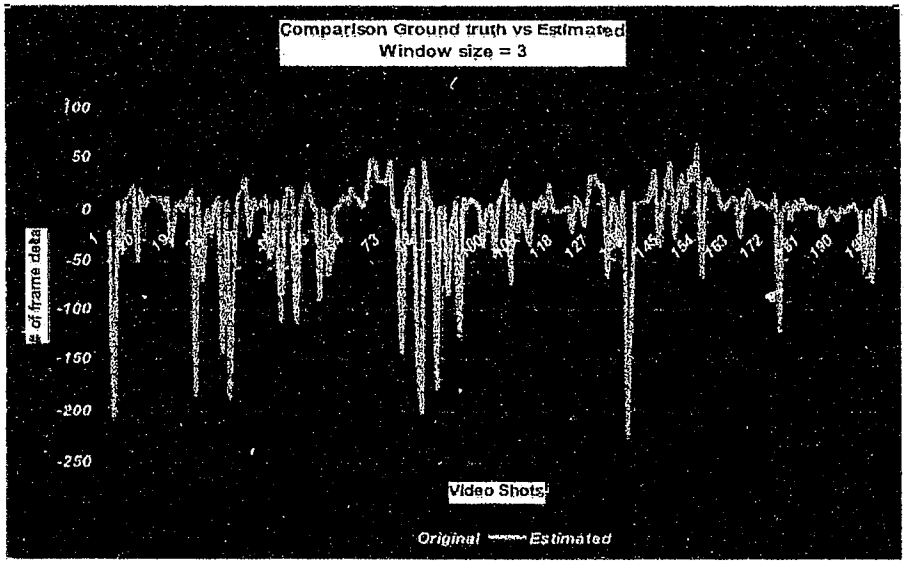

Figure 4.4: Deviation of estimated starting point from ground truth

It was noticed that we needed to develop some type of strategy to represent the results, since having a deviation chart only showed quantifiable results. The results

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

still needed to be evaluated in terms of what this deviation means in actual time. That is, we needed to evaluate if the algorithm is estimating a starting point too early or if it is estimating the starting point after a certain amount of delay.

Since MPEG-1 video has a frame rate of 30 frames/sec, building a histogram whose bin size was 30 frames would give a general idea of how apart the estimated frame numbers were from the ground truth in time domain. Figure 4.5, shows the performance of the algorithm using a window size of 3. The figure details what percentage of estimated frame numbers relate to early and delayed detection of the play event. From the figure we can see that the algorithm is able to detect 83% of the starting point within 1 second of the ground truth starting point.

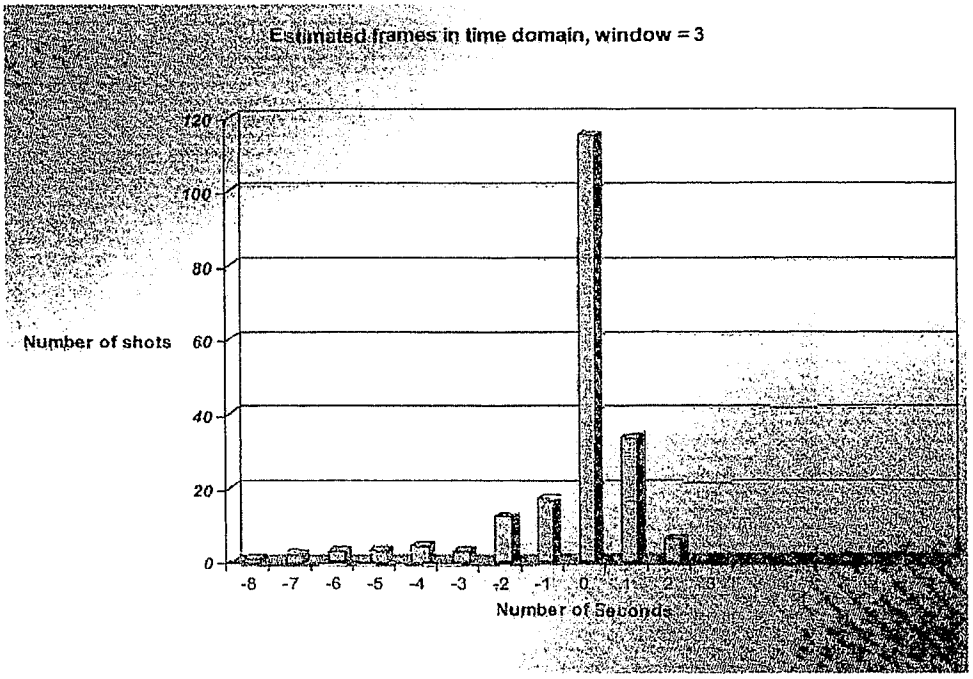

Figure 4.5: Performance of proposed algorithm in time domain

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

An experiment was also done by varying the window size to 4. Figure 4.6 shows the deviation between the estimated frame number and the ground truth. With the window size of 4, we are looking for the motion activity to be sustained for a longer period as compared to with window size of 3. Therefore, the algorithm will detect the starting point after a little delay. Thus in Figure 4.7 we see a lot more video shots in the range of +1 second and +8 second.

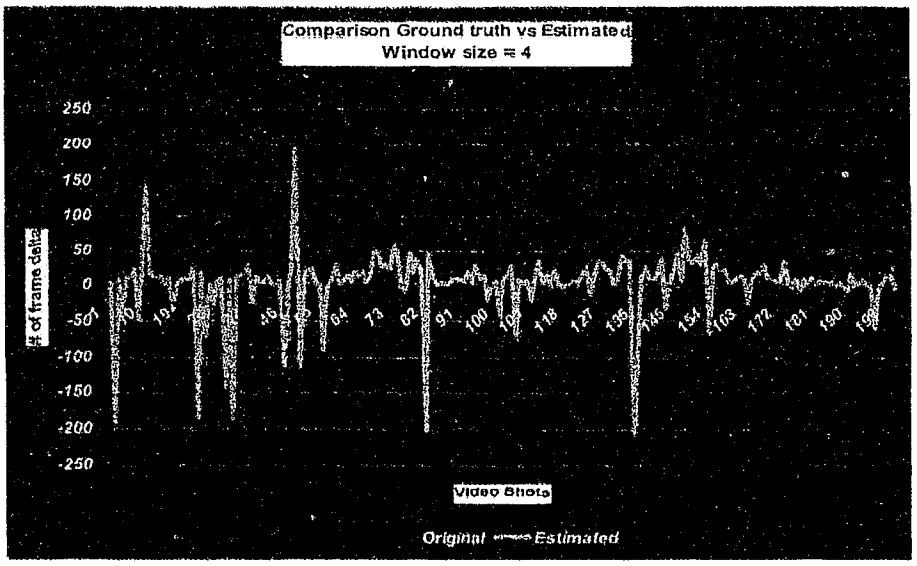

Figure 4.6: Deviation of estimated starting point from ground truth

4.5 Conclusions

In this chapter an algorithm was proposed to localize the play event within a video shot. MPEG-7 intensity of motion descriptor along with the mean of motion vector magnitudes was used to detect the starting point of the play. Keeping with our objective of fast and simple, we devised the algorithm that worked by analyzing

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

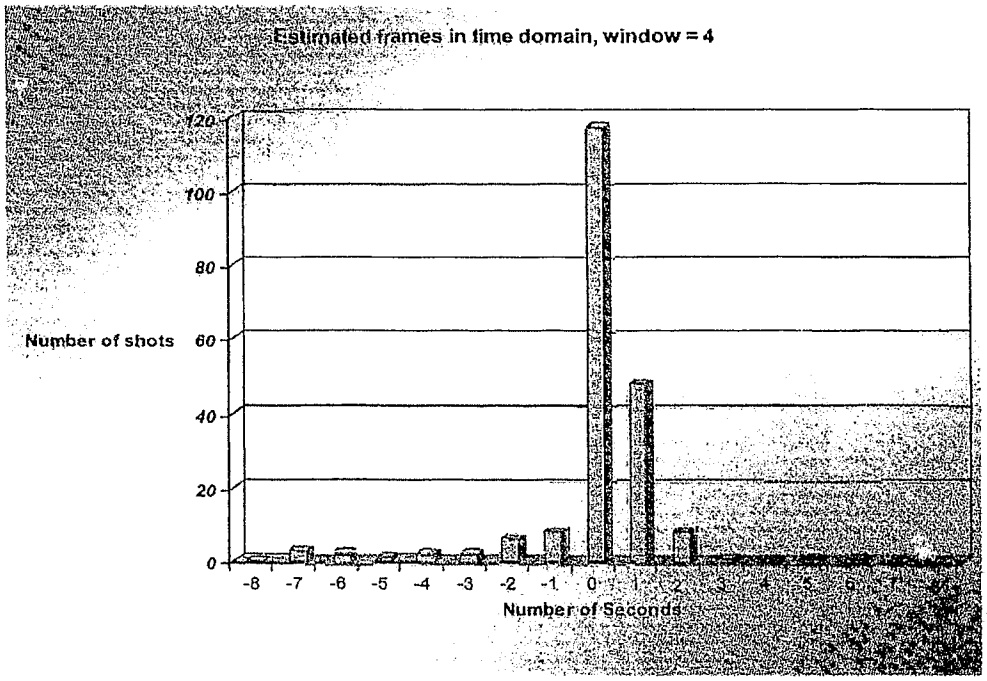

Figure 4.7: Performance of proposed algorithm in time domain

the mean and standard deviation of the motion vectors within a video shot. The algorithm relied on domain knowledge, that is in American football before the play starts the two teams face each other and stay in a particular position for a period of time. Thus providing low intensity motion just before the start of the play.

The algorithm detected the starting points of the play with 83% accuracy, that is 166 of the 200 video shots in the database had the starting points detected within ± 1 seconds of the original starting point. The accuracy of the algorithm can be increased to 86.5% by increasing the window size from 3 frames to 4 frames. But this change in window size has its own side effect. By increasing the window size we are looking for motion activity being sustained for a longer period of time. The trade off is that we get more shots that are detected after the play has started. In some of the cases the play starts without much motion activity happening in the surroundings.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In these cases, the starting point estimated by the algorithm is delayed a few seconds compared to the actual play start.

Table 4.1 shows a comparison of the performance of the algorithm utilizing the different window sizes.

Time index	WinSize=3	WinSize=4	Difference(3 vs 4)
-8	1	0	-1
-7	2	3	+1
-6	3	2	-1
-5	3	1	-2
-4	4	2	-2
-3	3	2	-1
-2	12	6	-6
-1	17	8	-9
0	115	117	+2
1	34	48	+14
2	6	8	+2
3	0	1	+1
4	0	0	0
5	0	1	+1
6	0	1	+1

Table 4.1: Comparison table of play detection performance using window size 3 vs 4

As seen from Table 4.1 columns 2 and 3, with a window size of 3 frames we have only 6 shots that are detected after the first second and all of them were detected within 2 seconds of the play start. But with a window size of 4, there are 11 shots that are detected after the first second, and one shot is detected 6 seconds after the play start. In the design of the system we used window size of 3, as most of the shots are detected within a couple of seconds of the play starting point. A 6 second delay is too much, as some critical feature information will be lost during this duration.

The robustness of the algorithm can be estimated by the fact that four different broadcasting station were used to construct the database, thus incorporating diverse styles of production and camera movements. Therefore it can be concluded that this algorithm can also be utilized for other sports in which the players take a specific position before starting a play.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 5

Indexing of American football

5.1 Introduction

One of the biggest application areas for MPEG-7 is multimedia indexing and retrieval. The feature formulation and classification strategy forms the essence of all indexing and retrieval applications. The standard does not specify how the features are extracted or how the the standard descriptors are used in a classification scheme. MPEG-7 creates a standard description made up of low level features, such that applications can be developed without regard to how the features were extracted. The applications take the standard descriptors and combine them with other descriptors in order to create a feature space which can be used to index and query the related database.

As evident from the review of works done in Section 3.2, indexing of sports video into semantic events requires a complicated strategy. It requires multi-modal features as well as domain knowledge in order to build an effective indexing system. Since the introduction of the MPEG-7 standard, there has been significant research effort put in developing applications based on MPEG-7 descriptors.

In Section 2.3, we presented a variety of commercial that utilize the standard descriptors. But to this date there has been only a few applications that utilize MPEG-7 descriptors for sports video indexing and retrieval. The application we are proposing is a first in the American football domain, which utilizes MPEG-7 motion and audio descriptors along with MFCC features.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Section 5.2 details the descriptors and features utilized to create the feature space of the proposed application system. Section 5.3 details the classification results obtained by utilizing a LDA technique. Section 5.4 presents the conclusions and observations on the proposed system.

5.2 Feature Extraction

From the review of sports indexing applications detailed in Section 3.2, we can see that audio and motion play an important role in providing discriminant features for sports domain video indexing and retrieval. In the case of American football, visual or motion features play a significantly dominant role in discriminating between different types of plays as shown in Figure 5.1. Therefore first we evaluate the efficacy of using motion descriptors for an American football video indexing system and then we evaluate the changes in system performance by adding audio descriptors and MFCC features.

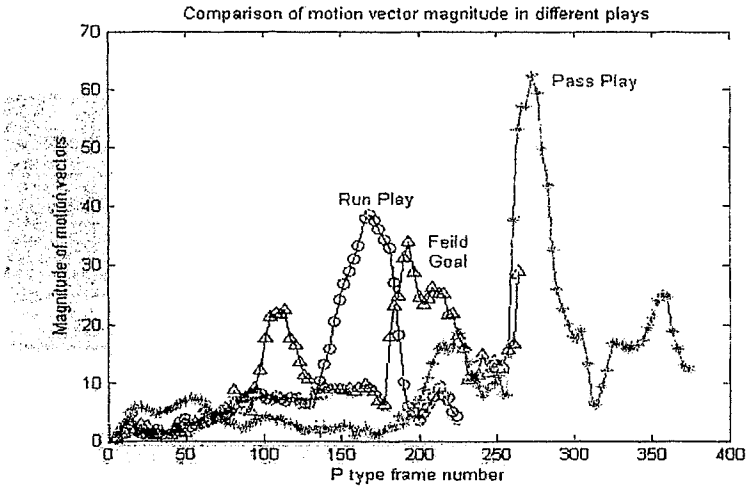

Figure 5.1: Different type of Motion Activity

PROPERTY OF
RYERSON UNIVERSITY LIBRARY

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5.2.1 MPEG-7 Motion Descriptors Feature Mapping

The motivation behind using the motion descriptors was due to the fact that in American football the global motion between different types of plays provides a variety of clues. In order to understand fully the difference in motion between the plays, first we require a detailed explanation of general motion involved in the plays:

- **Pass Plays:** During a pass play first the motion is lateral in order to track the movements of a quarterback who is going to throw the ball. Then it is followed by rapid zoom out and followed by a lateral movement to follow the throw. At the end of the play the motion is tracking the player to whom the ball was thrown. Therefore, the movements for a pass play involve first low intensity lateral movement followed by high intensity zoom out and lateral movement and then in the end low intensity lateral movement. Figure 5.2 shows some of the key frames from a pass play.

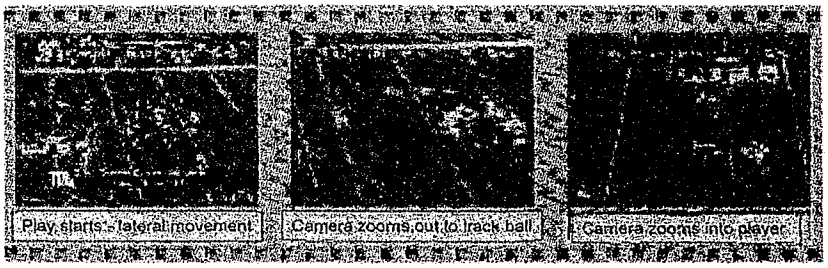

Figure 5.2: Key Frames of Pass Play

- **Run Plays:** During a run play first the motion is lateral as the runner gets the ball. Then the camera zooms in, to track the movements of the ball carrier. This zoom in provides the perception of high intensity motion. At the end

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the camera tracks laterally the movements of the ball carrier. Therefore, the movements of a run play involve first low intensity lateral movement, followed by short high intensity lateral movement and in the end low intensity lateral movement. Figure 5.3 shows some of the key frames from a run play.

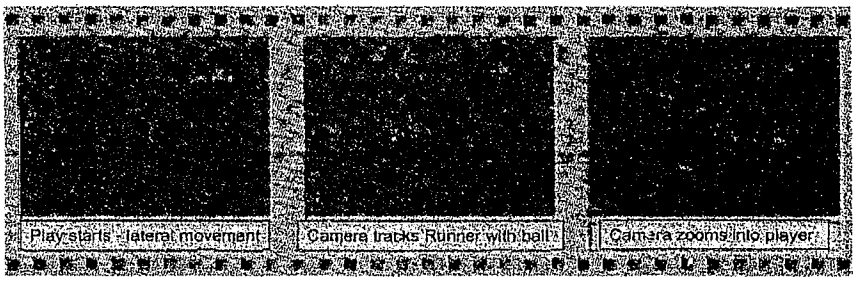

Figure 5.3: Key Frames of Run Play

- **Kicking Play:** In the kicking play category, there are two different types of kicks that take place. Each one has a completely different type of motion associated with it. The two types of kicking plays are detailed as follows:

  - **Kickoff/Punt (K/P):** In this category of kicking play, the kicker starts with kicking the ball high in the air. This motion causes the camera to rapidly zoom out to capture the kicked ball. After the kick the camera zooms into the player who has the ball and tracks the movements of the ball carrier. Therefore this play has movements that involve first high intensity motion of zooming out and zooming in with horizontal direction movement, followed by low intensity motion laterally. Figure 5.4 shows some of the key frames from a kickoff/punt play.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

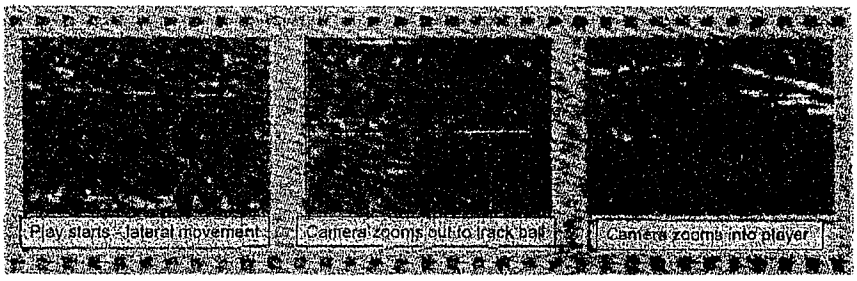

Figure 5.4: Key Frames of Kickoff/Punt Play

- **Field goal/Extra point (FG/XP):** In this category of kicking play, the ball is long snapped (short under hand throw) to a holder who sets the ball up to be kicked by a kicker. The majority of the movement is low intensity with most of it coming after the kick when the camera is tracking the kicked ball as it sails towards the goal post. Therefore the majority of motion in this category is vertical and low intensity. Figure 5.5 shows some of the key frames from a field goal/ extra point play.

The global motion of camera, the intensity of motion and the direction of motion provide valuable discriminating information regarding different types of plays. In this work we build the motion based feature set by utilizing intensity of motion descriptor and dominant direction descriptor of MPEG-7. The motivation behind using the two descriptor in combination comes from analyzing the explanation of different plays as formulated above.

The magnitude of motion vectors was calculated by extracting the encoded motion vector given by coordinates (x, y) from the macro blocks within P frames of the MPEG-1 video stream. The magnitude is given by the following equation:

$$MAG_{MV} = \sqrt{x^2 + y^2}\,. \tag{5.1}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

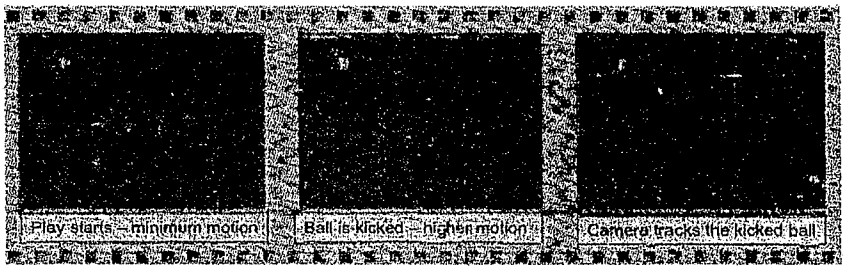

Figure 5.5: Key Frames of Field Goal/Extra Point Play

According to MPEG-7 description the standard deviation of the magnitudes of motion vectors formulates the intensity of motion descriptor. The descriptor takes on the value of 1 through 5, 1 meaning low intensity and 5 meaning high intensity. For the purpose of this work, the standard deviation of the magnitudes of motion vectors is quantized into 12 levels. Experiments showed that by using only 5 levels, the discrimination between plays was significantly lower. Thus to provide better resolution of motion activity, the magnitude of motion vectors were quantized into 12 levels.

Similarly the direction of each of the motion vectors encoded in the macro blocks of P frames of the MPEG-1 video stream was calculated. The direction of the motion vector is calculated by using the following equation:

$$\Theta_{MV} = \arctan(\frac{y}{x}) \,. \tag{5.2}$$

According to MPEG-7 description the dominant direction descriptor is calculated by quantizing the angles of the motion vectors into 8 levels as shown in Figure 2.2. For the purpose of this work, the same 8 quantization levels were used to define the dominant direction descriptor.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The fact that the intensity of motion descriptor and dominant direction descriptor cannot provide sufficient discriminating features if utilized independently, we decided to create a 2D feature map of intensity of motion descriptor and dominant direction descriptor. This 2D map consisted of 12 levels for intensity and 8 levels for direction, as defined above. Figure 5.6 shows the feature map developed based on the two motion activity descriptors for a video shot. In the feature map the blue colour corresponds to low values and the red colour corresponds to high values.

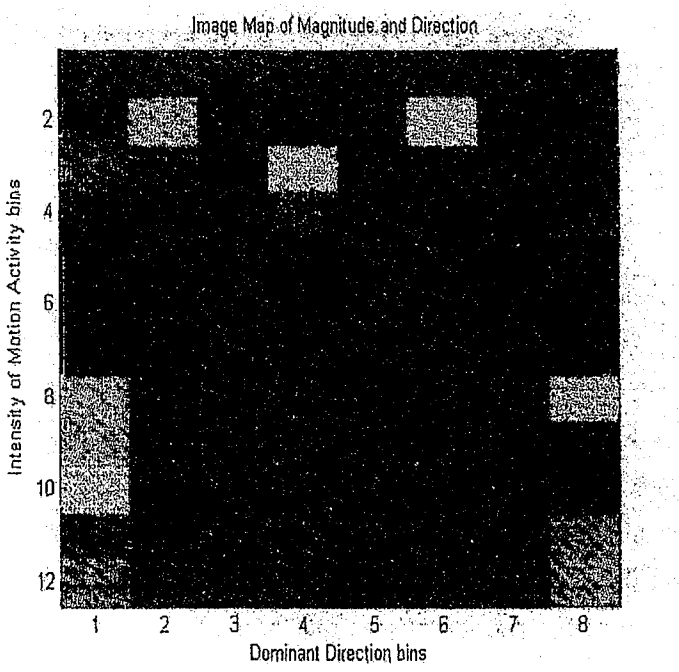

Figure 5.6: Motion feature map

The motivation behind this was to create a feature that was modeled by taking both the intensity and direction of motion into consideration; thus discriminating between high intensity motion in upward direction versus high intensity motion in the lateral direction.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The motion feature map provides a unique representation of only 96 dimension for both intensity of motion and direction of motion within a video shot. This compact representation can be used as input to a classifier to test the efficacy of motion descriptors in discriminating between American football plays.

5.2.2 MPEG-7 Audio Descriptors Feature Mapping

The motivation behind using audio descriptors is that most sports have a certain vocabulary associated with each event. Almost all the announcers will utilize some of the vocabulary to describe similar events. Therefore we wanted a compact representation of audio characteristics to describe the general tone and pitch of the announcer. The purpose was not to recognize all the spoken words, but only to analyze the similarity in the spoken sound between similar events.

As mentioned in Section 3.4.2.1, we used three MPEG-7 audio descriptors namely, Audio Spectrum Envelope, Audio Spectrum Centroid and Audio Spectrum Flatness. Figure 5.7 shows the input audio signal and the output of the three descriptors. Details on the extraction method of the audio descriptors is given below:

- **Audio Spectrum Envelope (ASE):** This descriptor represents the power spectrum of an audio signal. It is computed by calculating the Fourier transform of the audio signal which is windowed using a Hamming window with an overlap of 50% between adjacent audio frames or windows. The size of the Hamming window is taken to be 10ms. This descriptor is calculated using the following equations as given in [31]:

$$S(l,k) = \Sigma_{n=0}^{N-1} s(n + lM)w(n) \exp(-j(\frac{2\pi}{N})nk) \ , \tag{5.3}$$

  where N is the size of the short time fourier transform $S(l,k)$, k is the frequency bin index, l is the time audio frame index, $w(n)$ is the analysis window function of length lw and M is the hop size. The short time fourier transform $S(l,k)$ needs to be normalized by a factor of N in order to preserve Parseval's Theorem

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

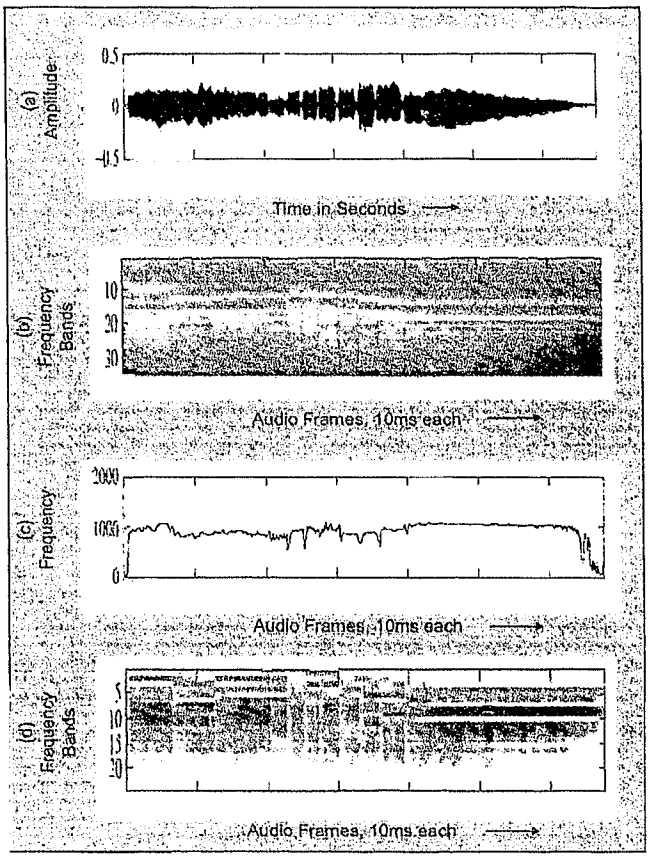

Figure 5.7: (a) Original audio signal; (b) Audio Spectrum Envelope descriptor output 1/4 octave resolution; (c) Audio Spectrum Centroid descriptor output; (d) Audio Spectrum Flatness descriptor output

and since ASE represents only the power spectrum, therefore we can estimate the ASE descriptor as follows:

$$\mathrm{ASE}(l,k) = \frac{1}{\alpha \cdot N}|S(l,k)|^2 , \tag{5.4}$$

where α is the window normalization factor. The number of frequency bins can be varied based on the octave resolution required. One bin is reserved for power

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

between 0 Hz and 62.5 Hz, while another one is reserved for power between 8 kHz and Nyquist rate. With 1/8 of octave resolution the frequencies in the middle are divided into 8 bins, thus providing a spectrum envelope consisting of 10 bins. Figure 5.7(b) shows Audio Spectrum Envelope description with 1/4 of octave resolution.

- **Audio Spectrum Centroid (ASC):** This descriptor represents the center of gravity of the power spectrum. That is, it shows the dominant frequencies in the power spectrum. This is calculated by adding the energy in each frequency bin by the total energy in the frame as given by the following equation:

$$\mathbf{ASC}(l) = \frac{\Sigma_{k=0}^{K-1} k \cdot \mathbf{ASE}(l,k)}{\Sigma_{k=0}^{K-1} \mathbf{ASE}(l,k)}, \tag{5.5}$$

where k is the frequency bins index. The **ASC** for each frame was then normalized between the values of 0 and 1, after which they were quantized into 10 bins in order to provide compact representation of the MPEG-7 descriptor. Figure 5.7(c) shows the **ASC** description. The figure shows that the audio signal has mainly low frequencies as the centroid is mainly below 1 kHz.

- **Audio Spectrum Flatness (ASF):** This descriptor represents the overall tonal component in the power spectrum of the audio signal. It is calculated by calculating the geometric mean of the audio frame and dividing it by the arithmetic mean of the audio frame as shown by the equation

$$\mathbf{ASF}(l) = \frac{(\Pi_{k=0}^{K-1} \mathbf{ASE}(l,k))^{\frac{1}{N}}}{\frac{1}{N}\Sigma_{k=0}^{K-1} \mathbf{ASE}(l,k)}, \tag{5.6}$$

where k is the frequency bins index and N is the size of the short time fourier transform window. Figure 5.7(d) shows the Audio Spectrum Flatness of an audio signal. The spectrum is then normalized and quantized into 10 bins.

5.2.3 MFCC Feature Mapping

Mel Frequency Cepstrum Coefficients (MFCC) have been widely used within the speech recognition community as a basic spectral feature set that provides robust classification of sound.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Due to the fact that most of the video shots contain a lot of crowd noise, and we want to extract the perceived rhythm and sound of the spoken content, we needed a feature that can model the human hearing and also works well under noisy conditions. Mel Frequency cepstrum has been used extensively in the speech recognition systems as it tries to emphasize the frequencies that are more easily perceived by the human ear.

The Mel scale first defined in [27] and revised in [28], was developed to model the pitch of a sound. Pitch is a non linear combination of both frequency and intensity. The Mel scale tried to put in perspective the relationship between pitch of the sound and its intensity. Therefore the pitch of 1000 Mels was half the intensity of the pitch at 2000 Mels. That is the Mel scale measures pitch in an absolute scale. The following equation shows the relationship between frequency and pitch.

$$\nu(f) = \frac{4491.7}{1 + \exp(7.1702 - 1.9824\log(f))} - 30.360\,, \tag{5.7}$$

where f denotes frequency in Hertz and v denotes pitch in mels. Figure 5.8 shows the Mel scale from which the above equation was derived by curve fitting.

The MFCC features are extracted by first de-multiplexing the audio stream from the MPEG-1 video. This audio stream is then input into the a MFCC feature extraction system, which first performs pre-processing on the raw audio data. After the pre-processing step the MFC Coefficients are extracted and finally the coefficient matrix is passed through a feature reduction or compaction step. This sub system is shown in Figure 5.9.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

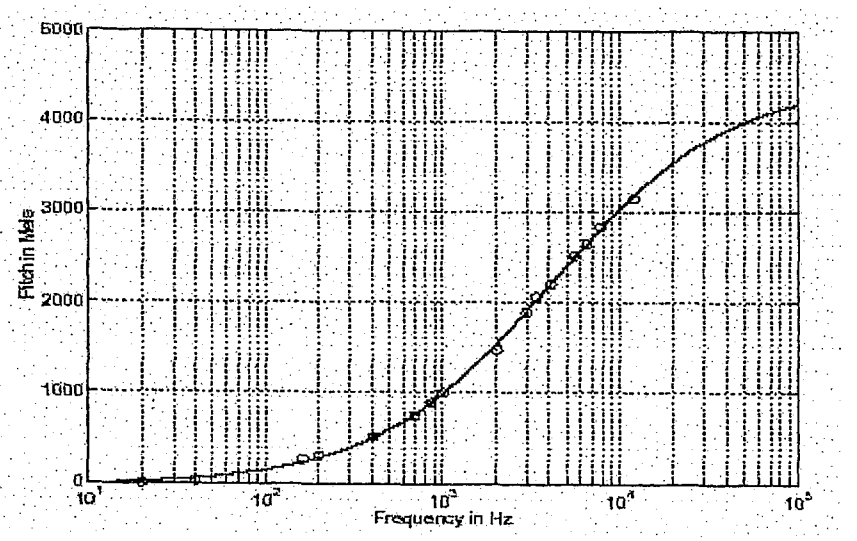

Figure 5.8: The MEL Scale

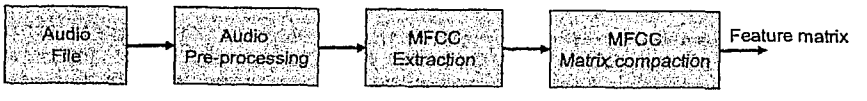

Figure 5.9: MFCC feature extraction sub system

MFCC Feature Mapping - Audio preprocessing

The objective of this stage is to remove any silent segments from the raw audio data. This is done in order to model the MFCC features for only the voiced segments of the audio signal. Following steps were taken to remove unvoiced segments from the audio data:

- Window audio signal in 25ms frames with 25% overlap between adjacent frames

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- Compute the mean of each window
- Remove frames where mean is less than a threshold

MFCC Feature Mapping - Coefficient calculation

In this step of MFCC feature modeling the Mel frequency coefficients are extracted using the algorithm proposed in [32]. Following are the step followed to extract the coefficients for a N sample audio signal given by $\mathbf{s} = s_0, \ldots, s_{N-1}$:

- **Pre-Emphasis:** This is a high pass filtering operation in order to compensate for the spectral tilt. In time domain this is performed by subtracting the original signal at a particular time instant from the signal in the previous time instant which is scaled by a constant. This constant is usually taken to be between 0.9 and 1. This step in time domain is given by the following equation:

$$s_i = s_i - \alpha s_{i-1} \quad for \quad 0.9 \leq \alpha \leq 1\,, \tag{5.8}$$

  where the constant α was taken to be 0.95.

- **Blocking and Windowing:** In this step the input signal is divided into frames of equal length. Each frame is made to overlap the previous audio frame. The overlap portion can vary between the ranges of 20% to 50%. The selection of the frame length is dependent on the specific use, but in most speech recognition applications the frame size is 10-40ms long. The individual frames are then windowed using a windowing function in order to reduce the spectral artifacts and also to smoothen the discontinuities in the signal edges. Usually a Hamming or Hanning type window function is used to perform this step. In this work we used an overlap window of 50% and Hamming window function given by the following equation:

$$y_{ij} = y_{ij} w_j \quad for\ j = 0, \ldots, W-1, \ and\ i = 0, \ldots, M-1\,; \tag{5.9}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Here y is the frame window, w is the Hamming window function as given by the equation below. M is the total number of frames in the audio signal and W is the size of the window:

$$w_j = 0.54 - 0.46\cos(\frac{2\pi j}{W-1}) \quad for\ j = 0, \ldots, W-1\ ; \tag{5.10}$$

- **Frequency domain transformation:** In this step each frame is transformed in the frequency domain using the Fourier Transform. The transformation into the frequency domain results in a signal with both real and complex parts. If only the power spectrum is to be utilized then the magnitude square of the Fourier coefficients are used. Therefore the output at this stage is the power spectrum coefficients based on the length of the transform for each frame. This is given by the following equation:

$$\mathbf{z}_i = |fft(y_i)|^2 \quad for\ i = 0, ..., M-1\ ; \tag{5.11}$$

- **Mel Filter Bank:** The Mel filter bank is designed to capture lower frequencies and emphasize the information in the speech signal at these frequencies. Most of the important and useful information in a speech signal is present in the lower end of the spectrum. The filter banks are constructed of triangular shaped filters and are made to overlap such that the lower frequency of the filter corresponds to the center frequency of the previous filter and the high frequency of the filter corresponds to the center frequency of the next filter. The filters below 1 kHz are spaced linearly and the filters above 1 kHz are spaced by increasing the distance 1.1 times after each filter. Usually the range of frequency covered by the filter bank lies between 20 Hz till half the sampling frequency of the signal. Figure 5.10 shows the arrangement of the Mel scale filter banks.

  When the power spectrum of the frequency coefficients is passed through the filter bank, the output is the inner product of the filter with the power spectrum coefficients. This provides the energy coefficients of each filter for every audio frame.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

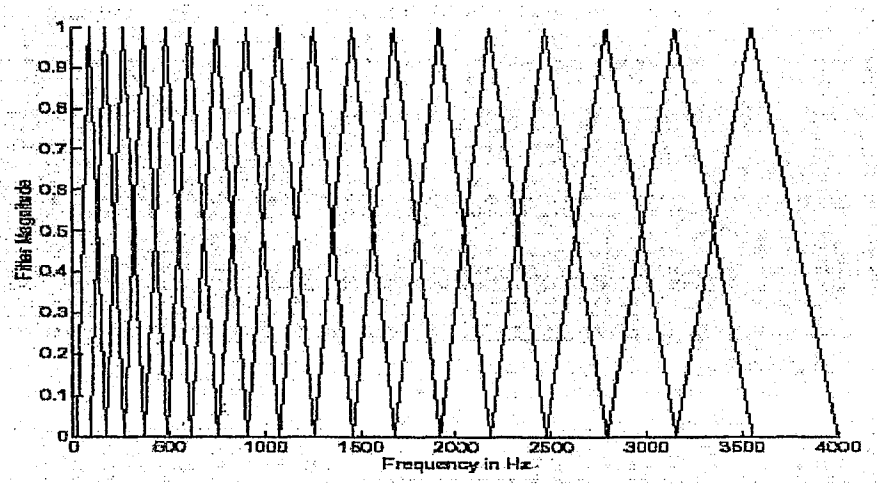

Figure 5.10: Mel filter bank

- **Log of energy coefficients:** By passing the signal through the Mel scale filter bank the coefficients are modeled according to human auditory system of pitch perception. In a very crude fashion the intensity-loudness relationship of the humane auditory system can be modeled by taking the log of the filter coefficients. This is done as shown in the equation below:

$$p_{ij} = \log(\frac{1}{A_j} \sum_{k=0}^{U-1} s_{ij} f_{ij}) \,, \tag{5.12}$$

where $i = 0, ..., M-1$; and $j = 0, ..., K-1$. U is the number of audio frames in the signal. A_j is the normalized energy of each frame and is given by the following equation:

$$A_j = \sum_{k=0}^{U-1} f_{ij} \,; \tag{5.13}$$

- **Inverse frequency domain transformation:** This step is performed to reduce the dimension of the coefficients and also to de-correlate the coefficients. Since only the power spectrum was used as input to the filter banks, the inverse frequency transformation only has real part. Because of the cosine transformation property most of the signal energy is compacted in the first few coefficients.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Thus only a first few coefficients are selected for the feature vector of each frame. Usually in speech recognition systems 9 to 13 coefficients are used. In this work we took the first 13 coefficients, the first coefficient was discarded as it contains the energy information of each frame.

MFCC Feature Mapping - Coefficient compaction

Since video shots for different plays varied in size, dimensions of the coefficient matrix were not uniform. Longer video shots had more audio frames compared to shorter video shots. Therefore to make the matrix dimension uniform we had to apply a matrix compaction strategy. Also there was redundancy between the coefficients of the adjacent frames, as seen in Figure 5.11. In the figure blue colour represents low values and the red colours represent high values.

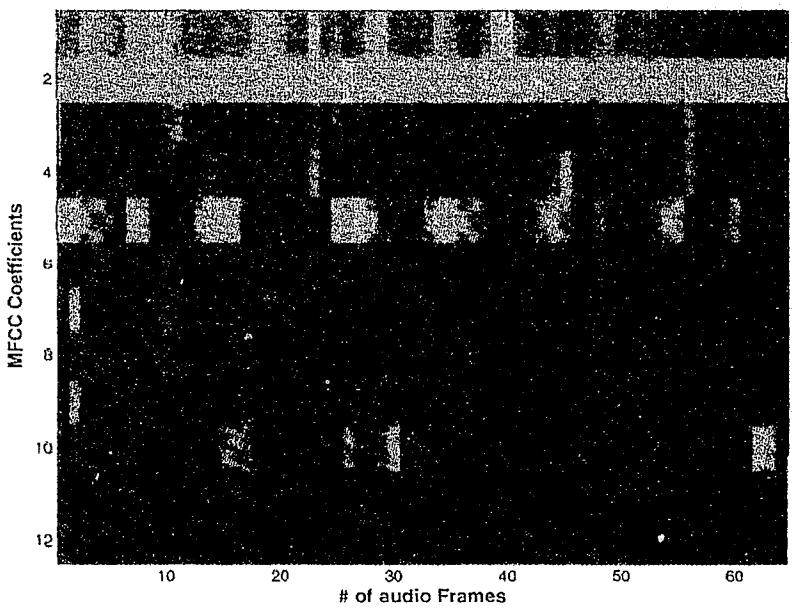

Figure 5.11: MFCC feature redundancy

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In order to make the matrix uniform and reduce redundancy, we decided to fix the number of audio frames. Euclidean distance between adjacent audio frames was calculated and a threshold was calculated empirically that would compact the coefficient matrix size into 64 segments. An experiment was conducted to find out the number of segments that would provide optimal results. In the experiment, segment sizes of 32,64,128 and 256 were used and the classification results showed that taking 64 segments provided best results. Thus this provided us with 12×64 feature matrix or a 768 dimension feature vector.

5.3 American Football RVS Event Classification

5.3.1 Linear Discriminant Analysis

As stated in Section 3.4.3, to perform classification of the feature set we utilized LDA, specifically Fisher's LDA [33]. In general for a multi-class problem, Fisher's LDA tries to choose those vectors $\mathbf{a}_i$ of the feature matrix A that will maximize the equation:

$$\frac{\mathbf{a}_i^T \mathbf{S}_B \mathbf{a}_i}{\mathbf{a}_i^T \mathbf{S}_W \mathbf{a}_i}, \tag{5.14}$$

subject to the orthogonality constraint $\mathbf{a}_i^T \mathbf{S}_W \mathbf{a}_i = \delta_{ij}$. In the equation $\mathbf{S}_W$ is the within class spread and $\mathbf{S}_B$ is between-class spread.

That is to say that the with-in class covariance matrix in the transformed space is an identity matrix. The first vector $\mathbf{a}_1$ is the Fisher's linear discriminant and the second vector $\mathbf{a}_2$ is orthogonal to $\mathbf{a}_1$ and so on.

The characteristics of Fisher's LDA can be stated as follows [33]:

- For C number of classes a transformation is done to a space of C-1 dimension.
- The transformation is computed with no prior assumptions about the distribution of the data set.
- Discrimination can be conducted by utilizing reduced dimension set of the feature set

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- For complex non linear classifiers, LDA can be used as a post processing step.

In order to calculate the transformation matrix which can be used to calculate the discriminant function of different classes, the following steps are used:

- First the mean of the features for each class is calculated and for the i^{th} class is represented by μ_i. Then the mean of the overall data set is calculated and is represented by $\boldsymbol{m}$.

- Then $\hat{\Sigma}$ the covariance matrix of each class is calculated using the following equation:

$$\hat{\Sigma} = (x_i - \mu_i)(x_i - \mu_i)^T \; ; \tag{5.15}$$

- The between class scatter matrix $\mathbf{S}_B$ is calculated using the following equation:

$$\mathbf{S}_B = \sum_{i=1}^{C} \frac{n_i}{n} (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \; ; \tag{5.16}$$

- Then the with-in class scatter matrix is calculated using:

$$\mathbf{S}_W = \sum_{i=1}^{C} \frac{n_i}{n} \hat{\Sigma}_i \; ; \tag{5.17}$$

- The solution that maximized the with-in class scatter and the between class is found by calculating the eigenvectors as given by:

$$\mathbf{S}_B \mathbf{A} = \mathbf{S}_W \mathbf{A} \Lambda \; , \tag{5.18}$$

  where $\mathbf{A}$ is the matrix whose columns are transformed vectors $\mathbf{a}_i$ and Λ is the diagonal matrix of eigenvalues.

- The eigenvectors corresponding to the highest eigenvalue are used for feature extraction.

The discriminant functions are devised by multiplying the new feature vector with the original vectors and finding the constant that will maximize the separability between classes, as given by the following equation.

$$\mathbf{g}_i = \mathbf{a}_i^T \mathbf{x} + a_0 \; , \tag{5.19}$$

where a_0 is the constant and $\mathbf{g}_i$ is the discriminant function of the ith class.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5.3.2 MPEG-7 Motion descriptor based classification

In order to evaluate the efficacy of using MPEG-7 motion descriptors for indexing of American football plays, the 98 dimensional motion feature vector was used to classify 4 events from the football game. The classification was done on a database of 200 video shots taken from 4 different games, as detailed in Section 3.5.

Table 5.1 shows the classification results obtained by using MPEG-7 motion descriptor feature model. This feature model consisted of a 96 dimensional feature map of motion magnitude and direction as well as 2 dimensional feature of mean and standard deviation of frames with the highest motion activity.

Play Category	Classification accuracy
Pass Plays	79.5%
Run Plays	92.5%
FG/Extra Point Plays	87.5%
Kickoff/Punt Plays	65.5%

Table 5.1: Classification Summary Table using MPEG-7 motion descriptor features

The overall classification accuracy of the system is 82.5%. As seen from Table 5.1 we get best classification results for running plays and the worst results for kickoff/punt plays. The kickoff/punt play category is classified only 65.5% correctly with most of the missed classifications falling in the pass play category. Table 5.2 shows the confusion matrix of the four category classification.

	Classified category			
Actual category	Pass	Run	FG/XP	K/P
Pass	70	13	0	5
Run	5	62	0	0
FG/Extra point	0	1	14	1
Kickoff/Punt	10	0	0	19

Table 5.2: Confusion matrix between categories using MPEG-7 motion descriptor features for classification

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Most of the confusion is between passing play category and kickoff/punt category. This can be due to the fact that the kickoff/punt category has the ball catching and running actions as well as the kicking action.

5.3.3 MPEG-7 Audio descriptor based classification

The MPEG-7 audio descriptor feature vector was extracted to complement the motion descriptor features. Before we evaluate the fusion of motion and audio descriptors, we first evaluate the pro and cons of using audio descriptors only for event classification.

Table 5.3 shows the classification results obtained by using MPEG-7 audio descriptor feature model on a database of 200 video shots as detailed in Section 3.5. This feature model consisted of a 10 dimension feature vector of audio spectrum envelope descriptor, a 10 dimension feature vector of audio spectrum centroid and a 10 dimension feature vector of audio spectrum flatness.

Play Category	Classification accuracy
Pass Plays	65.9%
Run Plays	32.8%
FG/Extra Point Plays	0.0%
Kickoff/Punt Plays	55.2%

Table 5.3: Classification Summary Table using MPEG-7 audio descriptor features

We can see that the classification rate is very low compared to the motion descriptor classification. The overall classification accuracy of 48.0% was achieved using only MPEG-7 audio descriptor feature vector. This feature vector tried to classify play events into four categories based on general spectral characteristics of the audio signal. It can be seen from Table 5.3 that the category of Field goal and Extra points had all the shots miss classified. This could be attributed to fact that during these plays the commentators only comment if the attempt to score was good or not. In some cases the commentators are talking about the previous plays, till the kick was made and by the time they mention the outcome of the play, the video is cut into another shot. Table 5.4 shows the confusion matrix of play classification.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

	Classified category			
Actual category	Pass	Run	FG/XP	K/P
Pass	58	22	0	8
Run	40	22	0	5
FG/Extra point	10	6	0	0
Kickoff/Punt	13	0	0	16

Table 5.4: Confusion matrix between categories using MPEG-7 audio descriptor features for classification

It can be seen from the confusion matrix in Table 5.4 that the audio descriptor features do not contain much discriminating power to categorize the play events, as most of the plays are classified into the first category of pass plays. But the last category of kickoff/punt plays achieved much better results than any other category. Therefore this feature can potentially be combined with motion descriptor features to improve the classification accuracy of the kick/punt category.

5.3.4 MFCC feature based classification

The MPEG-7 audio descriptor feature was purely a spectral feature, representing the various spectral characteristics of the audio signal. Thus we require another audio feature that has been proven to be robust in the speech recognition and general sound recognition classification problems. MFCC have been established to provide a feature set that can be used for general sound recognition.

Here we first evaluate the efficacy of using the MFCC features only before combining them with the MPEG-7 motion and audio descriptor features. Table 5.5 shows the classification results of using 12×64 feature matrix of MFCC features.

The overall classification accuracy is 57.5%. But using these features helps discriminate between the pass plays and run plays much better than only using MPEG-7 audio descriptors. These features are not very good in classifying the kickoff/punt category. Therefore using these features in combination with MPEG-7 audio descriptors will help in achieving better classification accuracy.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5.3.5 Multi Modal feature based classification

Here we compare the efficacy of combining the MPEG-7 motion descriptor features with MPEG-7 audio descriptor features and also the MFCC feature set. Table 5.6 summarizes the results of classification by combining the multi-modalities.

From Table 5.6 we can see the increase in classification accuracy by combining multi-modal features. In the case of combining the MPEG-7 audio with MFCC features we see an overall increase of 10%, while combining the audio features with motion descriptor features shows an increase of 5%. Combining all three features produce an overall classification result of 92.5%

All the results that were presented in this work are based on using Fisher's LDA classification technique. The database contained 200 video shots and in order to minimize the bias of the sample set we implemented leave-one-out classification. With this method one sample from the database sample set is removed and used as the test set. The classifier is trained with the rest of the samples. This process is repeated with each sample in the database. This process ensures that classification scheme does not contain bias due to sample set size. [34].

Feature selection was also performed using the Wilk's Lambda criterion in order to optimize the feature space. The dimension of our feature space is large and some of the features may not enhance discrimination between classes. Therefore in the feature selection phase the features that provide redundancy and deteriorate the performance of the overall classification accuracy are taken out of the equation.

Play Category	Classification accuracy
Pass Plays	68.2%
Run Plays	62.7%
FG/Extra Point Plays	43.8%
Kickoff/Punt Plays	20.7%

Table 5.5: Classification Summary Table using MFCC features

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

5.4 Conclusions

The primary design objective of the video indexing system was to utilize features and classification schemes that are fast, effective and simple to implement on hardware. The secondary objective was to evaluate the contribution of different multi-modal features. Also taken into consideration was the fact that domain knowledge plays an important part in the fine tuning of the system.

As mentioned before, the knowledge base model that we have proposed contains 3 categories of RVS events, namely run plays, pass plays and kicking plays. The kicking plays category is further sub divided into two categories namely, Field Goal/ Extra point and Kickoff / Punt. The reason for dividing them is the totally different type of motion each play category exhibits.

In this Chapter we have proposed and implemented an indexing system that utilizes MPEG-7 motion descriptors features, MPEG-7 audio descriptor features and MFCC features to classify American football plays into 4 categories mentioned above. The system first extract the features and then uses LDA to classify them.

In this Chapter we have shown the efficacy of using MPEG-7 motion descriptor features, MPEG-7 audio descriptor features and MFCC feature sets. We have also shown the classification results when the feature sets are used in a combination.

We have established that motion features best discriminate between the plays with an accuracy rate of 83%. But using only motion features cannot resolve the classification confusion between pass play category and kickoff/punt category.

We also established that some MPEG-7 audio descriptor features provide good

Play Category	MPEG-7 audio MFCC	MPEG-7 motion audio	MPEG-7 motion MFCC	MPEG-7 motion audio + MFCC
Pass	70.5%	85.2%	85.2%	94.3%
Run	59.7%	91.0%	92.5%	89.6%
FG/XP	75.0%	87.5%	87.5%	93.8%
K/P	69.0%	82.8%	82.8%	93.1%
Overall	67.0%	87.0%	87.5%	92.5%

Table 5.6: Classification Summary Table using multi-modal features

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

discrimination between kickoff/punt category and the other categories. Therefore we established that combining MPEG-7 audio features with MPEG-7 motion descriptors provides better classification accuracy. Table 5.6 shows that using MPEG-7 motion and audio in combination we can improve the classification accuracy from 82.5% to 87.0%. Most of the improvement is in the kickoff/punt category where the classification accuracy jumped from 65.5% to 82.8%.

We can conclude from our implementation of the system that using MPEG-7 motion and audio descriptors in combination with MFCC for classification of RVS events in American football can be very effective. We can also conclude that MPEG-7 descriptors can be readily used in sports indexing and retrieval applications.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Chapter 6

Conclusions

IN this thesis work we have proposed an American football video indexing system utilizing MPEG-7 motion and audio descriptors along with MFCC features. In Chapter 1, a knowledge base for American football was proposed, using the concept of Recurrent Visual Semantic (RVS) at its root. Chapter 2 provided an overview of the MPEG-7 standard and highlighted the descriptors that were most relevant to this work. Then in Chapter 3, the proposed system was outlined and an explanation on the motivation behind using the proposed techniques is provided. In Chapter 4, we proposed and implemented an algorithm to detect the play events within the video shots. A comparison of using different parameters in the algorithm is also done in the chapter. In Chapter 5, we implemented the indexing and classification phase of the proposed system. The classification results of using motion descriptors, audio descriptors and MFCC feature sets is also provided in the chapter. The chapter concluded by analyzing the effects of combining features from multiple modalities. In this Chapter, we will summarize the results of the overall system and also provide some recommendation for future enhancement of the system.

6.1 Summary of Thesis contribution

In this work we proposed a system that consisted of two main components: First was the localization phase of the system, which dealt with finding the starting point of the play event in the video shots. Second was the indexing and classification phase which was responsible for the extraction of features and classification of the video

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

shots into 4 categories. In the following sections the results of the two main phases of the system are summarized.

6.1.1 Play event detection

In Chapter 4 of this thesis work we proposed an algorithm to localize the play event within a video shot. This was done with the motivation that in some sports there are a lot of non-play events that occur before the play event occurs for only a short period of time. For example in football the total play time is 1 hour, but it takes approximately 3 hours to play the entire game. Also in between each play the teams have 40 seconds to setup and start the play. Similarly in sports like golf, bowling, baseball and tennis, the play event is followed by non-play events.

The algorithm we proposed took advantage of the fact that most of the play events are preceded with low motion intensity segments. The proposed algorithm utilized the mean and standard deviation of the motion vectors from the P frames of the MPEG-1 video stream. 83% of the time, the proposed algorithm was able to detect the starting point of the play events in the video shots within one second before or after the start of the play.

Based on our implementation and results obtained we can make the following conclusions:

- Localization of play events within a video shot is important as it removes non-essential data from our feature space and reduces the processing time.
- Collaboration of mean and standard deviation of the magnitude of motion vectors helped in enhancing the detection algorithm, as in some cases using only the mean may not be reliable.
- There is a trade off that has to be considered when detecting the starting point, accuracy versus delay in detection. As we saw by increasing the window size from 3 frames to 4 frames we got better accuracy but some of the plays were detected after six seconds from the actual starting point.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- We have established that MPEG-7 motion descriptors can be utilized to build applications that can automatically summarize sports events.
- The algorithm that we proposed is only viable for sports that contain low intensity motion just before the actual play event takes place. For example, tennis, baseball etc.
- The proposed algorithm can be plugged into a larger application system without much modifications in order to generate highlights of a game or detect interesting plays.

6.1.2 Play events classification

In Chapter 5 of this work we proposed a system to extract features based on MPEG-7 audio and motion descriptors as well as MFCC. The motivation was to develop an application of indexing and retrieval for American football games using primarily MPEG-7 descriptors, since one of the main objectives of MPEG-7 standardization efforts was to create an interface environment that can facilitate the application development of indexing and retrieval based systems.

MPEG-7 motion descriptors were primarily used to classify the events into the 4 RVS categories. MPEG-7 audio descriptors were used to complement and enhance the classification process. MFCC were used due to the fact that there were studies [35] that showed that MFCC performed better than MPEG-7 descriptors of Audio Spectrum Basis and Audio Spectrum Projections. These two MPEG-7 audio descriptors are very similar to features obtained through MFCC.

We were able to classify the events into 4 categories with an accuracy rate of 92.5% by using all the three feature sets. Using only MPEG-7 descriptor based features we were able to get classification accuracy of 87.0%. All the classification results were obtained by using Fisher's LDA and implementing a leave-one-out classification criteria in order to minimize the bias of the database which contains 200 video shots from 4 different games taken from 4 different networks.

Based on our implementation and results obtained we can make the following

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

conclusions:

- MPEG-7 motion descriptors are integral in classification of RVS events in American football. Using these simple features we were able to get 82.5% classification accuracy.

- Combining multi-modal features in a reasonable fashion can enhance the classification. But always there are trade-offs that need to be considered. Some features may reduce classification of a particular category but may enhance the overall performance of the system. Figure 6.1 shows the variations in classification results from adding audio features to the motion features.

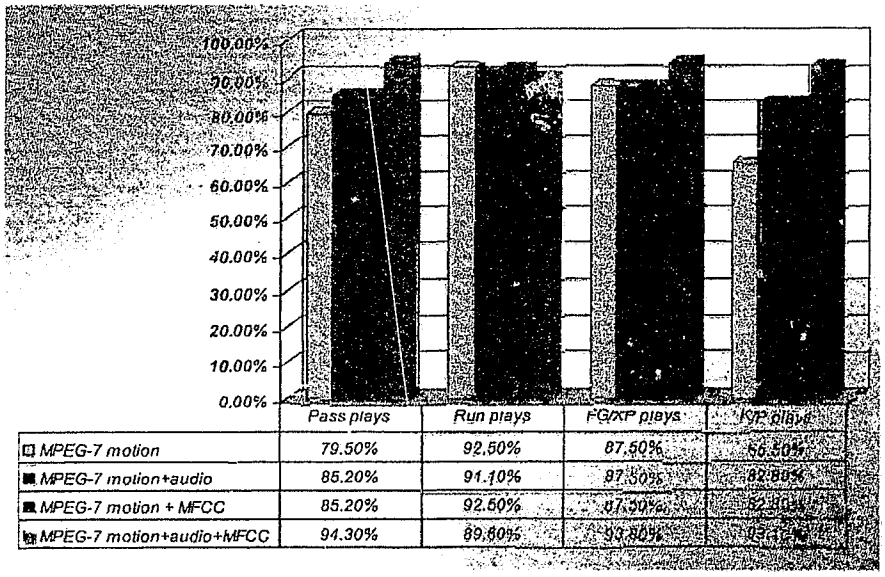

Figure 6.1: Multi-modal classification

- Although there is no baseline to compare our results with, since there is no standard database of American football. Somewhat similar works done in indexing and retrieval of American football events [6] [26] have shown precision accuracy of 81% and 84% respectively. In this work the system classification

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

accuracy is 82.5% by using MPEG-7 motion features only and increases up to 92.5% when all the audio visual features are combined. Table 6.1 shows the comparison between the proposed work and some of the previous works.

System	Events Classified	Performance
Proposed System	Pass,Run, FG/XP, K/P	92.5%
Miyauchi et. al. [26]	FG,TD	83.7%
T. Caelli et. al. [6]	3 types of formations	80.6%
Nitta et. al. [29]	Scrimmage,FG/XP, K/P	84.3%

Table 6.1: Performance Comparison of NFL Video Indexing System

- There is still some work that needs to be done in order to reduce the miss classification between kickoff/punt plays and pass plays. This is evident from the scatter plot shown in Figure 6.2. In the plot we can see that there is a lot of overlap between the categories 1 and 4.

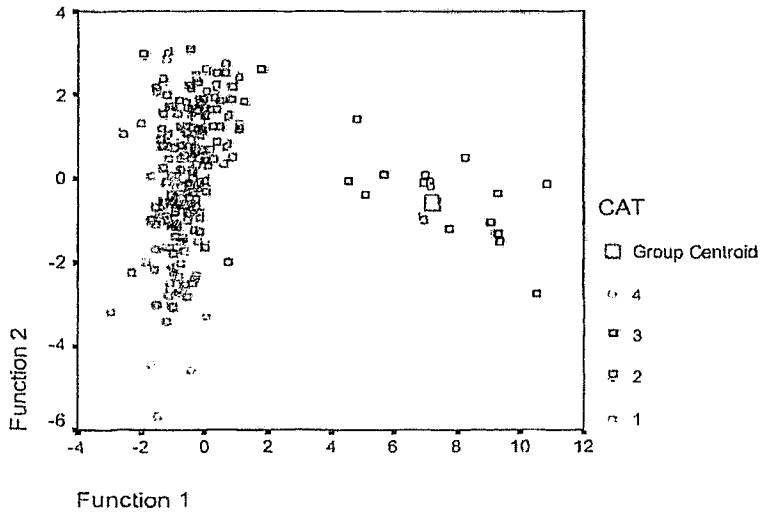

Figure 6.2: Scatter plot of classified data

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

- In the system we have utilized a supervised classification scheme. The performance of the system needs to be tested by using supervised as well as unsupervised classification schemes.

6.2 Future Directions

In the thesis work we have proposed a system for indexing of American football video shots utilizing mainly MPEG-7 motion and audio descriptors. The following directions can be taken to further the work undertaken in this thesis:

- In order to minimize the miss classifications between pass plays and kickoff/punt plays, we can utilize the MPEG-7 texture descriptors. This can help in discriminating the kickoff/punt plays as in these plays the ball is kicked high which in turn makes the camera zoom out to capture the trajectory of the ball. Thus in a video shot there is not only playing field but also audience in the stands. The frames with audience will have a higher texture compared to a frame with only playing field.

- Currently we are only working with video shots and not a whole footage of the game. One of the steps that can be implemented is parsing of video footage into shots, thus providing a more complete system.

- In the future a more complex classification scheme can be utilized which takes into consideration domain knowledge. This knowledge can be based on a state flow diagram of American football events.

- In this work we only classified the events at the root of our proposed knowledge base. In the future we can examine how to classify the inner nodes of the knowledge base tree.

- Further the work on retrieval by mapping the classification results to the MPEG-7 proposed visual descriptor objective measure of retrieval efficiency given by *Average Normalized Modified Retrieval Rate* (**ANMRR**).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bibliography

[1] Jurgen Assfalg, Marco Bertini, Carlo Colombo and Alberto Del Bimbo, "Extracting semantic information from news and sport video" *Proceedings of the International Symposium on Image and Signal Processing and Analysis*, 2001, pp. 4–11,

[2] A. Jaimes and S.-F. Chang, "Learning visual object filters and agents for on-line media" *ADVENT Project Technical Report*, Columbia University, June 1999.

[3] N. Babaguchi, Y. Kawai, and T. Kitahashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration" *IEEE Transactions on Multimedia*, vol. 4, pp. 68–75, 2002.

[4] Y. Chang, W. Zeng, I. Kamel and R. Alonso, "Integrated Image and Speech Analysis for Content-Based Video Indexing" *IEEE International conference on Multimedia computing and systems*, pp. 306–313, 1996.

[5] Z. Xioing, R. Radhakrishnan, and A. Divakaran, "Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Extraction Framework" *IEEE Conference on Image Processing (ICIP)*, vol. 1, pp. 29–32, September 2003.

[6] Terry Caelli, Mihai Lazarescu, Svetha Venkatesh and Geoff West, "On the automated interpretation and indexing of American football," *IEEE International conference on Multimedia computing and systems*, vol. 1, pp. 802–806, 1999.

[7] MPEG Requirements Group, "MPEG-7 Overview," Doc. ISO/MPEG N4317, Sydney MPEG Meeting July 2001.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[8] *http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.*

[9] B.S. Manjunath, Phillipe Salembier and Thomas Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley and Sons, England 2002.

[10] K. Yoon et al. "MPEG-7 based news browsing: description extraction, browsing and exchange," *Multimedia Systems and Application IV*, Proc. of SPIE, vol. 4518 Denver 2001.

[11] W.W. Cohen and W. Fan, "Web-collaborative filtering: recommending music by crawling the web," *Proc. Ninth International World Wide Web Conference*, pp. 685–698, Amsterdam 2000.

[12] A. Ghias, J. Logan, D. Chamberlin and B.C. Smith "Query by Humming: musical information retrieval in an audio database," *Proc. of Third ACM International Conference on Multimedia*, pp. 231–236, San Francisco 1995.

[13] E. Kasutani and A. Yamada, "The MPEG-7 color layout descriptor. A compact image feature description for high speed image/video segment retrieval," *Proc. of IEEE Internation Conference Image Processing ICIP2001*, vol. 1, pp. 674–677, 2001.

[14] Jurgen Assfalg, Marco Bertini, Carlo Colombo and Alberto Del Bimbo, "Semantic annotation of sports video" *IEEE Journal on Multimedia*, vol. 9, Issue 2, pp. 52–60, April–June 2002.

[15] Cees G.M. Snoek and Marcel Worring, "A Review on multimodal video indexing," *Procedings of IEEE International conference on Multimedia and Expo 2002, ICME'2002*, vol. 2, pp. 21–24, 2002.

[16] Arun Hampapur, "Semantic Video Indexing: Approach and Issues," *ACM SIGMOD Record*, vol. 28, no. 1, pp. 32–39, May 1999.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[17] J. Kittler, K. Messer, W.J. Christmas, B. Levienaise-Obadia and D. Koubaroulis, "Generation of semantic cues for sports video annotation," *Proceedings of IEEE International conference on Image Processing*, vol. 3, pp. 26–29, 2001.

[18] Wensheng Zhou, Asha Vellaikal and C.C. Jay Kuo, "Rule based video classification system for basketball video indexing," *http://www.acm.org/sigs/sigmm/MM2000/ep/zhou/*.

[19] Hisashi Miyamori and Shun-ichi Iisaku. "Video annotation for content based retrieval using human behaviour analysis and domain knowledge," *IEEE 4th International conference on Automatic face and gesture recognition* pp. 320–325 2000.

[20] Hong Lu and Yap-Peng Tan, "Sports video analysis and structuring," *IEEE 4th Workshop on Multimedia signal processing*, pp. 45–50. 2001.

[21] Milan Petkovic, Vojkan Mihajlovic, Willem Jonker, S. Djordjevic-Kajan, "Multi-Modal extraction of highlights from TV Formula 1 programs," *US Patent no. 3069654*, 2002.

[22] Chuan Wu, Yu-Fei Ma, Hong-Jiang Zhang and Yu-Zhuo Zhong, "Events recognition by semantic inference for sports video," *Procedings of IEEE International conference on Multimedia and Expo*, vol. 1, pp. 805–808, 2002.

[23] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati and P. Pala. "Detection and recognition of football highlights using HMM," *IEEE 9th International conference on Electronics, Circuits and Systems*, vol. 3, pp. 1059–1062, 2002.

[24] Mei Han, Peng Chang and Yihong Gong, "Extract highlights from baseball game video with hidden Markov models," *IEEE International conference on Image processing*, pp. 609–612, 2002.

[25] Mei Han, Wei Hua, Wei Xu and Yihong Gong, "An Integrated baseball digest system using maximum entropy method," *Proc. ACM conference on Multimedia*, pp. 347–350, December 2002.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[26] Shingo Miyauchi, Akira Hirano, Noboru Babguchi and Tadahiro Kitahashi, "Collaborative multimedia analysis for detecting semantical events from broadcasted sports video," *Proceedings of IEEE 16th International conference on Pattern recognition*, vol. 2, pp. 1009–1012 2002.

[27] S.S. Stevens, J. Volkmann and E.B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of Acoustical Society of America*, vol. 8, pp. 185–190, Jan. 1937.

[28] S.S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *American Journal of Psychology*, vol. 53, pp. 329–353, July 1940.

[29] N. Nitta, N. Babaguchi and T. Kitahashi, "Extracting actors, actions and events from sports video - A fundamental approach to story tracking," *Proc. IEEE Intl. Conf. on Pattern Recognition*, vol. 4, pp. 718–721, 2000.

[30] Baoxin Li and M. Ibrahim Sezan, , 2001, CBAIVL 2001, pg(s): 132-138. R. Suleesathira and L.F. Chaparro, "Event detection and summarization in sports video," *IEEE Workshop on Content-Based access of image and video libraries, CBAVIL 2001* , pp. 132–138, 2001.

[31] Hyoung-Gook Kim, Nicolas Moreau and Thomas Sikorau, "Audio Classification Based on MPEG-7 Spectral Basis Representations," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 5, May 2004.

[32] H. Combrinck and E. Botha, "On the Mel-scaled Cepstrum," *http://citeseer.nj.nec.com/524151.html*, 1996.

[33] Andrew R. Webb, *Statistical Pattern Recognition*, 2nd edition, Wiley Publishing USA, 2002.

[34] Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd edition, Academic Press, 1990.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

[35] Z. Xiong, R. Radhakrishnan, A. Divakaran, T.S. Huang, "Comparing MFCC and MPEG-7 Audio Features for Feature Extraction, Maximum Likelihood HMM and Entropic Prior HMM for Sports Audio Classification," *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 3, pp. 397–400, July 2003

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.